Supplement to "Combining xQTL and genome-wide association studies from ethnically diverse populations improves druggable gene discovery"

Noah Lorincz-Comi[1,2], Wenqiang Song[1,2], Xin Chen[1,2], Isabela Rivera Paz[1,2], Yuan Hou[1,2], Yadi Zhou[1,2], Jielin Xu[1,2], William Martin[1,2], John Barnard[3], Andrew A. Pieper[4,5,6,7,8,9], Jonathan L Haines[10], Mina Chung[11,12,13], Feixiong Cheng[1,2,13,*]

[1]Cleveland Clinic Genome Center, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[2]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[3]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[4]Department of Psychiatry, Case Western Reserve University, Cleveland, OH 44106, USA
[5]Brain Health Medicines Center, Harrington Discovery Institute, University Hospitals Cleveland Medical Center, Cleveland, OH 44106, USA
[6]Geriatric Psychiatry, GRECC, Louis Stokes Cleveland VA Medical Center; Cleveland, OH 44106, USA
[7]Institute for Transformative Molecular Medicine, School of Medicine, Case Western Reserve University, Cleveland 44106, OH, USA
[8]Department of Pathology, Case Western Reserve University, School of Medicine, Cleveland, OH 44106, USA
[9]Department of Neurosciences, Case Western Reserve University, School of Medicine, Cleveland, OH 44106, USA
[10]Department of Population & Quantitative Health Sciences, Cleveland Institute for Computational Biology, Case Western Reserve University School of Medicine, Cleveland, Ohio, USA
[11]Department of Cardiovascular Medicine, Heart, Vascular & Thoracic Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[12]Department of Cardiovascular and Metabolic Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA
[13]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH 44195, USA

Table of contents

S1. Stating the null distribution of the gene-based test (GenT) statistic

Let $H_0^\ell: \mathrm{E}(\mathbf{z}_\ell) = \mathbf{0}$ be tested against $H_1^\ell: \mathrm{E}(\mathbf{z}_\ell) \neq \mathbf{0}$ with the statistic $S_\ell = \mathbf{z}_\ell^\top \mathbf{z}_\ell$ for the ℓth gene, where $\mathbf{z}_\ell = \left(Z_{\ell j}\right)_{j=1}^{m^\ell}$ and $Z_{\ell j}$ is the Z-statistic used to test the null hypothesis of no marginal association between the jth SNP corresponding to the ℓth gene and the GWAS phenotype. We now drop the subscript ℓ for notational simplicity. Our primary goal is to find the distribution of $S \coloneqq S_\ell$ under H_0. We begin by demonstrating that this distribution has no closed-form solution when the LD matrix is estimated from a reference panel. We provide a highly precise approximation of the true distribution in this case which uses moment-matching properties of the Gamma distribution from Covo & Elalouf (2014) and Stewart (2007). Finally, we introduce a statistic F which is a scalar-transformed version of $\mathbf{z}^\top \mathbf{z}$ that has a known closed-form null distribution which can be used to provide valid inference of H_0 vs H_1. In the following subsections, $\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ is the true LD matrix between the m tested SNPs, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_i)$. As a result, $\mathbf{z}|H_0 \sim N(\mathbf{0}, \mathbf{R})$ We also consider the transformation $\mathbf{y} = \mathbf{U}^\top \mathbf{z} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$, with elements Y_j. This implies $Y_j^2 \sim \Gamma\left(\alpha_i = \frac{1}{2}, \xi_i = \frac{1}{2\lambda_j}\right)$ under the shape-rate parameterization of Γ. It follows that $\mathrm{Cov}\left(Y_j^2, Y_k^2\right) = 0$ for all $j, k = 1, \ldots, m$ by multiplication by orthogonal eigenvectors and $\mathbf{y}^\top \mathbf{y} = \mathbf{z}^\top \mathbf{z} = S$.

S1.1 Approximation of GenT null distribution

In this section, we introduce our moment-matching method to approximate the null distribution of S and demonstrate that this approximation is highly precise under a range of real-world conditions.

We first show that S does not exactly follow a Γ distribution using properties of the moment generating function under independence of $\left(Y_i^2, Y_j^2\right)$:

$$M_{Y_i^2}(t) = \left(1 - \frac{t}{2\lambda_i}\right)^{-\frac{1}{2}}, \quad t < 2\lambda_i$$

$$M_{\sum_i Y_i^2}(t) = \prod_{i=1}^{n} M_{Y_i^2}(t)$$

$$\neq \left(1 - t\left[2\sum_{i=1}^{n} \lambda_i\right]^{-1}\right)^{-\frac{n}{2}}, \quad t < 2\sum_{i=1}^{n} \lambda_i.$$

However, Stewart (2007) and Covo & Elalouf (2014) showed that a moment-matched Gamma density is a simple and highly accurate approximation to the distribution of S, with increasing precision as $m \to \infty$, even for $m > 3$. We use this section to state our moment-matching method and show in **Figure S1** that this method accurately characterizes the null distribution of S. Moment matching involves finding (α_0, ξ_0) of a

Gamma distribution such that $\mathrm{E}(S|H_0) = \alpha_0/\xi_0$ and $\mathrm{Var}(S|H_0) = \alpha_0^2/\xi_0$, which essentially assumes a priori that $S|H_0 \sim \Gamma(\alpha_0, \xi_0)$. It follows that

$$\xi_0 = \frac{\sum_{j=1}^{m^\ell} \frac{\alpha_i}{\xi_j}}{\sum_{j=1}^{m^\ell} \frac{\alpha_i}{\xi_j^2}} = \frac{\sum_{j=1}^{m^\ell} \lambda_j}{\sum_{j=1}^{m^\ell} 2\lambda_j^2} = \frac{m^\ell}{\sum_{j=1}^{m^\ell} 2\lambda_j^2} \tag{1.1}$$

and

$$\alpha_0 = \xi_0 \sum_{j=1}^{m^\ell} \frac{\alpha_j}{\xi_j} = \frac{\left(m^\ell\right)^2}{\sum_{j=1}^{m^\ell} 2\lambda_j^2}. \tag{2.1}$$

It is also apparent from (1.1) that $\xi_0 = m^\ell/\mathrm{tr}(2\mathbf{RR})$ and from (2.1) that $\alpha_0 = m^\ell \xi_0$. We performed extensive simulations to verify this distributional statement, even when $\mathbf{R}$ is randomly drawn from a Wishart distribution with finite degrees of freedom to emulate LD estimation in practice, the results of which are presented in **Figure S1** and demonstrate that the distribution stated above accurately models the simulated data under the null hypothesis. Full R code to perform these simulations is available from the webpage provided in the **Figure S1** legend.

S1.2 Null distribution simulations

We performed simulations to evaluate how well the aforementioned Gamma distribution can characterize the null distribution of the gene-based tests statistics under the null hypothesis. These simulations are described in the **Simulations** subsection of the **Methods** section in the main text, and here we describe how Type I error was evaluated using them. We began each simulation by generating SNP-level Z-statistics which were correlated according to a fixed structure, either first-order autoregressive or compound symmetry. We then calculated gene-based test statistic and the parameters of the Gamma distribution which are used to as a model under the null hypothesis. Figures S1-3, 6, and 7 show the distribution of the gene-based test statistics across all replicates as a histogram, and have overlaid in blue the theoretical density of them using the Gamma model under the null hypothesis. The displayed Type I error values are the proportion of simulated gene-based test statistics which are greater than the 95$^{\text{th}}$ quantile of the corresponding null Gamma distribution. After an infinite number of simulation replicates, these quantities are exactly the Type I error rates of our test since comparison of the observed gene-based test statistic to the 95$^{\text{th}}$ quantile, or in general the $[100 \times (1-\alpha)]^{\text{th}}$, is used to reject the null hypothesis. All simulation data were generated under the null, so the displayed Type I error values in Figures S1-3, 6, and 7 are also interpreted as the null hypothesis rejection frequencies under the null hypothesis, i.e., the Type I error rates.

S1.3 Exact GenT null distribution under transformation

In this section, we provide a statistic F which is a transformation of the original S that can be used for an exact inference of H_0 vs H_1. This transformation does not remove the LD structure between SNPs used to test H_0, and requires no additional information other than that required for S.

Our goal is to find the closed-form distribution of a scalar-transformed version of $\mathbf{z}^\top\mathbf{z}$ which can be used for valid inference of H_0 vs H_1. Note that $\mathbf{y}^\top\mathbf{y} = \mathbf{z}^\top\mathbf{U}^\top\mathbf{U}\mathbf{z} = \mathbf{z}^\top\mathbf{z}$ such that $\mathbf{y}^\top\mathbf{y}$ and $\mathbf{z}^\top\mathbf{z}$ will have the same distribution. It also follows that $\mathrm{Cov}(Y_i, Y_k) = 0$ for $i \neq k$ since they are each transformed by orthogonal eigenvectors. Therefore, $\mathbf{z}^\top\mathbf{z}$ is equal to the sum of independent Gamma-distributed random variates with constant shape and non-constant rate parameters. The distribution of $\mathbf{y}^\top\mathbf{y}$ has no simple closed-form expression, but it would if the rates were constant. We therefore seek a transformation of $\mathbf{y}$ by a constant matrix $\mathbf{C}$ such that the resulting $F_i \coloneqq (\mathbf{C}\mathbf{y})_i \sim \Gamma(\tilde{\alpha}_i, \tilde{\beta}_0)$, where α_i may or may not be constant across $i = 1, \ldots, m$, but $\tilde{\beta}_0$ is. First, we state that $\mathrm{E}(Y_i^2) = \lambda_i$ and $\mathrm{Var}(Y_i^2) = 2\lambda_i^2$ by the properties of the moments of the Gamma distribution, or the properties of a (univariate) quadratic form. Let $\mathbf{C} = \mathrm{diag}(c_i)$ and $c_i = \frac{\mathrm{E}(Y_i^2)}{\mathrm{Var}(Y_i^2)}$. This implies $F_i = c_i Y_i^2 \sim \Gamma\left(\tilde{\alpha}_i = \frac{1}{2}, \tilde{\beta}_0 = 1\right)$ by the following deduction:

$$\mathrm{E}(F_i) = \mathrm{E}(c_i Y_i^2) = \frac{\mathrm{E}(Y_i^2)^2}{\mathrm{Var}(Y_i^2)}, \qquad \mathrm{Var}(F_i) = \mathrm{Var}(c_i Y_i^2) = \frac{\mathrm{E}(Y_i^2)^2}{\mathrm{Var}(Y_i^2)}$$

$$\tilde{\beta}_0 = \frac{\mathrm{E}(F_i)}{\mathrm{Var}(F_i)} = 1$$

$$\tilde{\alpha}_i = \frac{\mathrm{E}(F_i)^2}{\mathrm{Var}(F_i)} = \frac{\left(\frac{\mathrm{E}(Y_i^2)^2}{\mathrm{Var}(Y_i^2)}\right)^2}{\left(\frac{\mathrm{E}(Y_i^2)^2}{\mathrm{Var}(Y_i^2)}\right)} = \frac{\mathrm{E}(Y_i^2)^2}{\mathrm{Var}(Y_i^2)} = \frac{\lambda_i^2}{2\lambda_i^2} = \frac{1}{2}.$$

This transformation therefore fixes both $\tilde{\alpha}_i \coloneqq \tilde{\alpha}_0$ and $\tilde{\beta}_0$ to be constant across $i = 1, \ldots, m$. We now state the moment-generating function (mgf) of F_i:

$$m_{F_i}(t) = (1 - t)^{-1/2}, \quad t < 1$$

and of $\sum_{i=1}^m F_i \coloneqq \mathbf{j}^\top\mathbf{f}$:

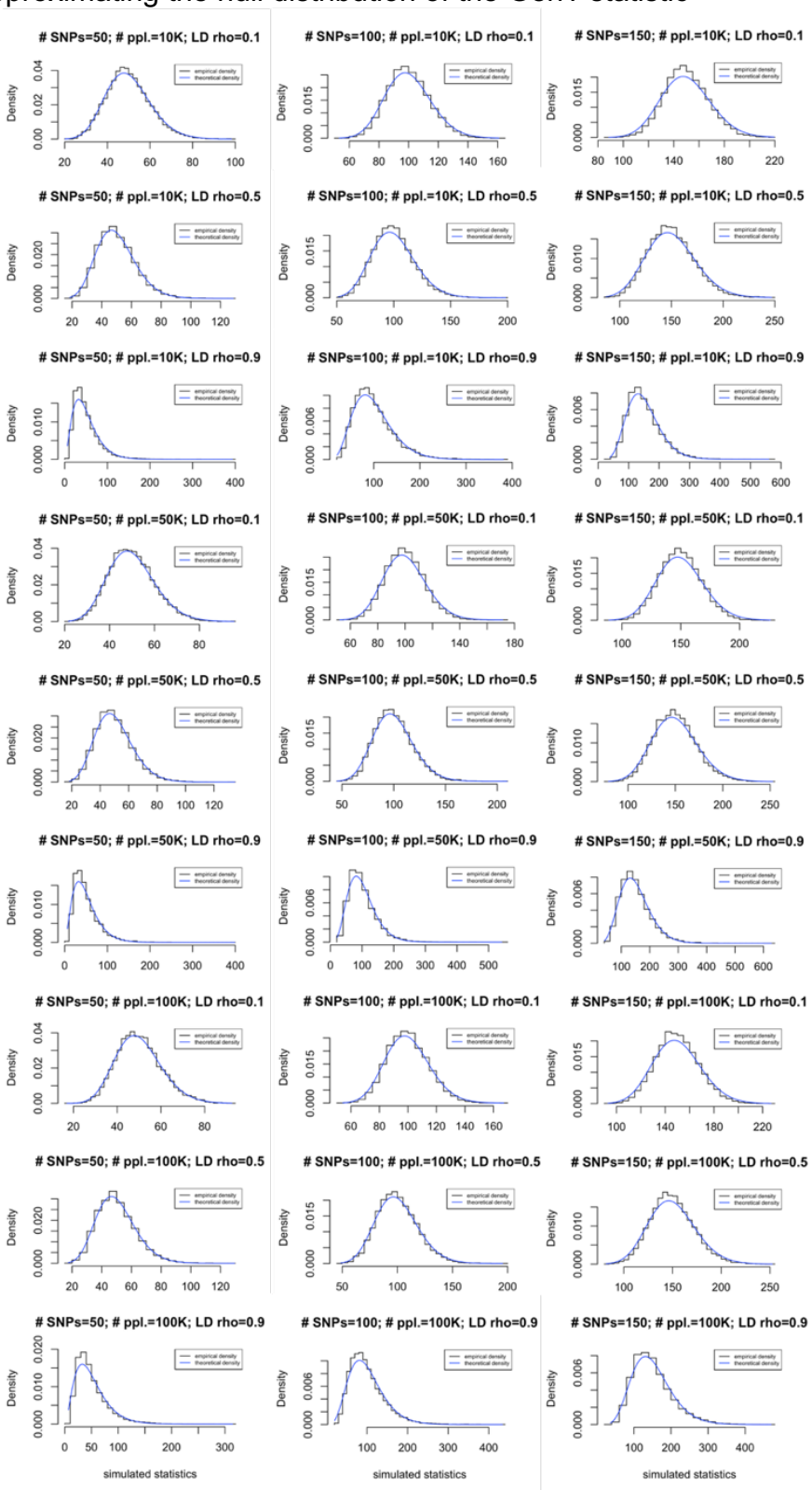
$$\begin{aligned} m_{\mathbf{j}^\top\mathbf{f}}(t) &= \prod_{i=1}^{m} (1 - t)^{-\frac{1}{2}} \\ &= (1 - t)^{-\frac{m}{2}}, \end{aligned}$$

which is the moment generating function of $\Gamma\left(\frac{m}{2}, 1\right)$. This result that

$$\sum_{i=1}^{m} F_i \sim \Gamma\left(\frac{m}{2}, 1\right)$$

suggests that the scalar-transformed version of $\mathbf{z}^\top\mathbf{z}$ has a closed-form distribution when $\mathrm{E}(\mathbf{z}) = \mathbf{0}$. This implies that $\mathbf{j}^\top\mathbf{f}$ can be used for inference of $H_0: \mathrm{E}(\mathbf{z}) = \mathbf{0}$ vs $H_1: \mathrm{E}(\mathbf{z}) \neq \mathbf{0}$. Relating back to the original random vector $\mathbf{z}$, we see that $\mathbf{Uy} = \mathbf{z}$, $\mathbf{y} = \mathbf{C}^{-\frac{1}{2}}\mathbf{f}$, and so $\mathbf{z} = \mathbf{UC}^{-\frac{1}{2}}\mathbf{f} \Rightarrow \mathbf{z}^\top\mathbf{z} = \mathbf{f}^\top\mathbf{C}^{-1}\mathbf{f}$ where $\mathbf{C}^{-1} = \mathrm{diag}[\mathrm{Var}(Y_i^2)/\mathrm{E}(Y_i^2)] = \mathrm{diag}(2\lambda_i)$. **Figure S2** and **S3** display the results of simulations to examine the distribution of $\mathbf{j}^\top\mathbf{f}$ when $\mathbf{R}$ is known and fixed vs when $(500 \times \mathbf{R})$ is drawn from a Wishart distribution with 500 degrees of freedom and covariance parameter $\mathrm{E}(\mathbf{R})$, respectively. **Figure S2** demonstrates that $\Gamma\left(\frac{n}{2}, 1\right)$ is indeed the distribution of $\mathbf{j}^\top\mathbf{f}$ when H_0 is true and $\mathbf{R}$ is known, as evidenced by the corresponding empirical and theoretical densities and by the controlled Type I error rate at 5%. **Figure S3** demonstrates that when $\mathbf{R}$ is only an [unbiased] estimate of true LD structure, inference based on $\mathbf{j}^\top\mathbf{f}$ has inflated Type I error.

