SCAMPI: A scalable statistical framework for genome-wide interaction testing harnessing cross-trait correlations

Shijia Bian, MS[1]; Andrew J. Bass, PhD[2]; Yue Liu, PhD[3]; Aliza P. Wingo, MD, MS[4,5]; Thomas Wingo, MD[3]; David J. Cutler, PhD[2]; Michael P. Epstein, PhD[2,*]

Affiliations:

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, 30329, USA

[2]Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA, 30329, USA

[3]Department of Neurology, University of California, Davis, Sacramento, CA 95817, USA

[4]Department of Psychiatry, University of California, Davis, Sacramento, CA 95817, USA

[5]Division of Mental Health, VA Northern California Health Care System, CA 95655, USA

* Correspondence: mpepste@emory.edu

Abstract

Family-based heritability estimates of complex traits are often considerably larger than their single-nucleotide polymorphism (SNP) heritability estimates. This discrepancy may be due to non-additive effects of genetic variation, including variation that interacts with other genes or environmental factors to influence the trait. Variance-based procedures provide a computationally efficient strategy to screen for SNPs with potential interaction effects without requiring the specification of the interacting variable. While valuable, such variance-based tests consider only a single trait and ignore likely pleiotropy among related traits that, if present, could improve power to detect such interaction effects. To fill this gap, we propose SCAMPI (Scalable Cauchy Aggregate test using Multiple Phenotypes to test Interactions), which screens for variants with interaction effects across multiple traits. SCAMPI is motivated by the observation that SNPs with pleiotropic interaction effects induce genotypic differences in the patterns of correlation among traits. By studying such patterns across genotype categories among multiple traits, we show that SCAMPI has improved performance over traditional univariate variance-based methods. Like those traditional variance-based tests, SCAMPI permits the screening of interaction effects without requiring the specification of the interaction variable and is further computationally scalable to biobank data. We employed SCAMPI to screen for interacting SNPs associated with four lipid-related traits in the UK Biobank and identified multiple gene regions missed by existing univariate variance-based tests. SCAMPI is implemented in software for public use.

Introduction

Genome-wide association studies (GWAS) have successfully improved our understanding of the role of common single-nucleotide polymorphisms (SNPs) on many complex human traits and diseases. Researchers can further use SNP data from a GWAS study to estimate a trait's narrow-sense heritability (proportion of trait variance due to additive genetic effects) using statistical techniques like GCTA and LD Score Regression (LDSC).[1; 2] Interestingly, SNP-based heritability estimates of a complex trait are routinely smaller than the corresponding family-based estimates of narrow-sense heritability based on kinship. For instance, studies have reported SNP-based estimates of narrow-sense heritability for body mass index (BMI) to be 0.3, which is considerably less than the narrow-sense heritability estimates of 0.47-0.90 for BMI reported in twin studies.[3; 4] For Alzheimer's Disease (AD), family-based heritability estimates of the disease range from 0.60-0.80, whereas the latest population-based AD GWAS meta-analyses estimated the narrow-sense heritability from SNP data to be between 0.06-0.41.[5-13] Likewise, a GWAS analysis of Amyotrophic Lateral Sclerosis (ALS) estimated SNP-based heritability of approximately 0.21, which is significantly less than the estimates of 0.38-0.85 observed in twin studies.[14]

The gap between family-based estimates of narrow-sense heritability and corresponding SNP-based estimates may be due to several factors, including rare causal variation poorly tagged by common SNPs as well as shared familial environmental effects ignored in traditional family-based heritability estimation.[15; 16] Here, we focus on another possible explanation for this gap - the presence of non-additive effects (including higher-order genetic interactions) on complex traits and diseases. As noted in the Supplemental Materials (S1), we can show that higher-order interactions of a complex trait inflate narrow-sense heritability estimates more among close

relatives (traditionally used for family-based estimates of heritability) than distantly related individuals (traditionally used to estimate GWAS heritability via LDSC/GCTA).[17] Thus, higher-order interactions can explain the discrepancy between family-based and SNP-based heritability estimates observed for many complex human traits. This motivates the search for genetic variants in large-scale genetic studies that demonstrate non-additive effects, including gene-gene and gene-environment interactions.

While studies have identified SNPs demonstrating interaction effects on complex traits,[18-23] genome-wide investigation of non-additive effects is inherently challenging.[24; 25] Comprehensive genome-wide testing of SNP-SNP (epistatic) interactions is computationally intractable as 10 million SNPs can lead to approximately 5×10^{13} potential interaction tests. Even if such analyses were tractable, the resulting multiple-testing adjustment cripples the power to detect epistatic effects. Gene-environment interaction analyses require fewer tests and are more computationally feasible, but measuring the right environmental determinants can be difficult and is often unknown.[26-29] To circumvent uncertainty about the right environmental factor yet still test for evidence of interaction, Paré et al. proposed an efficient variance-based method for a quantitative trait that screens for SNPs with possible interactive effects without requiring specification of the interacting factor.[30] Recognizing that a SNP with an interaction effect on a trait induces trait variance that differs by genotype (see Supplemental Figure S1), Paré screened for SNPs with potential interaction effects by testing for equality of variances across genotype categories using Levene's test.[31] Researchers have successfully applied this type of variance-based approach within the UK Biobank to identify genetic variants with interaction effects on obesity phenotypes and cardiometabolic serum biomarkers.[32; 33]