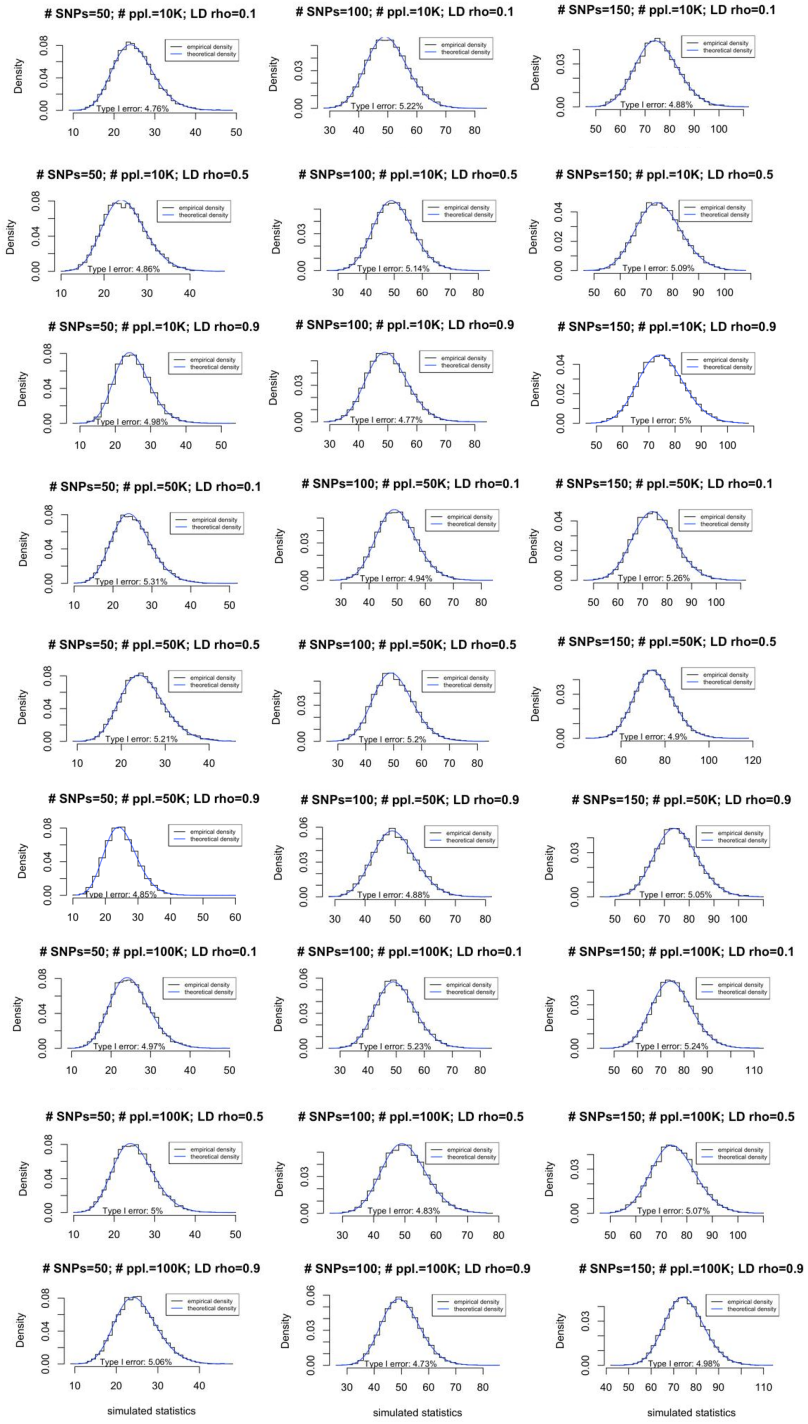
Figure S1: Approximating the null distribution of the GenT statistic



Each panel displays the distribution of simulated test statistics (S) under the GenT null hypothesis (black) and the theoretical density from the distribution parameterized using the derivation above (blue) for varying numbers of tested SNPs, GWAS sample sizes, and densities of the first-order autoregressive LD

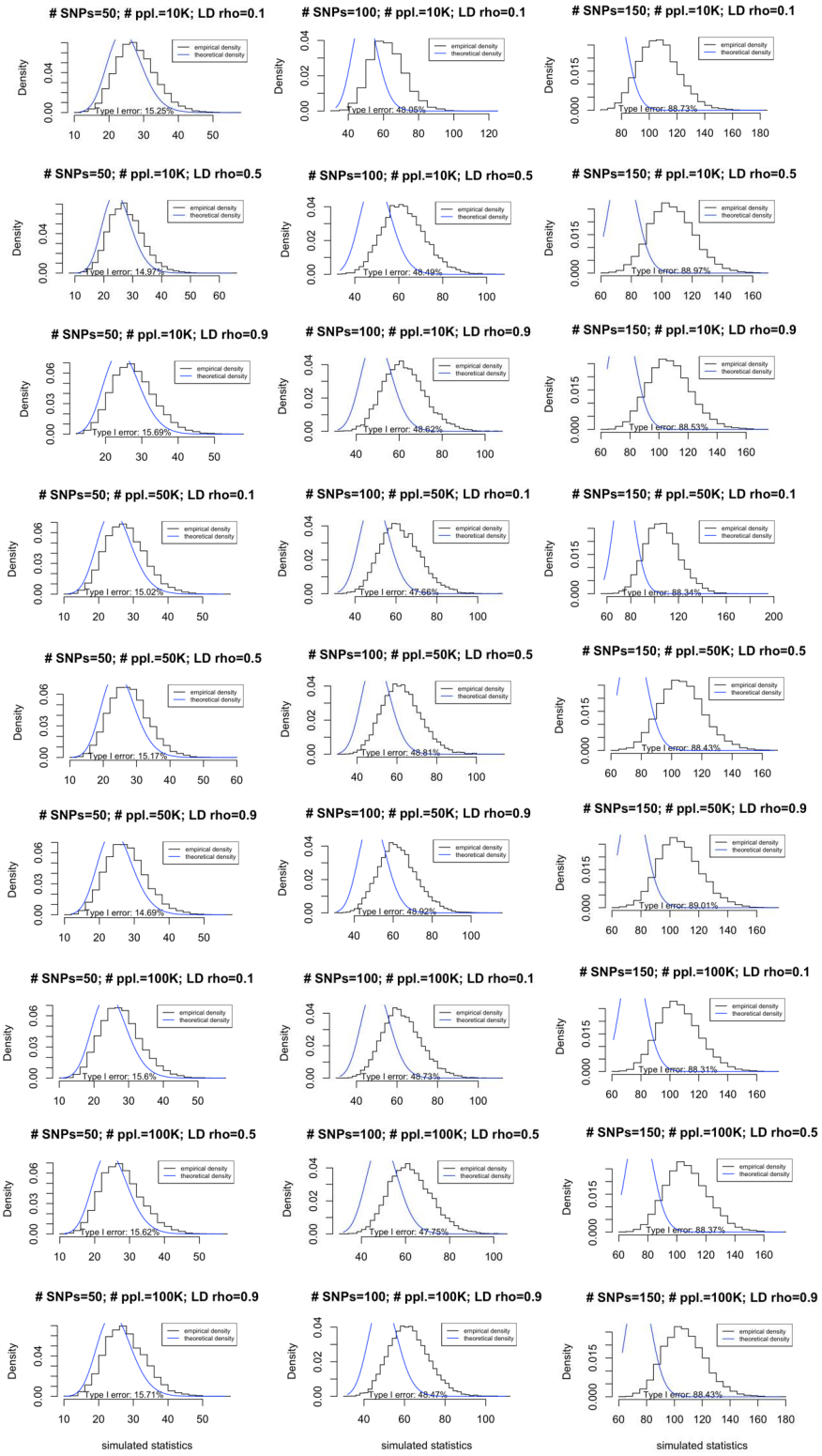
matrix, which was estimated without bias from a Wishart distribution with 500 degrees of freedom. In all cases we generated GWAS summary statistics directly using the procedure described in the Methods section of the main text and R code from github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/gent/gent_distribution.R.

Figure S2: The null distribution of the transformed gene-based test statistic with known LD



Each panel displays the distribution of simulated transformed test statistics ($\mathbf{j}^{\mathrm{T}}\mathbf{f}$) under the GenT null hypothesis (black) and the theoretical density from the distribution parameterized using the derivation in Section S1.3 (blue) for varying numbers of tested SNPs, GWAS sample sizes, and densities of the first-order autoregressive LD matrix, which was fixed and known. In all cases we generated GWAS summary statistics directly using the procedure described in the Methods section of the main text and R code from github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/gent/fgent_distribution.R.

Figure S3: The null distribution of the transformed gene-based test statistic with imprecisely estimated LD



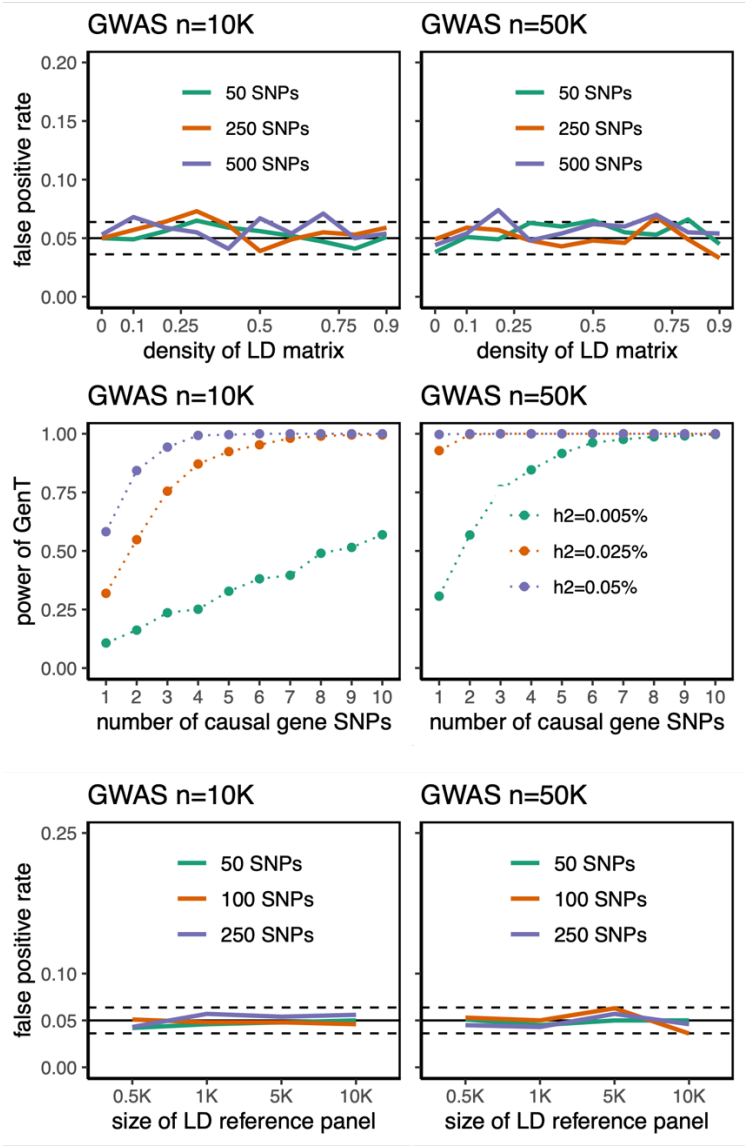
Each panel displays the distribution of simulated transformed test statistics ($\mathbf{j}^{\top}\mathbf{f}$) under the GenT null hypothesis (black) and the theoretical density from the distribution parameterized using the derivation in

Section S1.3 (blue) for varying numbers of tested SNPs, GWAS sample sizes, and densities of the first-order autoregressive LD matrix, which was estimated without bias from a Wishart distribution with 500 degrees of freedom. In all cases we generated GWAS summary statistics directly using the procedure described in the Methods section of the main text and R code from github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/gent/fgent_distribution.R.

S2. GenT simulations with compound symmetry LD structure

In the main text, all simulations performed using GenT used a first-order autoregressive LD structure, and we show in this section analogous results from simulations performed using a compound symmetry LD structure instead. These results are displayed in **Figure S4** and demonstrate that in this case, GenT controls the Type I error rate as the number of gene SNPs, the GWAS sample size, and the density of the LD matrix increases and does not have inflated Type I error when the size of the LD reference panel is relatively small compared to the number of SNPs used to test the gene-based null hypothesis. Similarly to the results presented in the main text, in the case of a compound symmetry LD structure, GenT has increasing power as the gene heritability, number of causal SNPs, and GWAS sample size increases. These results provide additional evidence that the performance of GenT, and the subsequent tests built upon it, are not sensitive to the structure of the LD matrix. As was shown above, the null distribution of the tests statistic used by GenT is based only on one parameter which is generally unknow, but precisely estimable, in practice, which is $\mathrm{tr}(\mathbf{RR})$, where $\mathbf{R}$ is the LD matrix for SNPs corresponding to the tested gene.

Figure S4: GenT performance with compound symmetry LD structure



False positive rates and power of GenT for varying numbers of tested SNPs, GWAS sample sizes, numbers of causal gene SNPs, and sample sizes of the LD reference panel, and densities of the compound symmetry LD matrix. In all cases we generated GWAS summary statistics directly using the procedure described in the Methods section of the main text and R code from github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/gent/gent_simulation.R.