The variance-based test of Paré is a univariate test that considers whether a SNP has an interactive effect with a single phenotype. However, biobanks routinely collect detailed information on a large collection of related phenotypes with shared genetic effects. Many recent methods of gene mapping illustrate the appeal of leveraging the ubiquitous phenomenon of pleiotropy across related traits when present.[34-37] Consequently, if pleiotropic genetic variants with interactive effects exist, we expect a multi-trait statistical method that leverages this information will have improved performance over existing univariate variance-based interaction procedures. Bass et al. recently showed that a SNP with an interaction effect induces not only variance but also covariance patterns between traits that differ by genotype (which we illustrate in Supplemental Figure S2).[38] Based on this observation, the authors developed a kernel framework for interaction testing that assessed where similarity in variance/covariance patterns among a group of modeled traits correlated with genotypic similarity at a test SNP. While more powerful than standard variance-based testing, the kernel framework of Bass lacks practical features for genetic analysis such as the inability to identify the specific phenotypes (among those modeled) that demonstrate interaction effects with the test SNP. Identifying these specific phenotypes are of substantial value for further downstream analyses.

To this end, we propose here an efficient screening method SCAMPI (Scalable Cauchy Aggregate test using Multiple Phenotypes to test Interactions) for identifying potential SNPs with interaction effects using multiple phenotypes. SCAMPI fits simple regression models relating SNP genotype to (standardized) cross products of all pairwise combinations of traits under consideration and then aggregates the correlated p-values from these separate regression tests together into an omnibus test using the Cauchy Combination Test.[39; 40] Similar to variance-based interaction tests, SCAMPI does not require specification of the factor that interacts with the SNP

of interest, thereby reducing the computational and testing burden and enabling the scaling of the method to biobank-size datasets. Moreover, SCAMPI scales to handle many related phenotypes and can identify the specific phenotype(s) that have interaction effects among those modeled. Using simulations, we show that SCAMPI can detect interactions under various scenarios and has improved performance over univariate variance-based interaction procedures. We also applied SCAMPI to lipid panel data (an indicator of risk of heart disease and stroke) in the UK Biobank (UKBB) and identified several genes with putative interaction effects that were missed by standard univariate variance-based procedures. For public use, SCAMPI is implemented as an R package.

Materials and Methods

Motivation: We first show that a SNP with a pleiotropic interaction effect yields trait correlation patterns that differ by genotype category. We could analogously show that a SNP with a pleiotropic interaction trait effect influences the covariance patterns between traits but chose to focus on correlation due to the scale-free nature of the latter measure. For subject i, define G_i as the subject's genotype at a test SNP and define W_i as some factor (either genetic or environmental) that interacts with the SNP to influence multiple traits. Suppose subject i possesses two correlated traits $Y_{i,1}$ and $Y_{i,2}$ that are generated under the relationships:

$$Y_{i,1} = \alpha_1 + \beta_1 G_i + \gamma_1 W_i + \delta_1 G_i W_i + \epsilon_{i,1};\ Y_{i,2} = \alpha_2 + \beta_2 G_i + \gamma_2 W_i + \delta_2 G_i W_i + \epsilon_{i,2}.$$

Here, $\beta_j, \gamma_j, \delta_j$ denote the main effect of genotype, the main effect of the factor, and two-way interaction effect between genotype and factor, respectively, on trait j ($j = 1, 2$). We further assume each of the error terms $\epsilon_{i,1}$ and $\epsilon_{i,2}$ has a standard normal distribution $\epsilon_{i,1}, \epsilon_{i,2} \sim N(0,1)$. Without loss of generality, further assume W_i is distributed as $W_i \sim N(0,1)$ and is independent of G_i.

Based on the trait models listed above, Paré previously showed that that variance of $Y_1(Y_2)$ differs by G when the genotype has an interaction effect on trait 1 (trait 2), respectively.[30] Additionally, when pleiotropic interaction effects exist, we can show the correlation of traits 1 and 2 also differ by genotype. In Supplementary Materials S2, we derive the correlation between $Y_{i,1}$ and $Y_{i,2}$ conditional on genotype G_i as:

$$cor\left(Y_{i,1}, Y_{i,2} \middle| G_i = g_i\right) = \frac{\gamma_1\gamma_2 + (\delta_1\gamma_2 + \delta_2\gamma_1)g_i + \delta_1\delta_2{g_i}^2}{\sqrt{(\gamma_1 + \delta_1 g_i)^2 + 1} \times \sqrt{(\gamma_2 + \delta_2 g_i)^2 + 1}} \qquad (1)$$

Equation (1) shows that the correlation between two traits differs by genotype when either a) the genotype interacts with the factor on both phenotypes or b) the genotype interacts with the factor on at least one of the phenotypes, provided the factor has a main effect on the other phenotype. We can see that if the SNP has no interaction effect on either phenotype ($\delta_1 = \delta_2 = 0$), the phenotypic correlation will not differ by genotype even when main effects for the factor exist ($\gamma_1 \neq 0, \gamma_2 \neq 0$).

The above result suggests an efficient strategy for screening SNPs with potential interaction effects. Instead of performing traditional interaction analyses, which mandates defining potential interacting factors W_i, we can instead screen for SNPs with interaction effects without having to specify W_i by examining whether the correlation between traits changes as a function of the linear and quadratic effects of genotype. Such modeling provides a workaround in situations where interacting covariates are uncollected or inaccurately recorded. The screening procedure further provides an efficient alternative strategy for genome-wide epistatic analysis in that it does not require direct modeling of the interacting genetic factor, which substantially reduces the number of tests to be considered. If we are analyzing *M* SNPs, SCAMPI requires only *M* tests

157 whereas comprehensive epistatic analysis requires $\binom{M}{2}$ tests. Thus, when M = 200K (M = 2M),
158 SCAMPI reduces the number of tests required by approximately 5 (6) orders of magnitude.