S3. Parameterizing the null distribution of the MuGenT test statistic

In the main text, we stated that for the ℓth gene the covariance matrix of $\text{vec}(\mathbf{Z}_\ell^\top \mathbf{Z}_\ell)$ in the multi-ancestry gene-based test (MuGenT) was equal to $\boldsymbol{\Upsilon}$. We now show the derivation of $\boldsymbol{\Upsilon}$, where $\mathbf{Z}_\ell = (\mathbf{z}_1, \dots, \mathbf{z}_p)$, $\mathbf{z}_k = \left(Z_{jk}\right)_{j,k=1}^{m,p}$, $\mathbf{z}_k \sim \text{Normal}(\mathbf{0}, \mathbf{R}_k)$ and $\mathbf{z}_k \perp \mathbf{z}_s$

for $s \neq k$ under $H_0^{\ell}: \mathrm{E}(\mathbf{Z}_{\boldsymbol{\ell}}) = \mathbf{0}$, and we will hereafter drop the subscript ℓ for notational simplicity. We start by visualizing $\mathbf{Z}^\top\mathbf{Z}$:

$$\mathbf{Z}^\top\mathbf{Z} = \begin{pmatrix} \mathbf{z}_1^\top\mathbf{z}_1 & \mathbf{z}_1^\top\mathbf{z}_2 & \cdots & \mathbf{z}_1^\top\mathbf{z}_p \\ \mathbf{z}_2^\top\mathbf{z}_1 & \mathbf{z}_2^\top\mathbf{z}_2 & \cdots & \mathbf{z}_2^\top\mathbf{z}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}_p^\top\mathbf{z}_1 & \mathbf{z}_p^\top\mathbf{z}_2 & \cdots & \mathbf{z}_p^\top\mathbf{z}_p \end{pmatrix}$$

and

$$\mathrm{vec}(\mathbf{Z}^\top\mathbf{Z}) = \left(\mathbf{z}_1^\top\mathbf{z}_1, \mathbf{z}_2^\top\mathbf{z}_1, \dots, \mathbf{z}_p^\top\mathbf{z}_1, \mathbf{z}_1^\top\mathbf{z}_2, \mathbf{z}_2^\top\mathbf{z}_2, \dots, \mathbf{z}_p^\top\mathbf{z}_2, \dots, \mathbf{z}_1^\top\mathbf{z}_p, \mathbf{z}_2^\top\mathbf{z}_p, \dots, \mathbf{z}_p^\top\mathbf{z}_p\right).$$

All derivation that follows is under H_0^{ℓ}. We now consider three cases for the values of $\boldsymbol{\Upsilon} \coloneqq \mathrm{Cov}[\mathrm{vec}(\mathbf{Z}^\top\mathbf{Z})]$: (i) $\mathrm{Cov}(\mathbf{z}_s^\top\mathbf{z}_s, \mathbf{z}_s^\top\mathbf{z}_k)$, (ii) $\mathrm{Cov}(\mathbf{z}_s^\top\mathbf{z}_k, \mathbf{z}_s^\top\mathbf{z}_k)$, (iii) $\mathrm{Cov}(\mathbf{z}_s^\top\mathbf{z}_s, \mathbf{z}_k^\top\mathbf{z}_k)$. Firstly,

$$\begin{aligned} \mathrm{Cov}(\mathbf{z}_s^\top\mathbf{z}_s, \mathbf{z}_s^\top\mathbf{z}_k) &= \mathrm{E}[\mathrm{Cov}(\mathbf{z}_s^\top\mathbf{z}_s, \mathbf{z}_s^\top\mathbf{z}_k|\mathbf{z}_s)] + \mathrm{Cov}[\mathrm{E}(\mathbf{z}_s^\top\mathbf{z}_s|\mathbf{z}_s), \mathrm{E}(\mathbf{z}_s^\top\mathbf{z}_k|\mathbf{z}_s)] \\ &= \mathrm{Cov}[\mathrm{E}(\mathbf{z}_s^\top\mathbf{z}_s|\mathbf{z}_s), \mathrm{E}(\mathbf{z}_s^\top\mathbf{z}_k|\mathbf{z}_s)] \\ &= \mathrm{Cov}(\mathbf{z}_s^\top\mathbf{z}_s, 0) \\ &= 0. \end{aligned}$$

Secondly,

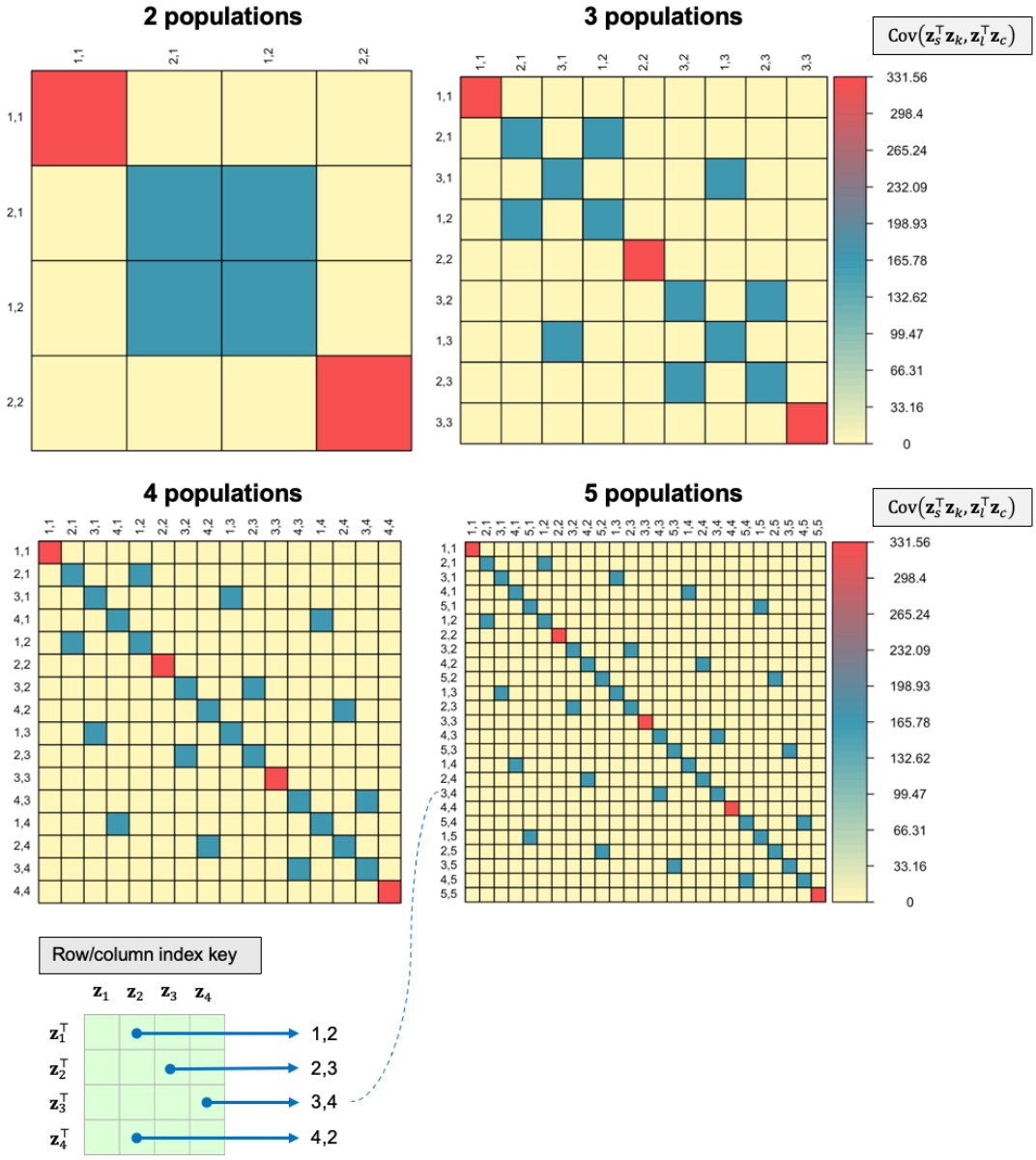
$$\begin{aligned} \mathrm{Cov}(\mathbf{z}_s^\top\mathbf{z}_k, \mathbf{z}_s^\top\mathbf{z}_k) &= \mathrm{E}(\mathbf{z}_s^\top\mathbf{z}_k\mathbf{z}_k^\top\mathbf{z}_s) \\ &= \mathrm{E}[\mathrm{E}(\mathbf{z}_s^\top\mathbf{z}_k\mathbf{z}_k^\top\mathbf{z}_s|\mathbf{z}_k)] \\ &= \mathrm{E}[\mathrm{tr}(\mathbf{R}_s\mathbf{z}_k\mathbf{z}_k^\top)] \\ &= \mathrm{E}(\mathbf{z}_k^\top\mathbf{R}_s\mathbf{z}_k) \\ &= \mathrm{tr}(\mathbf{R}_s\mathbf{R}_k). \end{aligned}$$

Thirdly,

$$\begin{aligned} \mathrm{Cov}(\mathbf{z}_s^\top\mathbf{z}_s, \mathbf{z}_k^\top\mathbf{z}_k) &= \mathrm{E}(\mathbf{z}_k^\top\mathbf{z}_k\mathbf{z}_s^\top\mathbf{z}_s) - \mathrm{E}(\mathbf{z}_k^\top\mathbf{z}_k)\mathrm{E}(\mathbf{z}_s^\top\mathbf{z}_s) \\ &= \mathrm{tr}\{\mathrm{E}[\mathrm{E}(\mathbf{z}_s\mathbf{z}_k^\top\mathbf{z}_k\mathbf{z}_s^\top|\mathbf{z}_k)]\} - \mathrm{tr}(\mathbf{R}_s)\mathrm{tr}(\mathbf{R}_k) \\ &= \mathrm{tr}\{\mathrm{E}[\mathbf{z}_k^\top\mathbf{z}_k\mathrm{E}(\mathbf{z}_s\mathbf{z}_s^\top)]\} - m^2 \\ &= \mathrm{E}(\mathbf{z}_k^\top\mathbf{z}_k)\mathrm{tr}(\mathbf{R}_\mathrm{s}) - m^2 \\ &= \mathrm{tr}(\mathbf{R}_k)\mathrm{tr}(\mathbf{R}_\mathrm{s}) - m^2 \\ &= m^2 - m^2 \\ &= 0. \end{aligned}$$

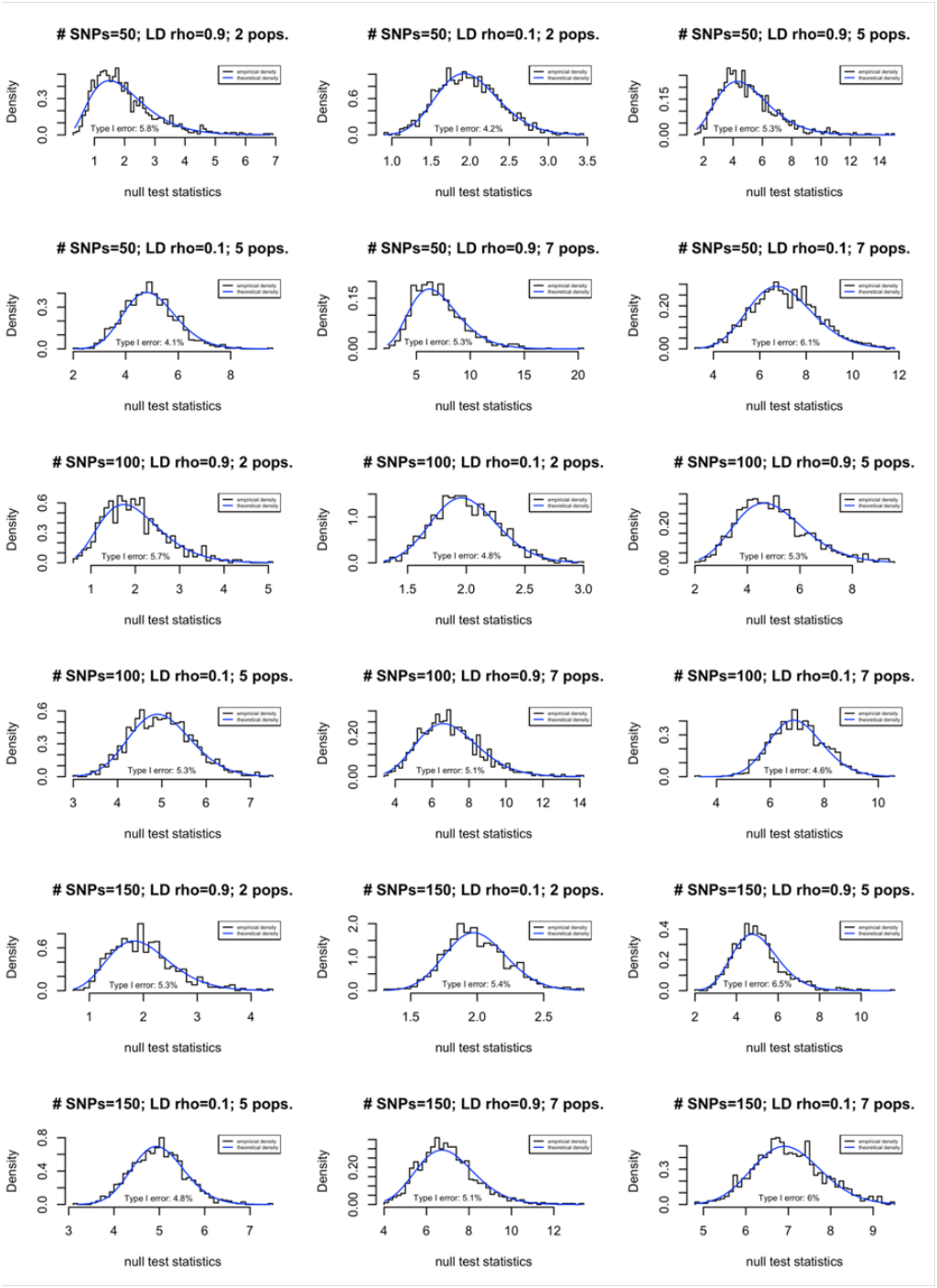
Examples of the structure of the resulting covariance matrix are presented in **Figure S5** and extensive simulations demonstrating that it correctly parameterizes the approximated null distribution of the MuGenT test statistic is displayed in **Figure S6.**

Figure S5: Structure of the MuGenT test statistic inner covariance matrix (Υ)



Example structures of the inner covariance matrix Υ derived in the text for 2-5 populations. R code used to produce these figures is available from github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/mugent/mugent_varmat_simulation.R.

Figure S6: Demonstrating the correct parameterization of the null distribution of the MuGenT test statistic



Each panel displays the distribution of simulated test statistics under the MuGenT null hypothesis (black) and the theoretical density from the distribution parameterized using the derivation above (blue) for varying numbers of tested SNPs, GWAS sample sizes, and densities of the first-order autoregressive LD matrix. Also displayed are the Type I error rates at the 5% significance level. In all cases we generated GWAS summary statistics directly using the procedure described in the Methods section of the main text and R code from github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/mugent/mugent_distribution.R.