159 Rather than model trait correlation as a function of linear and quadratic effects of genotype
160 mentioned above, we note that we can alternatively parameterize this relationship using a general
161 genotype model that allows for separate effects of each genotype relative to a baseline category.
162 That is, for some outcome Y^*, the coefficient estimates of $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\alpha}_3$ in the regression model
163 $Y^* = \alpha_1 + \alpha_2 G + \alpha_3 G^2 + \epsilon^*$ can be directly mapped to coefficient estimates $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ in a
164 model $Y^* = \beta_1 + \beta_2 G_1 + \beta_3 G_2 + \epsilon$, where G_1 and G_2 are genotype indicators for those with 1 and
165 2 copies of the reference allele, respectively (those with 0 copies are treated as baseline). Given
166 the familiarity of this general genotype model in GWAS, [41-44] we chose to use this alternative
167 parameterization in our method moving forward.

168 *Notation and Trait Standardization:* Assume a sample of N unrelated subjects that possess
169 J continuous phenotypes. Let $\boldsymbol{Y}_j = \left(Y_{1,j}, Y_{2,j}, \dots, Y_{N,j}\right)^T$ denote the N x 1 vector of observations for
170 trait j $(j = 1, \dots, J)$. Define $\boldsymbol{G} = (G_1, G_2, \dots, G_N)^T$ as an $N \times 1$ vector of genotypes for one test
171 SNP, where G_i represents $[0, 1, 2]$ copies of the minor allele that subject i possesses at the site. As
172 noted in the previous section, we are interested in applying a general genotype model for
173 interaction testing as it naturally captures the linear and quadratic effects of genotype shown in
174 equation (1). Consequently, further define $G_i^{(1)} = I[G_i = 1]$ and $G_i^{(2)} = I[G_i = 2]$ as indicator
175 variables for genotype categories 1 and 2, respectively (we treat genotype category 0 as baseline).
176 Finally, let $\boldsymbol{Z}$ be an $N \times K$ matrix of confounding variables. These confounding variables can be a
177 mixture of continuous or categorical features. Common confounder examples include age,
178 biological sex, batch ID, and principal components of ancestry to deal with population
179 stratification.

Our goal is to detect a SNP with an interaction effect that yields correlation patterns that differ by genotype. Such trait pattern differences can erroneously arise if the main effect of the genotype, as well as main and variance effects of confounders (such as population structure), are unaccounted for prior to analysis.[45; 46] To avoid this issue, we first standardize and adjust each $\boldsymbol{Y}_j (j = 1, \dots, J)$ prior to analysis using a double generalized linear model (DGLM) that corrects for the mean effects of the test SNP and confounders, as well as the potential variance effects of confounders.[47,48]

DGLM is composed of two sub-models, where the first sub-model controls population mean, and the second sub-model controls population variance. For our work, the first sub-model adjusts $\boldsymbol{Y}_j$ for the mean effects of $\boldsymbol{G}^{(1)}$ & $\boldsymbol{G}^{(2)}$ and confounders $\boldsymbol{Z}$ using the following framework:

$$\boldsymbol{Y}_j = [\mathbf{1} \quad \boldsymbol{G}^{(1)} \quad \boldsymbol{G}^{(2)} \quad \boldsymbol{Z}] \begin{bmatrix} \beta_{j,0} \\ \beta_{j,G^{(1)}} \\ \beta_{j,G^{(2)}} \\ \boldsymbol{\beta}_{j,Z} \end{bmatrix} + \boldsymbol{\varepsilon}_j$$

where $\beta_{j,0}$ is the intercept associated with the j^{th} trait. $\beta_{j,G^{(1)}}$ and $\beta_{j,G^{(2)}}$ are the regression coefficient for $\boldsymbol{G}^{(1)}$ and $\boldsymbol{G}^{(2)}$ respectively, and $\boldsymbol{\beta}_{j,Z}$ is a $K \times 1$ vector of regression coefficients for confounders $\boldsymbol{Z}$. Finally, $\boldsymbol{\varepsilon}_j$ is a $N \times 1$ vector of residual errors that follow

$$\boldsymbol{\varepsilon}_j \sim \boldsymbol{MVN}\left(\boldsymbol{\mu}_{\varepsilon_j} = \mathbf{0}_{\boldsymbol{Nx1}}, \boldsymbol{\Sigma}_{\varepsilon_j} = \begin{bmatrix} \sigma^2_{\varepsilon_{1,j}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{\varepsilon_{N,j}} \end{bmatrix}\right) \quad (\mathbf{2})$$

The second sub-model of the DGLM then models $\boldsymbol{\varepsilon}_j$ in (2) as a function of confounders $\boldsymbol{Z}$ using the following framework using the log link function:

$$\log\left(E\left(diag\left(\boldsymbol{\Sigma}_{\varepsilon_j}\right)\right)\right) = \boldsymbol{\varphi}_j = [\mathbf{1} \quad \boldsymbol{Z}] \begin{bmatrix} \gamma_0 \\ \boldsymbol{\gamma}_Z \end{bmatrix}$$

where $\boldsymbol{\varphi}_j$ is the *N x 1* column vector representing the expected residual variance of the j^{th} observed trait. Here, γ_0 is the intercept while $\boldsymbol{\gamma}_Z$ represents the $K \times 1$ column vector of confounder effects on the variance. The error distribution to be used in the two sub-models is Gaussian.

We fit the above DGLM using the R package "dglm". Let $\tilde{\boldsymbol{Y}}_{\boldsymbol{j}}$ $(j = 1, \dots, J)$ denote the adjusted and standardized form for trait *j* produced from the DGLM model fit. We subsequently use $\tilde{\boldsymbol{Y}}_{\boldsymbol{j}}$ $(j = 1, \dots, J)$ to construct appropriate measures for our downstream screening analyses for interaction effects.

Analysis Strategy: For $J = 2$ traits, we show in Supplemental Materials (S3) that we can approximate the sample Pearson correlation coefficient of traits $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ as the average of the $N \times 1$ vector of cross products of the traits after standardization, $\tilde{\boldsymbol{Y}}_1$ and $\tilde{\boldsymbol{Y}}_2$. That is, we estimate the Pearson correlation between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ as the sample average of

$$\tilde{\boldsymbol{Y}}_1 \odot \tilde{\boldsymbol{Y}}_2 = \left(\tilde{Y}_{1,1} \cdot \tilde{Y}_{1,2}\,, \tilde{Y}_{2,1} \cdot \tilde{Y}_{2,2}, \dots, \tilde{Y}_{N,1} \cdot \tilde{Y}_{N,2}\right)^{\boldsymbol{T}}$$

where $\odot$ denotes the row-wise product operator of two vectors. Similarly, we can estimate the variance of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ by $\tilde{\boldsymbol{Y}}_1 \odot \tilde{\boldsymbol{Y}}_1$ and $\tilde{\boldsymbol{Y}}_2 \odot \tilde{\boldsymbol{Y}}_2$, respectively.

Using these estimates, we construct a screening procedure to identify a SNP with an interaction effect on trait $\boldsymbol{Y}_1$ and/or $\boldsymbol{Y}_2$ by assessing whether SNP genotype $\boldsymbol{G}$ is associated with either $\tilde{\boldsymbol{Y}}_1 \odot \tilde{\boldsymbol{Y}}_1$, $\tilde{\boldsymbol{Y}}_2 \odot \tilde{\boldsymbol{Y}}_2$, or $\tilde{\boldsymbol{Y}}_1 \odot \tilde{\boldsymbol{Y}}_2$. Examination of the relationship of $\boldsymbol{G}$ with $\tilde{\boldsymbol{Y}}_1 \odot \tilde{\boldsymbol{Y}}_1$ (or $\tilde{\boldsymbol{Y}}_2 \odot \tilde{\boldsymbol{Y}}_2$) is similar to assessing whether trait variance differs by genotype (which Paré[30] investigated using Levene's test) while the study of $\boldsymbol{G}$ with $\tilde{\boldsymbol{Y}}_1 \odot \tilde{\boldsymbol{Y}}_2$ leverages additional information on interactions based on differences in trait correlations. To implement our procedure, we fit 3 separate linear regression models; each model treating one of $\tilde{\boldsymbol{Y}}_1 \odot \tilde{\boldsymbol{Y}}_1$, $\tilde{\boldsymbol{Y}}_2 \odot \tilde{\boldsymbol{Y}}_2$, or $\tilde{\boldsymbol{Y}}_1 \odot \tilde{\boldsymbol{Y}}_2$ as outcome with SNP genotype $\boldsymbol{G}^{(1)}$ and $\boldsymbol{G}^{(2)}$ as predictors. Each regression models

produces a p-value based on a two-degree-of-freedom test. Since the resulting 3 p-values from these regression tests are correlated, we can then combine them into an omnibus p-value (described in the next section) to assess whether the SNP has an interaction effect on at least one of the two traits under study.

The above example considered two traits under study. However, the strategy easily extends to the study of $J > 2$ correlated traits as well. Assuming J traits, we fit J regression models that regress $\widetilde{\boldsymbol{Y}}_j \odot \widetilde{\boldsymbol{Y}}_j$ on $\boldsymbol{G}^{(1)}$ and $\boldsymbol{G}^{(2)}$ $(j = 1, \dots, J)$ and further fit $\binom{J}{2}$ additional regression models that regress $\widetilde{\boldsymbol{Y}}_j \odot \widetilde{\boldsymbol{Y}}_l$ $(j, l = 1, \dots, J;\ j \neq l)$ on $\boldsymbol{G}^{(1)}$ and $\boldsymbol{G}^{(2)}$. We then can combine the $J_C = \left(J + \binom{J}{2}\right)$ p-values from these tests together to assess whether the SNP has an interaction effect on at least one of the J traits under study.