S4. Parameterizing the null distribution of the MuGenT-PH test statistic

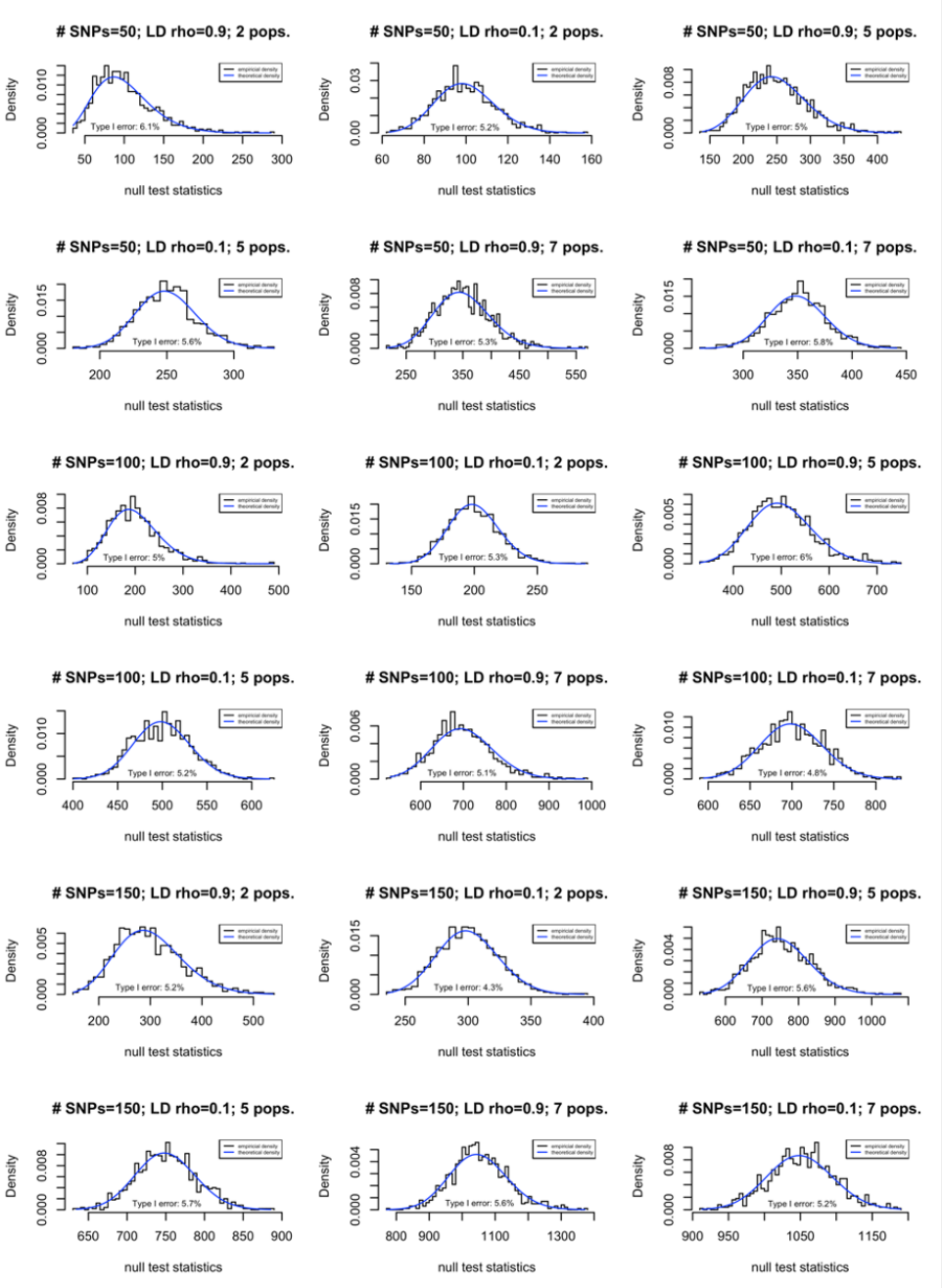
We now derive the quantity v^{ℓ} which was stated to be the variance of the statistic S^{ℓ}_{M-PH} in the main text, i.e. the statistic using by MuGenT-PH to test for population hetereogeneity in the effect sizes of association between the ℓth gene and the GWAS phenotype across p populations. Let $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jp})^{\top}$ be the p-length vector of associations between the jth SNP and GWAS phenotype in each of the p populations, of which there are m for the ℓth gene. MuGenT-PH begins by testing $H_{0j}: \cap_{1\le s<k<p}\, \beta_{jk} = \beta_{js}$ vs $H_{1j}: \cup_{1\le s<k<p}\, \beta_{jk} \neq \beta_{js}$ using the statistic $Y_j \coloneqq \sum_{k=1}^{p} Z_{jk}^2$ which follows $\chi^2(p)$ under H_{0j}. Tests of H_{0j} can be alternatively considered using a SNP-level grand mean effect size, τ_j, which implies that the former statement of the SNP-level hypotheses are equivalent to H_{0j}:$\cap_{k=1}^{p}\, \beta_{jk} - \tau_j = 0$ vs H_{1j}:$\cup_{k=1}^{p}\, \beta_{jk} - \tau_j \neq 0$ using the Z-statistics $(Z_{j1}, \dots, Z_{jp})$. MuGenT-PH tests $H_0: \cap_{j=1}^{m} \cap_{k=1}^{p}\, \beta_{jk} - \tau_j = 0$ vs $H_1: \cup_{j=1}^{m} \cup_{k=1}^{p}\, \beta_{jk} - \tau_j \neq 0$ using the statistic $S^{\ell}_{M-PH} = \sum_{j=1}^{m} Y_j$. It follows that $\mathrm{E}\left(S^{\ell}_{M-PH} \middle| H_0\right) = mp$ and it remains to find $\mathrm{Var}\left(S^{\ell}_{M-PH} \middle| H_0\right)$. Consider first $\mathrm{Var}(Y_j + Y_i) = 4p\left(1 + \mathrm{Corr}\left[Y_i, Y_j\right]\right)$ which follows from properties of the chi-square distribution. Assuming that S^{ℓ}_{M-PH} will follow a Gamma distribution, since it was shown that GenT uses a statistic which almost exactly does, and S^{ℓ}_{M-PH} is an extension of it, we first consider the covariance between SNP-level ANOVA chi-square statistics for SNPs j and s:

$$\begin{aligned}\mathrm{Cov}(Y_j, Y_s) &= \mathrm{Cov}\left(Z_{j1}^2 + \cdots + Z_{jp}^2, Z_{s1}^2 + \cdots + Z_{sp}^2\right) \\ &= \mathrm{Cov}\left(Z_{j1}^2, Z_{s1}^2\right) + \cdots + \mathrm{Cov}\left(Z_{jp}^2, Z_{sp}^2\right)\end{aligned}$$

since the populations (second indices) are independent of each other. Simulation shows that $\mathrm{Corr}([Y_1, \dots, Y_m]^{\top}) \coloneqq \mathbf{\Pi} = p^{-1} \sum_{k=1}^{p} \mathbf{\Sigma}_k$ where $\mathbf{\Sigma}_k = (r_{nk}^2)_{n=1}^{m}$ is the $m \times m$ matrix of squared LD correlations in the kth population. It follows that $\mathrm{Cov}([Y_1, \dots, Y_m]^{\top}) = \mathbf{D\Pi D}$ where $\mathbf{D} = \mathrm{diag}\left(\sqrt{2p}\right)$ and $v^{\ell} = \mathbf{1}_m^{\top}\mathbf{D\Pi D}\mathbf{1}_m = 2p\mathbf{1}_m^{\top}\mathbf{\Pi}\mathbf{1}_m$. Our use of simulation to find $\mathbf{\Pi}$ was for simplicity and extensive simulation results demonstrate that this variance correctly parameterizes the null distribution of S^{ℓ}_{M-PH} as shown in **Figure S7**. The parameters of the null distribution $\mathrm{Gamma}(\alpha, \xi)$ are:

$$\xi = \frac{1}{2}\frac{m}{\mathrm{tr}(\mathbf{\Pi\Pi})}, \qquad \alpha = \frac{m^2 p}{2}\frac{1}{\mathrm{tr}(\mathbf{\Pi\Pi})}.$$

Figure S7: Demonstrating the correct parameterization of the null distribution of the MuGenT-PH test statistic



Each panel displays the distribution of simulated test statistics under the MuGenT-PH null hypothesis (black) and the theoretical density from the distribution parameterized using the derivation above (blue) for varying numbers of tested SNPs, GWAS sample sizes, and densities of the first-order autoregressive LD matrix. Also displayed are the Type I error rates at the 5% significance level. In all cases we generated GWAS summary statistics directly using the procedure described in the Methods section of the main text and R code from github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/mugent_ph/mugent_ph_distribution.R.

S5. Setting the MuGenT-Pleio quantile to achieve target Type I error

In the main text description of MuGenT-Pleio, we stated that

$$P\left[\bigcup_{j=1}^{m}\left(\bigcap_{k=1}^{p}\beta_{jk}^{\ell}\in\mathcal{R}_{jk}\right)|H_0^{\ell}\right]\approx 1-(1-\alpha^p)^{2m_{eff.}}:=\tilde{\alpha} \tag{4.1}$$

where $m_{eff.}$ is the so-called effective number of independent SNPs (see below), p is the number of populations, β_{jk}^{ℓ} is the association between the jth SNP and GWAS phenotype in the kth population of the ℓth gene, and $\mathcal{R}_{jk}$ is the parameter space in which $H_{0jk}^{\ell}\colon\beta_{jk}^{\ell}=0$ is rejected based on significance level α. First, we provide the definition for $m_{eff.}$ from Jiang et al. (2022). Let $\mathbf{A}\in\mathbb{R}^{n\times n}$ be a positive semidefinite LD matrix with eigenvalues $(\lambda_1,\dots,\lambda_n)$ of arbitrary order and redundancy. Jiang et al. (2022) show that

$$m_{eff.}=\sum_{i=1}^{n}I(\lambda_i\geq 1)+\lambda_i I(\lambda_i<1)$$

is the approximate number of independent SNPs where the expression $I(a)$ equals 1 if the condition a is satisfied and 0 otherwise. To observe its application to our analyses, consider two extreme cases such that $\mathbf{A}=\mathbf{I}$ or $\mathbf{A}=(1)$, the identity matrix or an $n\times n$ matrix of 1s, respectively. $m_{eff.}$ in these two cases will respectively equal n and 1.

We stated that in practice $\tilde{\alpha}$ can either be smaller or greater than the nominal significance level based on $m_{eff.}$ and p since MuGenT-Pleio performs an intersection-union test. We now find a new $\ddot{\mathcal{R}}_{jk}$ which satisfies

$$P\left[\bigcup_{j=1}^{m}\left(\bigcap_{k=1}^{p}\beta_{jk}^{\ell}\in\ddot{\mathcal{R}}\right)|H_0^{\ell}\right]=\psi.$$

We begin by stating the equality

$$\begin{aligned}P\left[\bigcup_{j=1}^{m}\left(\bigcap_{k=1}^{p}\beta_{jk}^{\ell}\in\ddot{\mathcal{R}}\right)|H_0^{\ell}\right]&=1-P\left(\bigcap_{j=1}^{m}\left[\bigcup_{k=1}^{p}\beta_{jk}\notin\ddot{\mathcal{R}}\right]|H_0^{\ell}\right)\\&\approx 1-P\left(\bigcup_{k=1}^{p}\beta_{jk}\notin\ddot{\mathcal{R}}|H_0^{\ell}\right)^{2m_{eff.}}\end{aligned}$$

$$= 1 - \left[1 - P\left(\bigcap_{k=1}^{p} \beta_{jk} \in \ddot{\mathcal{R}}\right) | H_0^\ell\right]^{2m_{eff.}}$$
$$= 1 - \left[1 - P\left(\beta_{jk} \in \ddot{\mathcal{R}} \middle| H_0^\ell\right)^p\right]^{2m_{eff.}}, \quad (4.2)$$

where justification for the approximation is made at the end of this section and assumes that $P\left(\beta_{kl}^\ell \in \ddot{\mathcal{R}} | H_0^\ell\right)$ is not proportional to β_{jk}^ℓ. Now we simply set (4.2) equal to ψ and solve for $\ddot{\mathcal{R}}$ as below:

$$\psi = 1 - \left[1 - P\left(\beta_{jk} \in \ddot{\mathcal{R}} \middle| H_0^\ell\right)^p\right]^{2m_{eff.}}$$
$$\Rightarrow (1-\psi)^{\frac{1}{2m_{eff.}}} = 1 - P\left(\beta_{jk} \in \ddot{\mathcal{R}} | H_0^\ell\right)^p$$
$$\Rightarrow \left[1 - (1-\psi)^{\frac{1}{2m_{eff.}}}\right]^{\frac{1}{p}} = P\left(\beta_{jk} \in \ddot{\mathcal{R}} \middle| H_0^\ell\right).$$

Since $P\left(\beta_{jk} \in \ddot{\mathcal{R}} \middle| H_0^\ell\right) = P\left(Z_{jk}^2 > S^2 \middle| H_0^\ell\right) \coloneqq 1 - F_{\chi_1^2}(S^2)$ where $Z_{jk}^2 \sim \text{Normal}(0,1)$ under H_{0jk}^ℓ and S^2 is a quantile yet computed, it follows that $\ddot{\mathcal{R}}$ is defined by S^2 which is equal to:

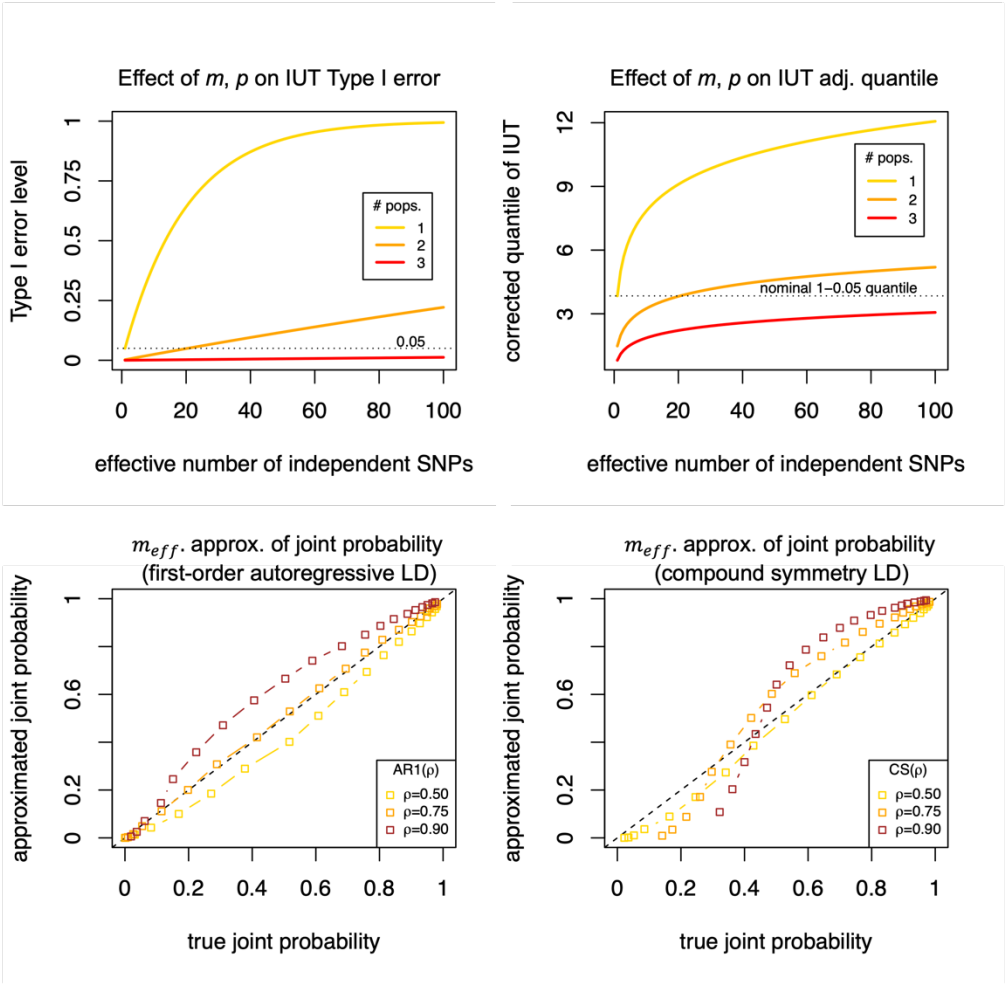
$$F_{\chi_1^2}^{-1}\left\{1 - \left[1 - (1-\psi)^{\frac{1}{2m_{eff.}}}\right]^{\frac{1}{p}}\right\},$$

which is what was shown in the main text. The quantile S^2 can be less than or greater than the nominal $(1-\alpha)$ χ_1^2 quantile corresponding to significance level α based on the relationship between $m_{eff.}$ and p. In practice, this translates to adjusting the nominal significance level α of the original tests, up or down depending, defining $\ddot{\mathcal{R}}$ such that overall level ψ is eventually achieved by MuGenT-Pleio.