Cauchy Combination Test (CCT): After obtaining the J_C p-values above, we create a final omnibus test for whether the test SNP has an interactive effect on any of the traits under consideration using the Cauchy Combination Test (CCT),[39; 40] which is a popular technique for aggregating many potentially dependent tests of high dimension together into an omnibus framework. CCT has provable type I error rate control for genome wide significance thresholds even when p-values are dependent. CCT is especially useful when an SNP signal is sparse and only affects a subset of the traits under consideration. The test statistics of CCT is a weighted sum of the Cauchy transformation of individual p-values in SCAMPI. Let p_r to denote the dependent individual p-value from the r^{th} regression test $(r = 1,2, \dots, J_c)$. The CCT statistic is defined as

$$T = \sum_{r=1}^{J_C} \frac{1}{J_c} \tan\{(0.5 - p_r)\pi\} \qquad (3)$$

Under the null hypothesis of no SNP interactive effect with any of the traits under consideration, T in (3) follows a standard Cauchy Distribution, i.e., $T \sim Cauchy(X_0 = 0, \gamma = J_C)$. This derived p-value is the SCAMPI p-value at the given genotype $\boldsymbol{G}$.

Overview of the SCAMPI Framework: Our SCAMPI framework aggregates the regression tests outlined earlier with the CCT to produce an omnibus p-value for testing whether the SNP has an interactive effect with at least one of the traits under study. SCAMPI, which is implemented in a public R package of the same name, requires the following inputs:

a. Multiple target traits are denoted as $\boldsymbol{Y}$. Should these traits not follow a normal distribution, users can apply a rank-based Inverse Normal Transformation to normalize the traits, if desired.
b. The confounding variables, represented by $\boldsymbol{Z}$;
c. One test SNP, represented by $\boldsymbol{G}$ and coded as $\boldsymbol{G}^{(1)}$ and $\boldsymbol{G}^{(2)}$.

SCAMPI then follows the workflow depicted in Figure 1.

Application to UK Biobank Data: We applied SCAMPI to identify SNPs with potential interaction effects on lipid measures within the UK Biobank (application ID 42223). We focused attention on four lipid-related measures: high-density lipoprotein cholesterol (HDL-C), low-density lipoproteins cholesterol (LDL-C), triglycerides (TGs), and Body Mass Index (BMI). Both the sample and SNP QC procedures are in accordance with Marderstein et al.[49] Similar QC procedures were also carried out in multiple studies.[50] From the cohort, we excluded individuals who either (1) had missing heterozygosity information, (2) were outliers in terms of heterozygosity or had missing genotype rates greater than 0.02, (3) had over 10 putative third-degree relatives in the kinship table, (4) were omitted from the kinship inference procedure, or (5) were either self-reported as anything other than 'White British' or did not show similar genetic ancestry to this

group based on a principal components analysis of the genotypes. After performing this quality control, 337,422 independent subjects remained (N_{Female}= 181,203; N_{Male}= 156,219). Moreover, the UKB employed two genotyping arrays. In this post-QC sample, we have the UK Biobank Axiom array (N_{UKBB}= 300,345) and the UK BiLEVE array (N_{UKBL}= 37,077). For the SNP QC, genotypes were discarded if they had an INFO score < 0.8, MAF < 0.05 and HWE p-value < 10^{-10}. After SNP QC procedures, 288,910 SNPs were retained. Finally, 277,653 SNPs were included for analysis using SCAMPI after applying a 10% missing rate threshold.

We first adjusted the four lipid-related traits for confounders, including the first six genetic principal components, biological sex, age, age squared (age^2), and the type of genotyping array, before applying SCAMPI to these traits. Notably, the first six principal components effectively captured population structure at subcontinental geographic scales.[51,52] Of the initial set of 337,422 independent subjects, 288,709 possessed complete information on all traits and confounders and were considered moving forward. We first transformed the four traits using the inverse normal transformation (INT) to align the traits, which is a common practice to ensure the residual of traits is normally distributed In a regression model such as DGLM.[53-56] The distribution of the four traits, both pre and post-INT, can be found in Supplemental Figure S3 (a) - (d). Correlation between post-INT traits was 0.1246 for HDL-C and LDL-C, -0.4938 for HDL-C and TG, -0.3809 for HDL-C and BMI, 0.2797 for LDL-C and TG, 0.0394 for LDL-C and BMI, and 0.3708 for TG and BMI.

Simulations: We conducted comprehensive simulations to evaluate the type-I error rate of SCAMPI under a variety of scenarios. For each scenario, we simulated a sample size of 300,000 to reflect biobank-scale datasets. Each scenario is analyzed based on 100,000 simulations. We assumed $J = 2,4,8$ traits and simulated the trait values for the i^{th} individual based on the multivariate normal distribution illustrated below:

$$\begin{pmatrix} Y_{i,1} \\ \vdots \\ Y_{i,J} \end{pmatrix} \sim MVN\left(\boldsymbol{\mu} = \begin{bmatrix} \alpha_1 + \beta_1 G_i + \gamma_1 W_i \\ \vdots \\ \alpha_J + \beta_J G_i + \gamma_J W_i \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & \dots & \sigma^2 \\ \vdots & \ddots & \vdots \\ \sigma^2 & \dots & 1 \end{bmatrix}\right) \quad (4)$$

For predictors, we generated the test SNP genotype G_i under Hardy-Weinberg equilibrium, assuming the SNP had a minor-allele frequency of either 0.05 or 0.25. We further generated a factor W_i that followed a standard normal distribution. For the choice of parameters in the equation, we simulated the intercept α_j from $N(0,5)$, the genotype main effect β_j from $Unif(0,\ 0.2)$, and the factor main effect γ_j from $Unif(0, 0.3)$ $(j = 1, \dots, J)$. In the covariance matrix Σ in (4), the off-diagonal covariance elements are assigned as σ^2. We performed different simulations assuming $\sigma^2 = 0.01$ (negligibly correlated traits), 0.25 (moderately correlated traits), and 0.5 (strongly correlated traits). For $J = 4$ traits, we conducted additional simulations where we considered a specific covariance matrix that mirrored the observed covariance structure of the lipid-related traits that we studied in the UKBB dataset. Finally, we conducted additional type-I error simulations based directly on our UKBB sample. Specifically, we randomly permuted the UKBB phenotype data (consisting of our four trait outcomes and confounding variables) across subjects and then re-ran SCAMPI on the genome-wide data. We repeated the permutation process four times, which resulted in a total of >1M SCAMPI p-values under the null hypothesis.