We now show that for a set of non-independent SNPs that $2m_{eff.}$ can be used to approximate their joint probability. We are forced to approximate this quantity because the m-part integral cannot be reliably approximated using current statistical packages (e.g., mvtnorm in R; Genz & Bretz, 2009) for numbers of SNPs greater than 10-20, which in our context would correspond to 10-20 SNPs. These results are displayed in **Figure S8** and demonstrate that the approximation of the joint probability is relatively precise for first-order autoregressive LD structures, especially when the true joint probability is large or small. For compound symmetry LD structures, the approximation of the joint probability is most precise when the true joint probability is approximately 0.5. **Figure S9** displays the Type I error and power of MuGenT-Pleio using this approximation and demonstrates that MuGenT-Pleio has a 0% false positive rate regardless of GWAS sample size and the number of populations, and that increasing

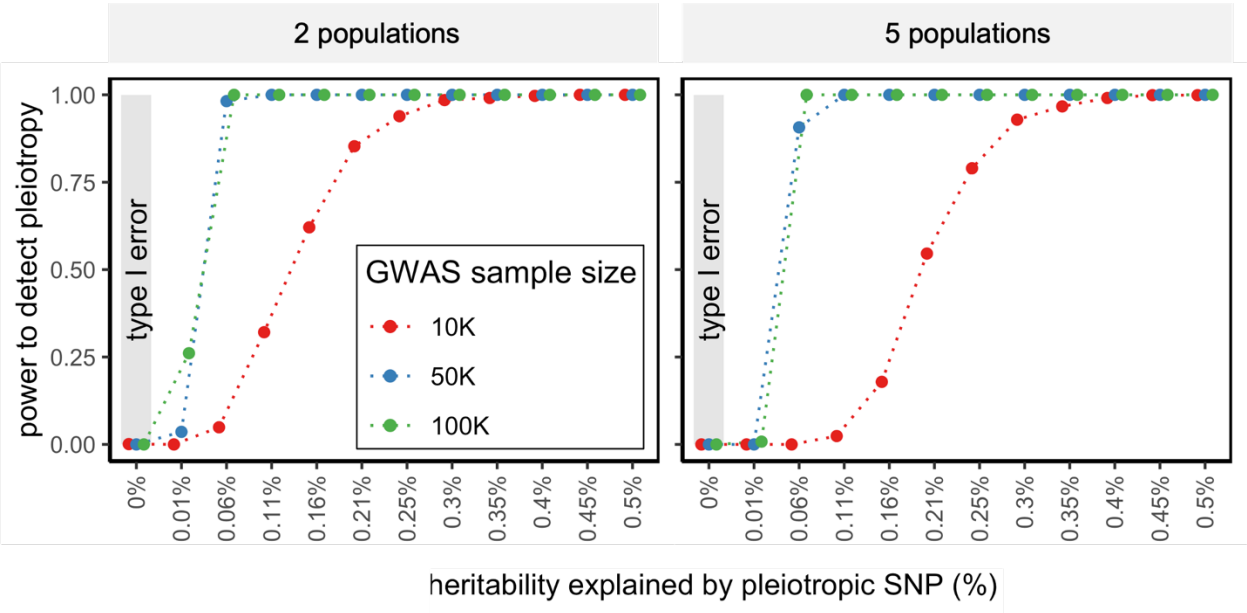
GWAS sample sizes and decreasing numbers of testing populations lead to greater power of MuGenT-Pleio.

Figure S8: Role of $m_{eff.}$ approximation when parameterizing the null distribution of MuGenT-Pleio



The top-left panel displays the relationship between the number of independent SNPs, number of populations, and Type I error rate of the intersection-union test (IUT) of the null hypothesis of no joint association across all populations. The top-right panel displays the correction that must be applied to the critical region of the IUT to maintain a fixed 5% Type I error rate. The bottom-left and bottom-right panels display the approximation of a generic joint probability that 30 correlated chi-square statistics are greater than threshold τ which varied from 1×10^{-6} to 0.1. The true probability of this event was determined using the proportion of 1,000 simulations for which the condition was satisfied. The approximated joint probability was estimated using $m_{eff.}$ calculated from the first-order autoregressive (left) or compound symmetry (right) covariance matrix of the corresponding Z-statistics with correlation parameter ρ. R code used to produce these figures is available from github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/mugent_pleio/mugent_pleio_type1_of_IUT_demo.R (top two figures) and github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/mugent_pleio/meff_vs_true_joint_probability.R (bottom two figures).

Figure S9: Type I error and power of MuGenT-Pleio

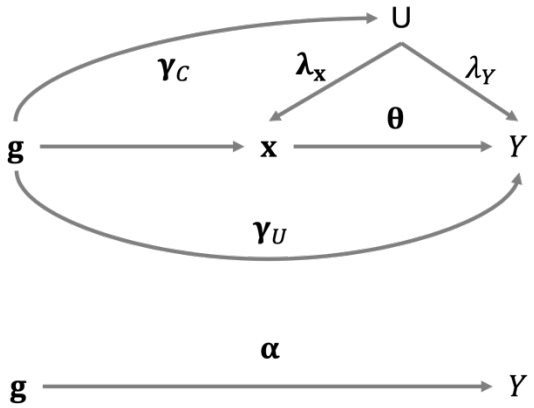


Displayed are false positive rates ('type I error') and power of the MuGenT-Pleio test performed on simulation GWAS summary statistics from 2 and 5 populations of varying sample size with explained heritability from a single population-shared causal SNP ranging from 0% to 0.5%. R code used to produce these figures is available from github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/mugent_pleio/mugent_pleio_simulation.R.

S6. Connection between xGenT and Mendelian Randomization

We briefly stated in the main text that when the assumptions of Mendelian Randomization (MR) hold, that the expected value of the test statistic used by xGenT is proportional to the sum of total causal effects of the corresponding gene products (e.g., expression, protein abundance) on the disease phenotype. Let the notation of the main text hold such that $\widehat{\boldsymbol{\beta}}^{\ell} = \left(\hat{\beta}_1^{\ell}, \ldots, \hat{\beta}_{m^{\ell}}^{\ell}\right)^{\mathsf{T}}$ and $\hat{\boldsymbol{\nu}}_k^{\ell} = \left(\hat{\nu}_{1k}^{\ell}, \ldots, \hat{\nu}_{m^{\ell}k}^{\ell}\right)^{\mathsf{T}}$ represent the m^{ℓ}-length vectors of estimated marginal associations between the m SNPs corresponding to the ℓth gene and the disease phenotype and the kth gene product, respectively. We begin by stating the structural equations representing the full set of possible relationships between SNP genotypes for the ℓth gene, the gene product, the disease phenotype, and any confounders which are represented by the DAG in **Figure S10**.

Figure S10: Causal diagram connecting xGenT and Mendelian Randomization



Key

$\mathbf{g}$: $m\times1$ vector of SNP genotypes for a gene
$\mathbf{x}$: $p\times1$ vector of gene product phenotypes (e.g., expression, protein abundance)
U: generic confounder
Y: disease phenotype
$\boldsymbol{\gamma}_C$: $m\times1$ vector of correlated horizontal pleiotropy in Mendelian Randomization
$\boldsymbol{\gamma}_U$: $m\times1$ vector of uncorrelated horizontal pleiotropy in Mendelian Randomization
$\boldsymbol{\theta}$: $p\times1$ vector of causal effects of $\mathbf{x}$ on Y
$\boldsymbol{\alpha}$: $m\times1$ vector of total associations between $\mathbf{g}$ and Y
$\boldsymbol{\lambda}_\mathbf{x}$: $p\times1$ vector of effect of confounder on gene product phenotypes
λ_Y: effect of confounder on disease phenotype

The structural equations are

$$\begin{aligned}\mathbf{x} &= \mathbf{B}^\top\mathbf{g} + \boldsymbol{\lambda}_\mathbf{x}U + \boldsymbol{\epsilon}_\mathbf{x}\\ &= (\mathbf{B}^\top + \boldsymbol{\lambda}_\mathbf{x}\boldsymbol{\gamma}_C^\top)\mathbf{g} + \boldsymbol{\lambda}_\mathbf{x}\epsilon_U + \boldsymbol{\epsilon}_\mathbf{x},\\ Y &= (\boldsymbol{\theta}^\top\mathbf{B}^\top + \boldsymbol{\theta}^\top\boldsymbol{\lambda}_\mathbf{x}\boldsymbol{\gamma}_C^\top)\mathbf{g} + \boldsymbol{\theta}^\top\boldsymbol{\lambda}_\mathbf{x}\epsilon_U + \boldsymbol{\theta}^\top\boldsymbol{\epsilon}_\mathbf{x} + \lambda_Y U + \boldsymbol{\gamma}_U^\top\mathbf{g} + \epsilon_Y\\ \Rightarrow \boldsymbol{\alpha} &= \mathbf{B}\boldsymbol{\theta} + \boldsymbol{\gamma}_C\boldsymbol{\lambda}_\mathbf{x}^\top\boldsymbol{\theta} + \boldsymbol{\gamma}_U.\end{aligned}$$

The goal of MR is to estimate the causal effects $\boldsymbol{\theta}$ using estimates of $\boldsymbol{\alpha}$ and $\mathbf{B} = \left(\boldsymbol{\beta}_j\right)_{j=1}^m$, but these causal estimates can be severely biased, and indeed the wrong inference can be made at greater than nominal rates, when there is uncorrelated horizontal pleiotropy ($\boldsymbol{\gamma}_U$) or correlated horizontal pleiotropy ($\boldsymbol{\lambda}_\mathbf{x}\boldsymbol{\gamma}_C$). There is no need to distinguish marginal from joint association estimates from the standpoint of estimating $\boldsymbol{\theta}$, and so we assume throughout that the parameters $\mathbf{B}$ represent marginal association estimates. The above models imply $\alpha_j = \boldsymbol{\beta}_j^\top\boldsymbol{\theta}$ for the jth SNP, and that $\boldsymbol{\alpha} = \boldsymbol{\beta}_k\theta_k$ for the kth gene product phenotype, of which there are p. In this context, and where $\mathbf{z} = \left[\frac{\widehat{\alpha}_j}{\mathrm{SE}(\widehat{\alpha}_j)}\right]$ our xGenT gene-based test statistic is proportional to $\mathbf{z}^\top\mathbf{L}\mathbf{z}$ where $\mathbf{L} = p^{-1}\sum_{k=1}^p \boldsymbol{\beta}_k\boldsymbol{\beta}_k^\top = \left(\sum_{k=1}^p \beta_{jk}\right) = \left(\sum_{k=1}^p \alpha_{jk}\,\theta_k\right)$. This implies that in the absence of uncorrelated and correlated horizontal pleiotropy, $\mathbf{z}^\top\mathbf{L}\mathbf{z}$ is proportional to the total causal effects of the gene product phenotypes on the disease phenotype. This was demonstrated in Figure 4a in the main text which showed that the xGenT null hypothesis is only rejected at rates exceeding its

nominal significance threshold when there is shared heritability among the gene product and disease phenotypes, which implies a causal effect if the SNPs explaining the heritability have correlated effect sizes according to the principles under Mendelian Randomization.

S7. De-correlating GTEx Z-statistics from sample overlap

We used GTEx summary data to estimate genetic correlations between gene expression in 49 different tissues, but many individuals in GTEx v8 (2020) contributed multiple tissues and so Z-statistics within SNPs and between tissues are correlated. We removed this correlation using the procedure described in LeBlanc et al. (2018) and Lorincz-Comi et al. (2024). We briefly describe this procedure here. Let $\mathbf{z}_j = (Z_s)$ represent the length 49 vector of Z-statistics of marginal association for the jth SNP and each of the 49 GTEx v8 tissues. LeBlanc et al. showed that Z_s and Z_k have covariance ψ which is proportional to the number of individuals present in both cohorts. Lorincz-Comi et al. showed that ψ can be estimated empirically from non-significant (P>0.05) GWAS summary statistics from both cohorts, and LeBlanc et al. showed that Z_s and Z_k can be de-correlated by pre-multiplying the vector $(Z_s, Z_k)^\top$ by $\mathbf{H}^{-1/2}$ where $\mathbf{H} = \mathbf{I} + \mathbf{F}$ and $\mathbf{F}$ only has off-diagonal elements equal to ψ. We applied this procedure using the full vector $\mathbf{z}_j$ and estimated the corresponding matrix $\mathbf{H}$ using a random sample of 11,854 non-significant (P>0.05) SNPs on chromosome 1 taken uniformly from each SNP-gene pair. That is, GTEx v8 summary data provide estimated associations between SNP-gene pairs for multiple SNPs and multiple genes. From each of the genes on chromosome 1, we stored the estimated associations with gene expression in each tissue for 10 randomly selected non-significant SNP-gene pairs. We then used the subset of SNPs for which SNP-gene association estimates were non-significant for all tissues and used them to estimate $\mathbf{H}$. The full code to perform this task is available at github.com/noahlorinczcomi/gent_analysis/blob/main/data_analysis_scripts/making_gtex_sample_overlap_correlation_matrix.r.

S8. Gene clumping and correlation between GenT statistics

S8.1 Gene clumping

We described in the Methods section that users of our software have the option, when performing gene-based association testing genome-wide, to adjust for inflation of test statistics genome-wide either numerically or analytically. We first introduce numerical adjustment. Let $\mathbf{s} = (S_1, \dots, S_i, \dots, S_M)^\top$ be the vector of statistics used to test associations between M genes and a disease phenotype using any of the gene-based tests we introduced. The null distributions of these statistics F_i are parameterized by elements in the shape vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_i, \dots, \alpha_M)^\top$ and rate vector $\boldsymbol{\xi} = (\xi_1, \dots, \xi_i, \dots, \xi_M)^\top$. Note that each distribution is parameterized by different sets of values which are determined completely by the local LD structure. We find the expected 0.5 quantile of each distribution as $\tau_i = F_i^{-1}(0.5; \alpha_i, \xi_i)$ and calculate the observed genomic control inflation

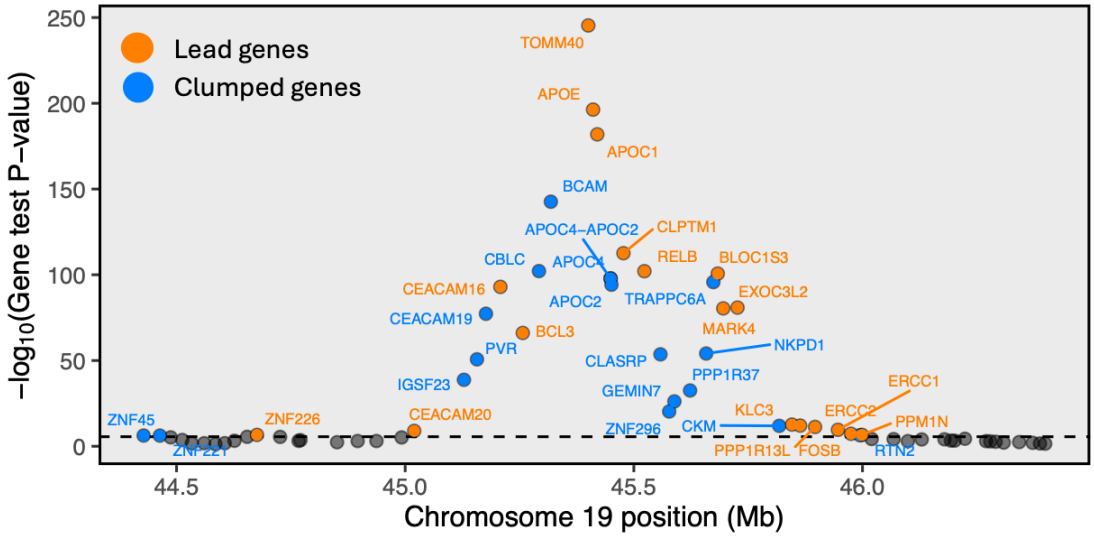
factor of $\lambda = \text{median}\left(\frac{S_i}{\tau_i}\right)$ for $i = 1, \ldots, M$. We then return to users the inflation-adjusted test statistics $\ddot{S}_i = S_i/\lambda$.

We now describe an approach which can be used to prioritize a subset of genes in a large set which share correlated GenT test statistics. This procedure is modelled after the `clump` procedure in PLINK (Chang et al., 2015) and follows these steps for each chromosome separately using P-values from either GenT or MuGenT:

1. **Require**: *d*, significance threshold of gene-based association test; *k*, megabase window size in which genes will be clumped; r_{js}, correlation of GenT or MuGenT test statistics for genes *j* and *s*.
2. **Initialize**: Create empty sets *A, C, L*; add all genes with P-value>*d* to *A*
3. **Iterate**: Repeat steps 3a-3e until all genes on the chromosome are in *A*
    3a. Select all genes not in set *A* and place in set *V*
    3b. Select the top gene from this set, i.e. the one with the smallest P-value
    3c. Filter genes in *V* to those within *k*Mb of the midpoint of the top gene which have a squared GenT test statistic correlation with the top gene that is greater than r^2, excluding the top gene
    3d. Add remaining genes from the filtered set of *V* to sets *C, A*
    3e. Add top gene to sets *L, A*
4. **Return**: *L*, list of lead genes; *C*, list of 'clumped' genes, which are significant at level *d* but not lead genes in their locus.

The `gene_clump()` function in the `gent` R package performs this automatically for users using pre-computing chromosome-specific matrices of correlations between GenT and MuGenT test statistics stored at https://github.com/noahlorinczcomi/gent. **Figure S11** displays an example of how this procedure prioritizes AD risk genes in the *APOE* locus using GenT test statistics, providing evidence that *APOE*, *APOC1*, and *TOMM40* each have independent associations with AD risk, which those of *BCAM*, *APOC2*, and *APOC4* are 'clumped' to one of these more strongly associated genes.

Figure S11: Applying the gene clumping procedure to GenT AD associations in the *APOE* locus identifies LD-independent gene associations



Displayed are gene-based association test P-values for genes in the APOE locus and association with AD. 'Lead' genes are the most significant genes not correlation with any other sin the locus beyond level 0.01. 'Clumped genes' are significant, but a more significant gene is present in the locus with which the gene is corelated beyond level 0.01. Genes with P<0.05/18257 are colored in gray.

S8.2 Correlation between GenT statistics due to shared LD

We stated earlier that the correlation between GenT/MuGenT test statistics is required to perform the gene clumping procedure. It is also required to perform gene-based fine-mapping which was mentioned in the Methods section of the main text as well as **Supplement Section S9**. We begin by deriving the correlation between GenT test statistics for genes ℓ and k, respectively denoted S_ℓ and S_k which test $H_0^\ell: \mathrm{E}(Z_{i\ell}) = 0$ for $i = 1, \ldots, m_\ell$ SNPs in the set $\mathcal{L}_\ell$ and $H_0^k: \mathrm{E}(Z_{ik}) = 0$ for $i = 1, \ldots, m_k$ SNPs in the set $\mathcal{L}_k$ where $(Z_{i\ell}, Z_{ik})$ are Z-statistics from same GWAS. Under H_0^ℓ and H_0^k,

$$\bar{\mathbf{z}} \coloneqq \begin{pmatrix} \mathbf{z}_\ell \\ \mathbf{z}_k \end{pmatrix} \sim \mathrm{Normal}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \boldsymbol{\Sigma} \coloneqq \begin{bmatrix} \mathbf{R}_\ell & \mathbf{R}_{\ell k}^\top \\ \mathbf{R}_{\ell k} & \mathbf{R}_k \end{bmatrix} \right),$$

where $\boldsymbol{\Sigma}$ is the matrix of LD correlations between SNPs in the union of SNP sets $\mathcal{L}_\ell$ and $\mathcal{L}_k$. Let $\mathbf{F}_\ell = (\mathbf{I}_{m_\ell}, \mathbf{0})$ and $\mathbf{F}_k = (\mathbf{0}, \mathbf{I}_{m_k})$ where $\mathbf{I}_o$ is an identity matrix of size o. We aim to compute $\mathrm{Cov}(S_\ell, S_k)$, which is equivalent to $\mathrm{Cov}(\bar{\mathbf{z}}^\top \mathbf{F}_\ell^\top \mathbf{F}_\ell \bar{\mathbf{z}}, \bar{\mathbf{z}}^\top \mathbf{F}_k^\top \mathbf{F}_k \bar{\mathbf{z}})$. By the properties of the multivariate normal distribution under H_0^ℓ and H_0^k,

$$\mathrm{Cov}(\bar{\mathbf{z}}^\top \mathbf{F}_\ell^\top \mathbf{F}_\ell \bar{\mathbf{z}}, \bar{\mathbf{z}}^\top \mathbf{F}_k^\top \mathbf{F}_k \bar{\mathbf{z}}) = 2\mathrm{tr}(\mathbf{F}_\ell^\top \mathbf{F}_\ell \boldsymbol{\Sigma} \mathbf{F}_k^\top \mathbf{F}_k \boldsymbol{\Sigma}).$$

We can first see that

$$\mathbf{F}_\ell^\top \mathbf{F}_\ell \boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{I}_{m_\ell} \\ \mathbf{0} \end{pmatrix} (\mathbf{I}_{m_\ell} \quad \mathbf{0}) \boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{I}_{m_\ell} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{R}_\ell & \mathbf{R}_{\ell k}^\top \\ \mathbf{R}_{\ell k} & \mathbf{R}_k \end{pmatrix} = \begin{pmatrix} \mathbf{R}_\ell & \mathbf{R}_{\ell k}^\top \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and

$$\mathbf{F}_k^\top \mathbf{F}_k \boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_{m_k} \end{pmatrix} (\mathbf{0} \quad \mathbf{I}_{m_k}) \boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m_k} \end{pmatrix} \begin{pmatrix} \mathbf{R}_\ell & \mathbf{R}_{\ell k}^\top \\ \mathbf{R}_{\ell k} & \mathbf{R}_k \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{R}_{\ell k} & \mathbf{R}_k \end{pmatrix}.$$

It follows that

$$\begin{aligned} 2\mathrm{tr}(\mathbf{F}_\ell^\top \mathbf{F}_\ell \boldsymbol{\Sigma} \mathbf{F}_k^\top \mathbf{F}_k \boldsymbol{\Sigma}) &= 2\mathrm{tr} \begin{pmatrix} \mathbf{R}_{\ell k}^\top \mathbf{R}_{\ell k} & \mathbf{R}_{\ell k}^\top \mathbf{R}_k \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &= 2\mathrm{tr}(\mathbf{R}_{\ell k}^\top \mathbf{R}_{\ell k}). \end{aligned}$$

These results imply that under H_0^ℓ and H_0^k,

$$\mathrm{Corr}(S_\ell, S_k; \mathrm{GenT}) = \frac{\mathrm{tr}(\mathbf{R}_{\ell k}^\top \mathbf{R}_{\ell k})}{\sqrt{\mathrm{tr}(\mathbf{R}_\ell \mathbf{R}_\ell)\mathrm{tr}(\mathbf{R}_k \mathbf{R}_k)}}.$$

This result demonstrates that the magnitude of correlation between GenT test statistics from two genes is strictly positive and is proportional to the LD between their SNPs. To estimate the correlation between MuGenT test statistics S_k and S_ℓ for genes k and ℓ, we follow the derivation of Supplementary Section S4 and the GenT result from above to produce the following estimator

$$\widehat{\mathrm{Corr}}(S_\ell, S_k; \mathrm{MuGenT}) = \frac{\mathrm{tr}(\boldsymbol{\Pi}_{\ell k}^\top \boldsymbol{\Pi}_{\ell k})}{\sqrt{\mathrm{tr}(\boldsymbol{\Pi}_\ell \boldsymbol{\Pi}_\ell)\mathrm{tr}(\boldsymbol{\Pi}_k \boldsymbol{\Pi}_k)}}$$

where $\boldsymbol{\Pi}_\ell = p^{-1} \sum_{i=1}^p \mathbf{R}_k^i$, $\boldsymbol{\Pi}_k = p^{-1} \sum_{i=1}^p \mathbf{R}_k^i$, and $\boldsymbol{\Pi}_{\ell k} = p^{-1} \sum_{i=1}^p \mathbf{R}_{\ell k}^i$ where $\mathbf{R}_{\ell k}^i$ is a matrix of LD correlations between the ℓth and kth genes in the ith ancestral population, of which there are p. If MuGenT is used for multi-trait analyses in a single ancestral population, $\boldsymbol{\Pi}_\ell = \mathbf{R}_\ell^i$, $\boldsymbol{\Pi}_k = \mathbf{R}_k^i$, and $\boldsymbol{\Pi}_{\ell k} = \mathbf{R}_{\ell k}^i$.

S9 Fine-mapping gene-based test statistics

We stated in the Methods section that we can perform fine-mapping of gene-based test statistics if they are transformed to be asymptotically normal. In **Subsection 9.1**, we provide the motivation for this transformation and show that its limiting distribution is multivariate normal. In **Subsection S9.2**, we perform simulations to assess the performance of using SuSiE (Zou et al., 2020) to perform fine-mapping.

S9.1 Asymptotic distribution of transformed statistics

In a single locus, multiple genes exist. The gene-based test statistics calculated from these genes may be correlated if the gene-specific SNP sets share SNPs which are in LD with each other (see **Section S8**). Let S_k represent the gene-based test statistic for the kth gene in this locus, $\mathbf{s} = (S_k)_{k=1}^K$ be the vector of them, and $\boldsymbol{\Sigma} = \mathrm{Corr}(\mathbf{s})$. We have

already stated that $S_k \sim \Gamma(\alpha_k, \xi_k)$ with high accuracy under H_{0k}:$\cap_{j=1}^{m_k} \beta_{jk} = 0$ where β_{jk} is the GWAS effect size for the kth gene and jth SNP in its set, of which there are m_k. Recall $S_k = \sum_{j=1}^{m_k} \hat{\beta}_{jk}^2 s_{jk}^{-2}$ where $s_{jk}^2 = \mathrm{Var}(\hat{\beta}_{jk})$, i.e., S_k is the sum of SNP chi-square statistics for SNPs in the kth gene's SNP set. Recall that $\mathrm{Corr}(\hat{\beta}_{1k}, \ldots, \hat{\beta}_{m_k k}) = \mathbf{R}_k$ and that S_k is distributionally equivalent to $\sum_{j=1}^{m_k} \lambda_{kj} Y_j$ where $Y_j \sim \chi^2(1)$, $Y_j \perp Y_i$, and $\mathbf{R}_k = \mathbf{U}_k \mathrm{diag}(\lambda_{kj}) \mathbf{U}_k^\top$. We derive the limiting distribution of m_k^{-1} under the distribution equivalence $S_k \sim \sum_{j=1}^{m_k} \lambda_{kj} Y_j$:

$$\sqrt{m_k}\left(\frac{1}{m_k}\sum_{j=1}^{m_k}\lambda_{kj}Y_j - \frac{1}{m_k}\sum_{j=1}^{m_k}\mathrm{E}(\lambda_{kj}Y_j|H_{0k})\right) \xrightarrow{D} N\left(0, \frac{1}{m_k}\sum_{j=1}^{m_k}\mathrm{Var}(\lambda_{kj}Y_j|H_{0k})\right).$$

We can rewrite this quantity using the properties of the chi-square distribution as

$$\sqrt{m_k}\left(\frac{1}{m_k}S_k - 1\right) \xrightarrow{D} N\left(0, \frac{1}{m_k} 2\mathrm{tr}(\mathbf{R}_k\mathbf{R}_k)\right)$$

which can be further rewritten as

$$\frac{1}{\sqrt{2\mathrm{tr}(\mathbf{R}_k\mathbf{R}_k)}}(S_k - m_k) \xrightarrow{D} N(0,1).$$

The final expression is what we presented in the main text. It follows that the correlation between S_k and S_r will be the same as that between $m_k^{-1}S_k$ and $m_r^{-1}S_r$ so that we can use the matrix of correlations from **Section S8** above. Under H_{1k}: $\cup_{j=1}^{m_k} \beta_{jk} \neq 0$, $m_k^{-1}S_k$ will have mean greater than m_k and variance larger than $2\mathrm{tr}(\mathbf{R}_k\mathbf{R}_k)$. Future adaptations of GenT and its asymptotic transformation presented above may consider a variance-stabilizing transformation such that the asymptotic mean parameter under H_{1k} has constant variance.

S9.2 Simulations

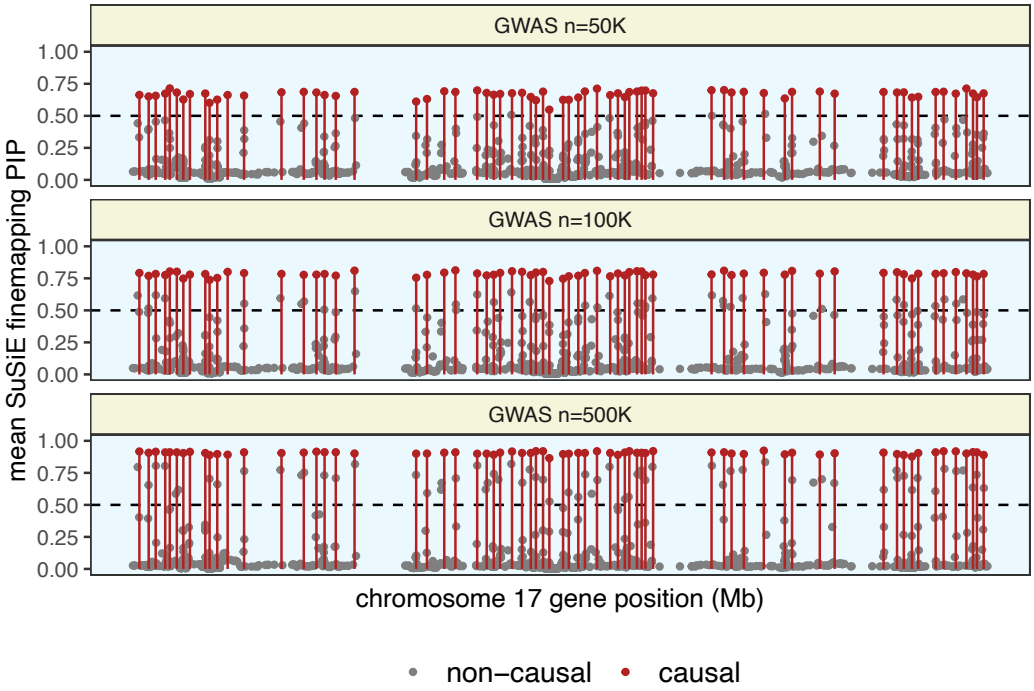
We performed simulations in which we simulated gene-based test statistics for genes from empirically-defined uncorrelated blocks under a model which was causal for 2 genes in a locus and non-causal for the rest then performed fine-mapping of genes using SuSiE (Zou et al., 2022). We first calculated correlations between gene-based test statistics for all genes on chromosome 17 using SNPs in the 1000 Genomes Phase 3 European reference panel and defined independent blocks of them using the correlation block-sorting method of Prive (2022) in the `bigsnpr` R package (https://github.com/privefl/bigsnpr). The arguments we used were: `bigsnpr::snp_ldsplit(…,thr_r2=0.3, min_size=10, max_size=300, max_K=1e6, max_r2=0.3)`. In each uncorrelated block, we used the correlation matrix $\boldsymbol{\Sigma}$ calculated using the method in **Section S8** and the following data-generating model to generate correlated gene-based test statistics

$$\mathbf{z} \sim N(\mathbf{0}, \mathbf{\Sigma}), \qquad \mathbf{z} = \left(Z_{\mathrm{j}}\right)$$
$$p_j = \Phi\left(Z_j\right)$$
$$S_j = F_{\Gamma}^{-1}\left(p_j; \alpha_j, \xi_j\right)$$
$$\ddot{S}_j = \frac{S_j - \alpha_j \xi_j^{-1}}{\alpha_j \xi_j^{-2}}$$

where Φ is the standard normal cumulative density function and $F_{\Gamma}^{-1}\left(\cdot\,; \alpha_j, \xi_j\right)$ is the quantile function of a Gamma distribution with shape parameter α_j and rate parameter ξ_j for the jth gene. We performed fine-mapping on the set of statistics $\left\{\ddot{S}_j\right\}$, which are asymptotically normal under the derivation above. We fixed α_j and ξ_j for non-causal genes at $\xi_j \approx 0.1406255$ and $\alpha_j \approx 52.03143$, which were calculated assuming 370 SNPs with first-order autoregressive LD structure with a correlation of 0.75 between neighboring SNPs. We chose 370 because this was the median size of gene-specific SNP sets in our real data analysis with Alzheimer's disease. We defined α_j and ξ_j for each of the two causal genes using a reduced form of the random effect model of Lorincz-Comi et al. (2025) without minor allele frequency and LD scores. For causal genes, $\xi_j \approx (370 + N\tau)/(34381.381 + 2N^2\tau^2 + 4N\tau)$ and $\alpha_j \approx (200 + N\tau)\xi_j$ where N is the GWAS sample size and $\tau = 0.005/2$, i.e., the SNP heritability explained by each causal gene divided by the number of causal SNPs in the gene-specific SNP set. We simulated a single causal gene in each of the 67 loci on chromosome 17 we analyzed and were always set to be nearest to the center of the locus in base pair position. Note that in this simulation for demonstrative purposes each locus was considered as independent and contained a single causal gene. Our goal was to apply fine-mapping to the transformed gene-based association test statistics to prioritize each causal gene. The full R code used to perform the simulations and create the figures presented below is available at https://github.com/noahlorinczcomi/gent_analysis/simulations/finemapping.