For power simulations, we implemented a similar simulation design as for our type-I error simulations but introduced additional parameters to model the effect of the interaction between

SNP and the factor on the simulated traits. Specifically, we generated J traits based on the multivariate normal distribution as presented in Eq. (5):

$$\begin{pmatrix} Y_{i,1} \\ \vdots \\ Y_{i,J} \end{pmatrix} \sim MVN\left(\boldsymbol{\mu} = \begin{bmatrix} \alpha_1 + \beta_1 G_i + \gamma_1 W_i + \delta_1 G_i W_i \\ \vdots \\ \alpha_J + \beta_J G_i + \gamma_J W_i + \delta_J G_i W_i \end{bmatrix} \quad , \quad \Sigma = \begin{bmatrix} 1 & \dots & \sigma^2 \\ \vdots & \ddots & \vdots \\ \sigma^2 & \dots & 1 \end{bmatrix} \right) \tag{5}$$

δ_j $(j = 1, \dots, J)$ in equation (5) represents the interaction effect of the SNP and factor on trait j. For a given simulation scenario, we vary the percentage of traits that possess such an interaction (i.e. the sparsity of the interaction signal) among the values 25%, 50%, 75%, and 100%. For those traits with an interaction effect, we vary the value of δ_j across a range of values from 0.01 to 0.50 to study how the power trends change as δ_j increases for each scenario. The settings for the number of traits, MAF, α_j, β_j align with those in the Type I error simulations. However, γ_j is held at fixed values for all traits instead of being simulated from a distribution. Without loss of generality, this approach eliminates the potential for power fluctuations arising from the randomness in γ_j. We simulated the results for various combinations under different parameter sets. To illustrate the overall pattern of the power simulation, we selected the simulation with $\gamma = 0.05$ and 0.25, and $\sigma^2 = 0.1, 0.3$ and 0.5. For each simulation scenario, we assumed a sample size of 20K and generated 10K replicates for inference.

We chose to benchmark SCAMPI against an enhanced multi-phenotype version of Levene's test that was originally restricted to a single phenotype.[30] This enhanced version is termed as the multivariate Levene's test in our context. The multivariate Levene's test applies Levene's test (described in Supplemental Materials S4) to each trait separately, resulting in J p-values. These J p-values are then aggregated together into an omnibus test using the CCT methodology detailed in the prior section (see Supplemental Figure S4 for an outline of the framework). While this

benchmark examines how variances vary by genotype across different traits, it does not consider difference in correlation patterns among traits that SCAMPI integrates within its framework.

Results

Simulation Studies: Table 1 provides empirical type 1 error rates for SCAMPI summarized at a nominal rate α of 10^{-2} and 10^{-3} across varying numbers of phenotypes, MAF and Σ when γ is simulated from $Unif(0, 0.3)$. As described in Supplemental S5, we focused primarily on studying the empirical type-I error rate at 10^{-3} based on the number of simulations performed and observed that SCAMPI was well calibrated at such a threshold. To examine whether SCAMPI was well calibrated at more stringent thresholds, we studied type-I error rates based on permutation of the UKBB data, which yielded > 1M tests under the null hypothesis. For these null simulations, we observed the type I error rates of SCAMPI to be 1.08×10^{-2}, 1.06×10^{-3} and 9.81×10^{-5} at α of $10^{-2}, 10^{-3}$ and 10^{-4}, respectively. SCAMPI p-values generally followed the same pattern as p-values of other statistical methodology that employs CCT.[39; 57-59]

We assessed the power of SCAMPI in different scenarios. Figure 2 provides representative power results at a genome-wide significance threshold of 6.25×10^{-8} (based on a multiple-testing correction for the total number of ~800,000 SNPs in the UKBB) assuming *J*=4 traits and a correlation matrix that mirrored the observed correlation structure of the lipid-related traits that we studied in the UKBB dataset. Figure 2 is comprised of four sub-figures, with each sub-figure presenting simulation results and assuming a different level of sparsity for the interaction effect among the traits modeled. For example, Figure 2a assumes the test SNP has an interaction effect with only one of the four traits, while Figure 2d assumes the test SNP has an interaction effect on all four traits. Within each sub-figure, the yellow solid line represents the power of SCAMPI while the dashed green line represents the power of Multivariate Levene's test. Within each sub-figure,

results show, as expected, that the power of both SCAMPI and Multivariate Levene's test increases as the magnitude of the interaction effect increases. Further, the power of each method increases as the number of traits the SNP has an interaction effect with increases (or, similarly, the sparsity of the interaction effect decreases). However, across all four sub-figures, SCAMPI consistently shows improved power over Multivariate Levene's test. We note that such improved power of SCAMPI over Multivariate Levene's test holds even when the SNP has an interaction effect on only one of the traits under study (Figure 2a), which suggests that the inclusion of traits with no interaction effects still contributes valuable information to the SCAMPI test via their correlation with the trait that does have an interaction effect. We do see that, in Figure 2b and Figure 2c, SCAMPI experiences a pattern at $\delta=0.4$ and $\delta=0.25$, respectively, where power dips slightly at the parameter value; this pattern emerges under conditions where interaction effects are present in multiple, but not all, traits. It results from randomly assigning interaction effects to a subset of traits, provided that the pairwise correlation among the traits are distinct. While the Multivariate Levene's test does not exhibit this behavior (since it only considers the variance of the traits under study), we find that SCAMPI is still more powerful in these situations. We also overlay the power curve of SCAMPI and Multivariate Levene's test with varying sparsity for better visualization in the same plot in Supplemental Figure S5.