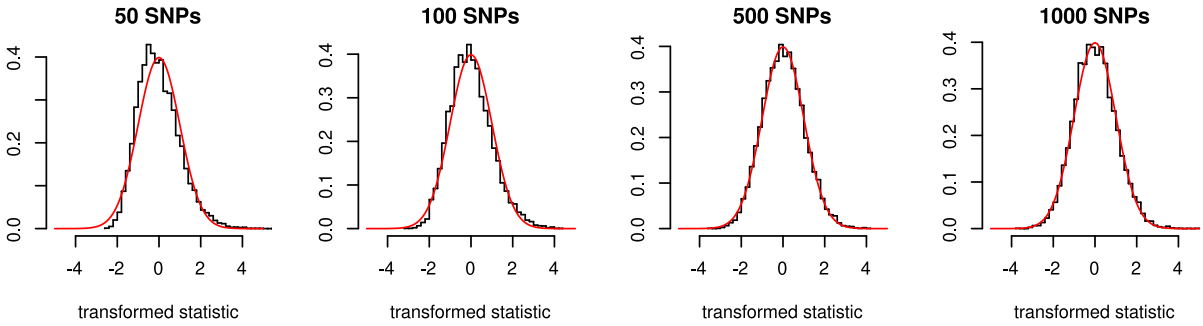
Figure S12 demonstrates that causal genes on average have larger posterior causal probabilities of causality (PIPs) under the SuSiE model than non-causal genes. Larger GWAS sample sizes generally increased PIPs for all genes, not just causal genes, owed partly we hypothesize to the increase in variance of the gene-based test statistics for causal genes compared to non-causal genes. On average, causal genes had the largest PIPs in each locus, demonstrating the utility of applying SuSiE fine-mapping to the transformed gene-based test statistics.

Figure S12: Simulation results for fine-mapped gene-based test statistics



Each point is a single gene. Correlations between gene-based test statistics were set by using a multivariate normal copula and empirical matrices of correlations between gene-based test statistics on chromosome 17 using approximately uncorrelated blocks. Full R code used to perform this simulation is available at https://github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/finemap/chr17_simulation.R

Figure S13: Asymptotic convergence of transformed gene-based test statistic



The x-axis is the transformed GenT test statistic generated under the null hypothesis with varying sizes of SNP sets containing SNPs with first-order autoregressive correlation structure and correlation parameter 0.5. Y-axes represent the empirical densities of the empirical distributions denoted by the black line histograms and red lines denote the true density of a standard normal distribution. These results suggest that as few as 50 SNPs are sufficient for close convergence of the transformed statistic to its limited distribution. under this LD structure. R code is available at https://github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/finemap/asymptotic_convergence.R

S10: Genetic correlation matrices

We estimated marginal genetic correlations between 18 complex traits and gene expression in 49 tissues and present the results in Supplementary Figures S19-S35. These genetic correlations are calculated as the pairwise Pearson correlations between Z-statistic vectors for a pair of traits. For example, let $(\mathbf{z}_1, \mathbf{z}_2)$ represent the $m \times 2$ matrix of Z-statistic vectors for m SNPs with LD matrix $\mathbf{R} = (\mathbf{G}\mathbf{G})^{-1}$ for traits 1 and 2. We estimated the genetic covariance between these traits as $(m-1)^{-1}\tilde{\mathbf{z}}_1^\top \mathbf{G}\mathbf{G}\tilde{\mathbf{z}}_2^\top$ where $\tilde{\mathbf{z}}_k = m^{-1}\mathbf{1}_m^\top \mathbf{G}\mathbf{z}_k$ for $k = 1{,}2$. We then converted this to a sample correlation. The matrix $\mathbf{R}$ was estimated from the European reference panel of 1000 Genomes Phase 3 and was updated as $\mathbf{R} \leftarrow 0.99\mathbf{R} + 0.01\mathbf{I}_m$.

We also estimated a sparse matrix of conditional cis-eQTL genetic correlations between gene expression in 49 different tissues in the region of the *RIPK2* gene and displayed the results in **Figure S53** (Supplementary Figures file) as a network. To produce this network, we began by extracting eQTL summary data for all SNPs within $\pm$50Kb of the start and end positions of *RIPK2* from GTEx v8. We then restricted this set of SNPs to just those present in the trans-ancestry 1000 Genomes Phase 3 reference panel, from which we estimated LD between the remaining m SNPs using PLINKv1.9. Where $\mathbf{Z}$ is the $m \times 49$ matrix of Z-statistics for the 49 tissues, we updated the raw LD matrix estimate $\mathbf{R}$ with $\mathbf{R} \leftarrow 0.99\mathbf{R} + 0.01\mathbf{I}$ and estimated genetic correlations using $\mathbf{Z}_f = \mathbf{G}\mathbf{Z}\mathbf{F}$ where $\mathbf{G}^\top\mathbf{G} = \mathbf{R}^{-1}$ and $\mathbf{F}$ is a matrix used to de-correlate GTEx Z-statistics due to sample overlap between tissues (see Section S7). We then applied graphical LASSO (Friedman et al., 2008) to $\mathbf{S} \coloneqq C\left(\mathbf{Z}_f^\top \mathbf{Z}_f\right)$ where C is a function to convert a sample covariance matrix to the corresponding sample rank correlation matrix. Graphical LASSO used the tuning parameter λ to achieve a sparse network of estimated conditional genetic correlations. We computed λ following Gao et al. (2012) by searching a grid of values which was found to minimize $-m\log\left(\left|\widehat{\boldsymbol{\Theta}}\right|\right) + m\mathrm{tr}\left(\widehat{\boldsymbol{\Theta}}\mathbf{S}\right) + 3\log(m)\, d_\lambda/2$ where $\widehat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \mathrm{tr}(\boldsymbol{\Theta}\mathbf{S}) - \log(|\boldsymbol{\Theta}|) + \lambda\|\boldsymbol{\Theta}\|_1$, d_λ is the number of nonzero elements in $\widehat{\boldsymbol{\Theta}}$, and m is the number of SNPs. Use of the constant 3/2 in the penalty imposed additional sparsity on the genetic correlation matrix.

S12 Window sizes of SNP-gene mappings

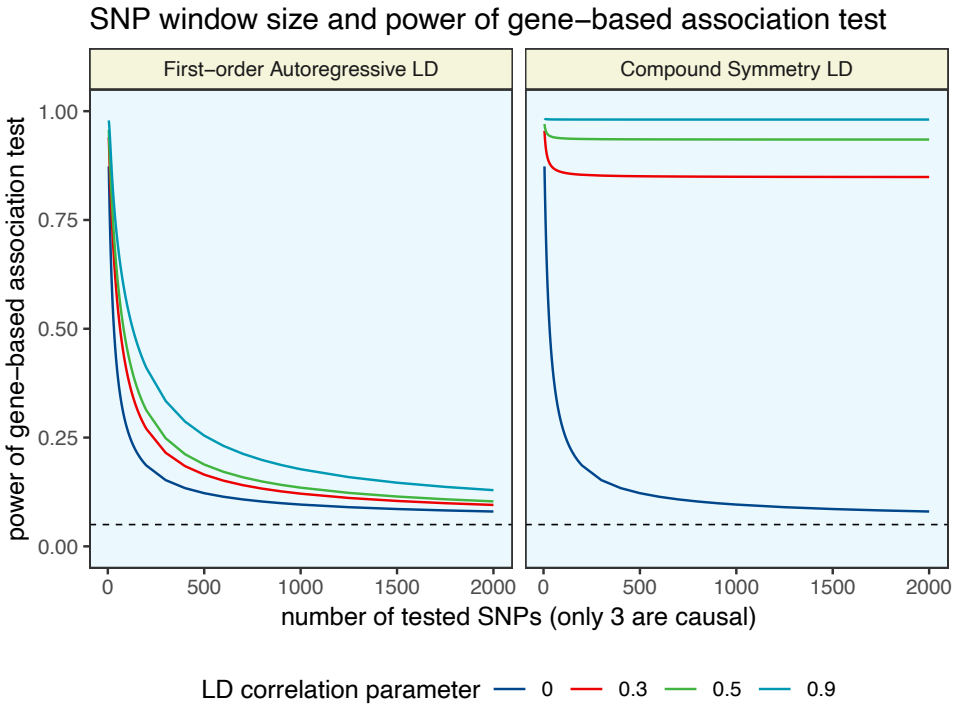
S12.1 Window size of gene-specific SNP sets and GenT power

In this section, we evaluated the extent to which the statistical power of the gene-based association test changes as the number of SNPs included in the gene-specific SNP set increases. In all scenarios, we set the GWAS sample size to 30K, SNP heritability explained by three causal SNPs in the gene-specific set at 0.05%, and set the causal SNPs to be those which are as close to the center of the median base pair position while being as far away from each other as possible. We varied the strength and pattern of LD among the SNPs in the gene-specific set and the number of non-causal SNPs included. LD patterns were either first-order autoregressive (AR1) or compound symmetry (CS). In the AR1 settings, SNPs were in LD with each other at a value which was a decaying function of their distance from each other. In the CS settings, all SNPs

in the gene-specific set, of which only three were causal in all cases, were correlated with each other at the same level. Increasing numbers of SNPs included in the gene-specific set were intended to emulate increasing window sizes in which SNPs are assigned to genes. In the main text, we used a window size of ±50Kb from the gene body.

The results in **Figure S14** demonstrate that for CS LD structures in which all noncausal SNPs are in LD with at least one causal SNP, increasing numbers of including SNPs (i.e., from using kilobase windows of increasing size) does not greatly reduce the power of the gene-based association test if the CS correlation parameter is nonzero. However, if LD is a decaying function of base pair distance as in the AR1 scenario, which is commonly observed in real data, increasing numbers of included SNPs (i.e., from using kilobase windows of increasing size) can greatly decrease the power of the gene-based association test. These results suggest that increasingly large window sizes that form gene-specific SNP sets may reduce the statistical power of the gene-based association test if no additional causal SNPs are captured in the increasingly large window.

Figure S14: Theoretical power as the SNP set kilobase window size increases

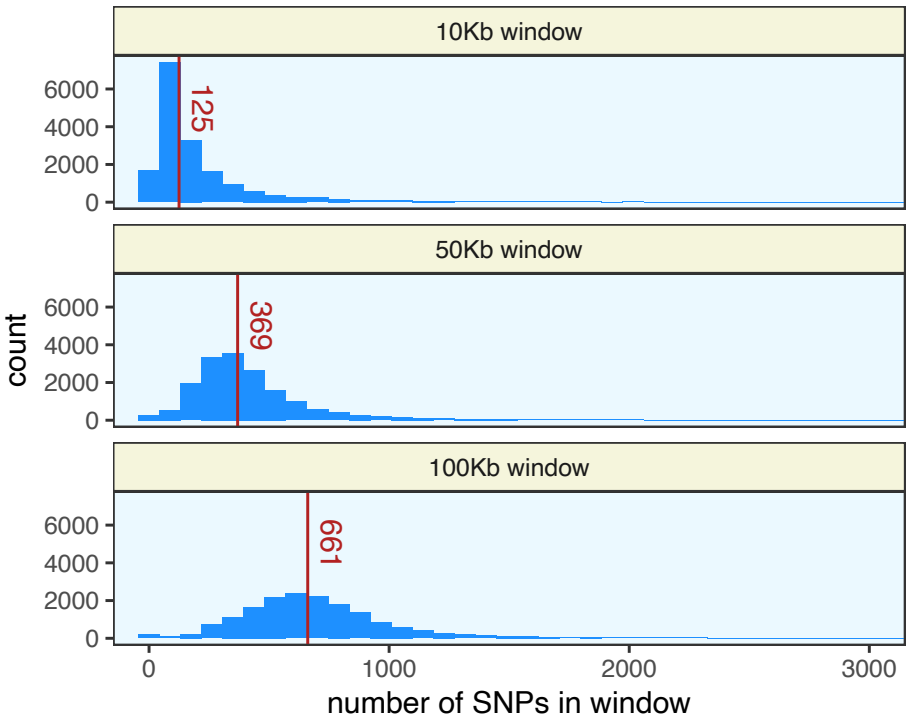


Displayed are the theoretical power values of the gene-based association test (GenT) as the number of included SNPs increases (e.g., from increasingly large sizes of the base pair windows which assign SNPs to genes) under different scenarios of LD strength and structure (see text above figure). The R code used to recreate this figure is available at https://github.com/noahlorinczcomi/gent_analysis/blob/main/GenT_power_changing_SNP_window_size.R

S12.2 Empirical SNP set sizes from varying window sizes

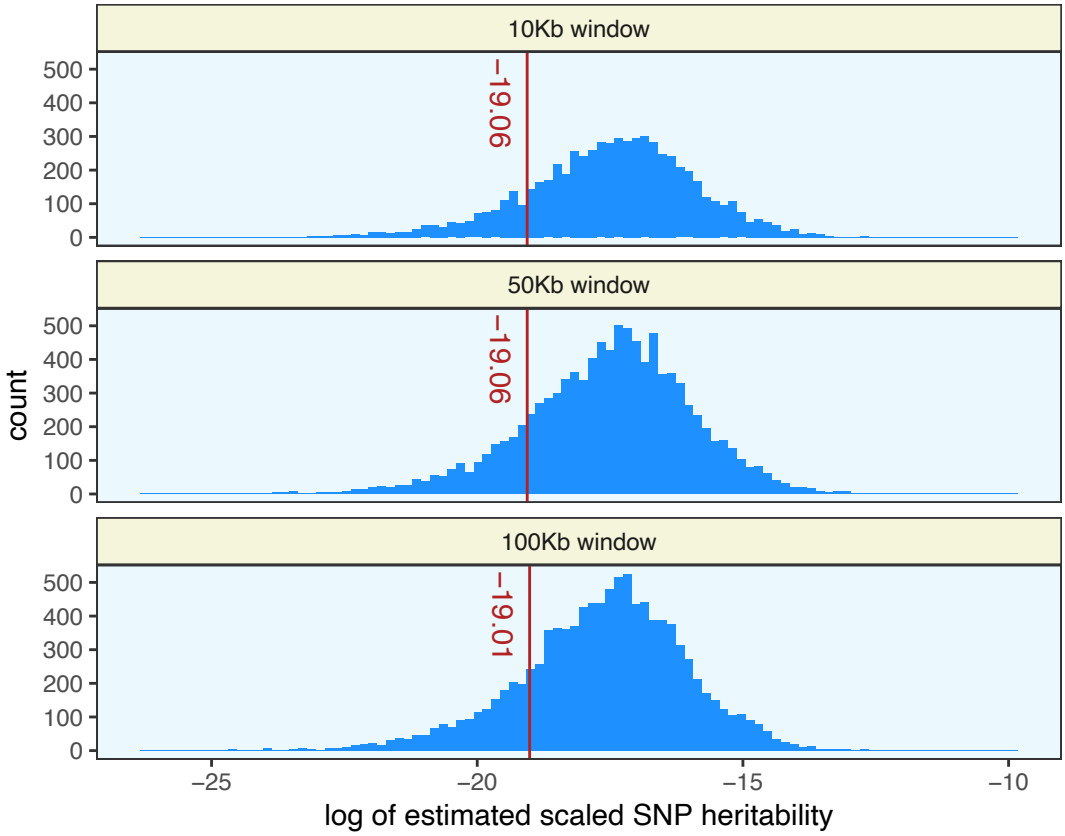
In this subsection, we show the distributions of the number of SNPs captured by windows of various kilobase sizes in gene-based association testing. We used all SNPs in the Type 2 Diabetes (T2D) European cohort (Suzuki et al., 2023) that were also in the 1000 Genomes Phase 3 European reference panel. For each window size of 10Kb, 50Kb, and 100Kb, we included all SNPs that had a build37 base pair location within the window distance with the build37 start or end position of each of 18,260 gene bodies on chromosomes 1-22. Genes and the build37 base pair locations of their starts and ends were downloaded from the Ensembl Biomart tool (https://useast.ensembl.org/biomart/martview). **Figure S15** shows these distributions and suggests that the median SNP set size when using a 10Kb window is 125 SNPs, when using a 50Kb window is 369 SNPs, and when using a 100Kb window is 661 SNPs. **Figure S16** shows the distribution of estimated scaled SNP heritability values in each SNP set using the LD score regression model (Bulik-Sullivan et al., 2015). These quantities are exactly the estimate coefficient from linear regression of SNP chi-square statistics on LD scores calculated from ±1Mb windows. These results suggest that increasing window sizes do not on average produces estimates of partitioned heritability which are larger. Together with the results of **Figure S15** and **Figure S17**, these results suggest that widows even as small as 10Kb may also be suitable for gene-based association testing and that SNPs in a 10Kb window may explain more per-SNP average heritability than SNPs in larger windows.