In addition to the power simulations inspired by the UKBB, Supplemental Figure S6 provides power results for SCAMPI and multivariate Levene's test under a broader range of models that vary the number of traits considered, the sparsity of the interaction effect, the correlation among traits, and the main effect of the variable interacting with genotype. Overall, we find the power of SCAMPI increases with a decrease in the sparsity of the interaction effect, a decrease in the trait correlation, and an increase in the effect size of the interaction variable.

Assuming these three inputs are fixed, we find that the power of SCAMPI increases as the number of traits modeled increases. Regarding the power comparisons between SCAMPI and the Multivariate Levene's test under this broader range of models, Supplemental Figure S6 also reaffirms the trends observed in our UKBB-inspired power simulations. Across the spectrum of scenarios tested, SCAMPI consistently exhibited superior performance when compared to the Multivariate Levene's test, largely because the former method accounts for correlation among traits that the latter method ignores.

Application to UKBB: Figure 3 provides the Manhattan plot of SCAMPI results for detecting interaction effects on four lipid-related traits. SCAMPI identified 210 SNPs across 68 genes and intergenic regions at a study-wide significance level ($\alpha = 1.67 \times 10^{-7}$, i.e., multiple comparison correction for 300,000 SNPs). Table 2 highlights the SNPs with the smallest SCAMPI p-value on each chromosome from the 210 SNPs. A comprehensive list of the 210 SNPs is available in the Supplemental Table S1. The Q-Q plot for SCAMPI (Supplemental Figure S7) shows no evidence of inflation. SCAMPI is an omnibus test that, by aggregating p-values (outputs of Step 3 in Figure 1) from association tests of trait correlation, pinpoints the specific traits that influence the overall signal. Thus, for every lead SNP in Table 2, we examined the p-values linked to each trait variance and cross-trait correlation at a genome-wide significance threshold of 1.67×10^{7}. Significant variance and correlation terms among traits are noted in the "Significant Variance/Correlation Components" column of the Table. For example, SNP rs7528419 on CELSR2 is significantly associated with the correlation of triglycerides and LDL, as well as the variance of LDL alone, suggesting the SNP may have an interaction effect with other genetic or environmental factors on these two specific traits that merit further investigation.

We also cross-referenced our findings in Table 2 with PheWAS results based on the GWAS Catalog or UK Biobank from the Open Targets Platform (v22.10) , which confirmed many of our initial findings.[60] For instance, SNP rs738409 in PNPLA3 (which SCAMPI identified to be associated with the correlation of triglycerides and BMI as well as triglyceride variance) is reported by Open Targets Platform to be significantly linked with BMI. These results of the lead SNPs are cross listed in the "PheWAS" column of Table 2. Beyond the lead SNPs, Supplemental Table S2 includes the p-values for all correlation components related to the 210 SNPs.

Overall, SCAMPI identified several established lipid- and BMI-related genes that also demonstrate potential interaction effects. For example, *APOC1*, which contained the smallest SCAMPI p-value ($p = 8.1 \times 10^{-61}$), has pleiotropic effects on lipid metabolism, influencing various processes through its actions on lipoprotein receptors and enzyme activity modulation. By controlling the lipids plasma level, the influence of *APOC1* spans several disease areas, including cardiovascular physiology, inflammation, immunity, sepsis, diabetes, cancer, viral infectivity, and cognition.[61] Furthermore, *CETP*, which contained a SNP demonstrating a possible interaction effect with HDL ($p = 7.61 \times 10^{-39}$), may prevent plaque buildup and protect from atherosclerotic cardiovascular disease.[62] There are also mixed results regarding the modifying effects of *CETP* on cardiovascular events.[63-65] Another top gene identified by SCAMPI was *LIPC*. Evidence suggests the *LIPC* promoter polymorphism (T-514C) affects the activity of Hepatic lipase (HL) and, in concert with other factors, modifies the therapeutic response in coronary artery disease (CAD) patients, with those having the CC genotype benefiting the most from intensive lipid-lowering treatments due to their predisposition to high HL activity and smaller, denser LDL particles.[66] SCAMPI also identified SNPs in *CELSR2* with interaction effects predominantly on lipids. Research has shown *CELSR2* deficiency impacts intracellular Ca^{2+} levels, possibly due to

compromised endoplasmic reticulum (ER) function and unfolded protein response (UPR). The depletion of *CELSR2* affects the expression of UPR sensors and the splicing of XBP-1, a critical transcription factor for hepatic lipogenesis, as demonstrated by reactions to various cellular stresses.[67]