Figure S15: Empirical sizes of SNP sets as kilobase window increases in size



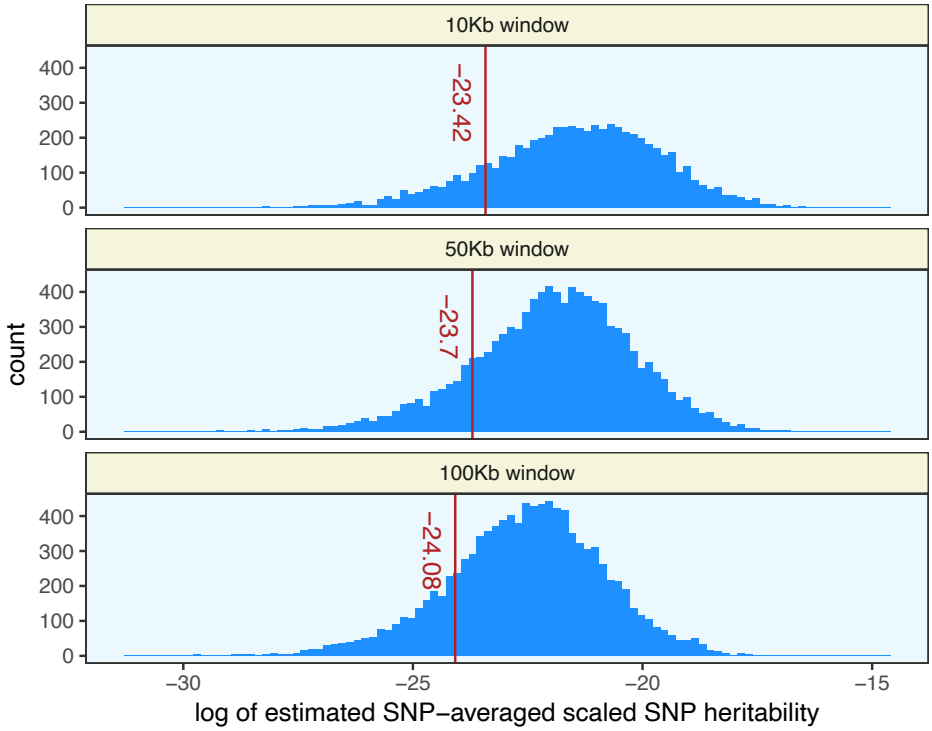
Displayed are the distributions of SNP set size as the size of the kilobase window increases from 10Kb to 100Kb using read data (see text above). Red vertical lines and the text to the right of them represent the median SNP set sizes. SNPs sets containing more than 3000 SNPs (0.69%) were excluded from the plot for viewability.

Figure S16: Distribution of scaled SNP heritability in SNP set windows of varying size



Displayed are the distributions of log scaled SNP heritability – the quantity h^2M^{-1} in Equation (1) of Bulik-Sullivan et al. (2015) – for varying sizes of SNP set windows used in gene-based association testing. In each set of SNPs formed by the various windows, we estimated scaled heritability using LD score regression and display the distribution of their log transformed values in this plot. Vertical lines together with the text to the left of them indicate the median values. Full R code used to perform these analyses and create Figures S15-S16 is available at https://github.com/noahlorinczcomi/gent_analysis/blob/main/window_size_SNP_set_size_h2_T2D.R

Figure S17: Distribution of SNP-averaged heritability in SNP set windows of varying size



Displayed are the distributions of log scaled SNP heritability – the quantity h^2M^{-1} in Equation (1) of Bulik-Sullivan et al. (2015) – divided by the number of observed SNPs in the window for varying sizes of SNP set windows used in gene-based association testing. This quantity is proportional to the average heritability explained per SNP in the window. Vertical lines together with the text to the left of them indicate the median values.

S13 Type I/II error comparisons with other methods

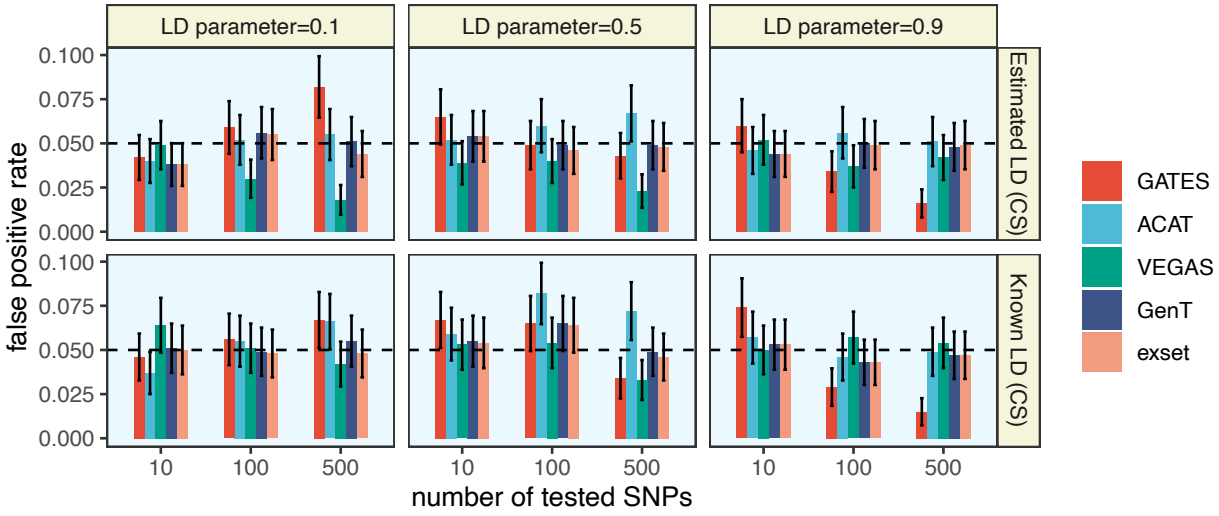
In this section, we compare the Type I and II error rates of GenT to that of existing methods including VEGAS (Liu et al., 2010), GATES (Li et al., 2011), ACAT (Liu et al., 2019), and exset (Lorincz-Comi, 2025c). For both Type I and Type II (power) simulations, which each used 1000 replicates, we generated data under the following model for a single gene

$$\begin{aligned}
&\mathbf{z} \sim N(\mathbf{z}_0, \mathbf{R}_0)\\
&\mathbf{R} \sim \text{Wishart}_m(505, \mathbf{R}_0)\\
&\mathbf{R} \leftarrow \text{C}(\mathbf{R})\\
&S = \mathbf{z}^\top \mathbf{z}\\
&m \in \{10{,}100{,}500\},
\end{aligned}$$

where $\text{C}(\mathbf{A})$ converts the sample covariance matrix $\mathbf{A}$ to a sample correlation matrix, m is the number of SNPs used, and $\mathbf{R}_0$ had a compound symmetry structure. We let $\mathbf{z}_0 = \mathbf{0}$ in the Type I error simulations and one element of $\mathbf{z}_0$ to equal $8 + \frac{1}{3}$ in the Type II error simulation, which implies 3 causal SNPs corresponding to one gene explaining 0.05%

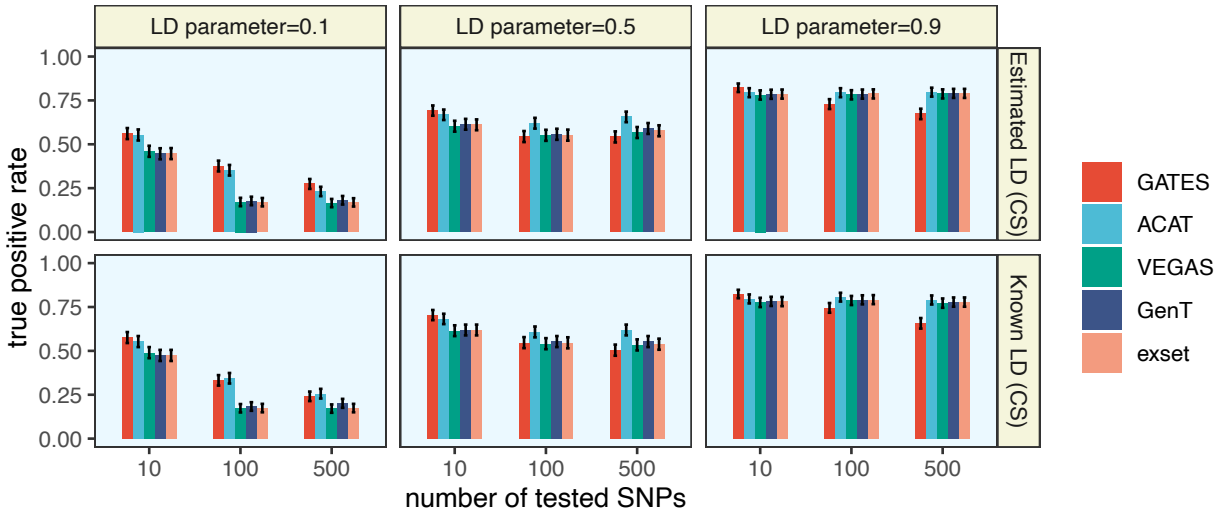
heritability in a GWAS of 50K subjects (see Lorincz-Comi et al., 2024). Each method required eigenvalues of the true LD matrix, which in one set of conditions we assumed were unavailable. In their place, we used the empirical eigenvalues of the matrix $\mathbf{R}$ assumed to be estimated from 505 individuals, the number of European ancestry individuals in the 1000 Genomes Phase 3 reference panel. The results in **Figure S18** suggest that GenT an exset control their Type I error at 5% across all simulation scenarios. GATES had inflated Type I error when 500 SNPs were used and the correlation between them was 0.1. When the correlation between them was 0.9 and 500 SNPs were used, GATES had deflated Type I error suggested reduced power. The VEGAS method, which uses a numerical approximation to the null distribution, had deflated Type I error when 500 SNPs were used and there was 0.5 LD between SNPs. The ACAT method, which performs the Cauchy combination test (Liu & Xie, 2020), had slightly inflated Type I error when using 500 SNPs correlated at known level 0.1, 100 and 500 SNPs correlated at known level 0.5, and 500 SNPs correlated at the estimated level 0.5. The results in **Figure S19** suggest that GATES and ACAT display similar patterns of power across most simulation scenarios, and GenT, VEGAS, and exset display similar patterns of power. ACAT and GATES on average have greater than other methods, though this may be at least partially explained by their inflated Type I errors.

Figure S18: Type I error comparison between GenT and other methods



Displayed are the proportions of null test statistics which were significant at the 0.05 level after 1000 replicates of the simulation described above. The full R code used to reproduce these results is available at:
https://github.com/noahlorinczcomi/gent_analysis/tree/main/simulations/type1_comparisons

Figure S19: Power comparison between GenT and other methods



Displayed are the proportions of non-null test statistics which were significant at the 0.05 level after 1000 replicates of the simulation described above. The full R code used to reproduce these results is available at:
https://github.com/noahlorinczcomi/gent_analysis/blob/main/simulations/power_comparisons/simulation.R

References

Bellenguez, C., Küçükali, F., Jansen, I. E., Kleineidam, L., Moreno-Grau, S., Amin, N., ... & Goldhardt, O. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. Nature genetics, 54(4), 412-436.

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, ... & Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics, 47(3), 291-295.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742-015.

Covo, S., & Elalouf, A. (2014). A novel single-gamma approximation to the sum of independent gamma variables, and a generalization to infinitely divisible distributions.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432-441.

Gao, X., Pu, D. Q., Wu, Y., & Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statistica Sinica*, 1123-1146.

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities* (Vol. 195). Springer Science & Business Media.

Jiang, L., Miao, L., Yi, G., Li, X., Xue, C., Li, M. J., ... & Li, M. (2022). Powerful and robust inference of complex phenotypes' causal genes with dependent expression quantitative loci by a median-based Mendelian randomization. *The American Journal of Human Genetics*, *109*(5), 838-856.

LeBlanc, M., Zuber, V., Thompson, W. K., Andreassen, O. A., Schizophrenia and Bipolar Disorder Working Groups of the Psychiatric Genomics Consortium, Frigessi, A., & Andreassen, B. K. (2018). A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. *BMC genomics*, *19*, 1-15.

Li, M. X., Gui, H. S., Kwan, J. S., & Sham, P. C. (2011). GATES: a rapid and powerful gene-based association test using extended Simes procedure. The American Journal of Human Genetics, 88(3), 283-293.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., ... & Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. The American Journal of Human Genetics, 87(1), 139-145.

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., & Lin, X. (2019). ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. The American Journal of Human Genetics, 104(3), 410-421.

Lorincz-Comi, N., Yang, Y., Li, G., & Zhu, X. (2024). MRBEE: A bias-corrected multivariable Mendelian randomization method. *Human Genetics and Genomics Advances*, *5*(3).

Lorincz-Comi, N., & Cheng, F. (2025a). Bayesian estimation of shared polygenicity identifies drug targets and repurposable medicines for human complex diseases. medRxiv, 2025-03.

Lorincz-Comi, N. (2025b). Exact set-based hypothesis testing of correlated normal statistics using a distributional equivalence in a high-dimensional space. Retrieved from https://github.com/noahlorinczcomi/exset.

Stewart, T., Strijbosch, L. W. G., Moors, H., & Batenburg, P. V. (2007). A simple approximation to the convolution of gamma distributions.

Suzuki, K., Hatzikotoulas, K., Southam, L., Taylor, H. J., Yin, X., Lorenz, K. M., ... & de Silva, H. J. (2023). Multi-ancestry genome-wide study in> 2.5 million individuals reveals heterogeneity in mechanistic pathways of type 2 diabetes and complications. medRxiv.

Zou, Y., Carbonetto, P., Wang, G., & Stephens, M. (2022). Fine-mapping from summary data with the “Sum of Single Effects” model. PLoS genetics, 18(7), e1010299.