Interestingly, SCAMPI identified several SNPs (shown in Supplemental Table S3) exclusively through the correlation among traits (such that they were not detected by the multivariate Levene's test that only considered variance terms). Noteworthy among these are rs2228603 (*NCAN*), rs58542926 (*TM6SF2*), and rs10415849 (*GATAD2A*). For each of these three SNPs, SCAMPI detected a significant effect exclusively via the correlation of BMI and triglycerides (each $p < 10^{-8}$); the SNP was not significantly associated with the variance of either trait and, as such, was not picked up by Levene's test. Prior PheWAS studies show an association between these SNPs and triglycerides.[68-71] A similar pattern is observed for three SNPs in *NECTIN2*; each SNP is associated with the correlation of LDL and HDL (each $p < 10^{-8}$) but not with the variance of either trait. PheWAS analysis previously demonstrated the association of these SNPs with LDL. Beyond PheWAS, we also want to highlight that the SNPs identified by SCAMPI have been implicated in other studies of lipid traits and BMI. For example, numerous studies suggest that rs2228603 and rs58542926 are risk alleles associated with an increased likelihood of liver inflammation and fibrosis that is closely associated with weight change, indeed impacting BMI.[72-74] rs10415849 is significantly associated with α-Tocopherol (one type of vitamin E), which interacts with biological sex to modify BMI.[75] The two SNPs rs519113 and rs6859, which are *BCL3-PVRL2-TOMM40* SNPs, imply gene-gene and gene-environment interactions on dyslipidemia, which pathophysiology is characterized by reverse cholesterol transport in HDL metabolism.[76; 77] Even though there are not many direct studies showing the association between

rs3852860 and HDL, rs3852860 is a well-known predictor in Alzheimer's disease, and Alzheimer's disease progressed with HDL change.[78-80]

SCAMPI Analysis in UKBB Adjusting for APOE: In our applied analyses of lipid traits and BMI in the UKBB, the strongest signal detected by SCAMPI was located within *APOC1*, which is in close physical proximity to *APOE*, a gene with established relevance to the lipid traits we examined. Given APOE's prominence as a biomarker in lipid panels,[81] we determined whether the signals we observed at *APOC1* were independent of those at *APOE*. To assess this, we repeated our SCAMPI analyses conditioning on the main and variance effects of *APOE* SNPs. Specifically, we selected all SNPs on *APOE*, located within 45,409,113 and 45,412,532 on chromosome 19, based on the Genome Reference Consortium Human Build 37 (GRCh37). Five SNPs (rs440446, rs769449, rs769450, rs429358, and rs7412) within this region passed the SNP level QC. We adjusted for the effects of the five *APOE* SNPs on the phenotypic outcomes' mean and variance and then reapplied the SCAMPI methodology. We note that the sample size for our adjusted SCAMPI analysis dropped from 288,709 samples to 241,167 samples due to missing genotypes at the five *APOE* SNPs.

We provide the Manhattan and Q-Q plots for the APOE-adjusted SCAMPI analyses in Supplemental Figure S8. Overall, SCAMPI identified 150 SNPs (see Supplemental Table S4) that remained significant after adjusting for *APOE* genotypes. Our original top hits in *APOC1* remain significant after adjusting for *APOE* genotypes (minimum $p = 3.35 \times 10^{-38}$), which suggests an independent relationship between this gene and lipid traits. This underscores the potential for *APOC1* to be a locus of interest in interaction analyses, with implications for lipid metabolism and associated phenotypes. We note that the initial UKBB analysis identified *APOC1* as the top gene and *LDLR* as the second top gene on Chromosome 19. Upon adjusting for *APOE*, we note that the

rankings of the two genes switch; the SNP with the lowest SCAMPI p-value is now rs55791371 ($p = 4.11 \times 10^{-48}$), located in an intergenic region near *LDLR*.

Computational Performance: We benchmarked the computational performance of SCAMPI across varying sample sizes and numbers of traits for analyzing a single genotype using the High-Performance Computing (HPC) cluster hosted by Emory University Rollins School of Public Health (RSPH), whose infrastructure consists of 25 nodes: twenty-four equipped with 32 compute cores and 192GB of RAM, and one outlier with 1.5TB of RAM. We provide average computational run times per genotype in Figure 4. For instance, in our applied analysis of UKB data, SCAMPI processed a single genotype in an average of 20.17 seconds for four lipid-related traits with 300,000 participants. In general, computational run time of SCAMPI increased linearly with sample size and exhibited quadratic growth with the number of traits. While using SCAMPI on the RSPH HPC, we distribute the computational workload into one job array with 1,000 simultaneous job instances (1,000 job instances are the maximum allowance per job array on RSPH cluster), which effectively partitions the analysis of 300,000 SNPs into 1000 instances of 300 SNPs each. Figure 4 also depicts the number of hours required to complete analyses under various sample sizes and trait quantities by assigning 1,000 job instances on RSPH HPC. Notably, our computational configuration can complete the UKB analysis in approximately 1.68 hours. The figure also shows that processing times grow only modestly with the expansion of the dataset; for instance, a dataset featuring 8 traits and 300,000 samples is estimated to take about 3.98 hours, underscoring SCAMPI's effectiveness for large-scale genetic analyses. Moreover, for the users who are interested in applying SCAMPI to analyze the UKB imputed dataset of over 90 million SNPs, which has approximately 6,000,000 SNPs after QC using the same QC procedure we have discussed in the previous session,[49] supplemental Figure S9 depicts the number of hours required

to complete analyses of 6,000,000 SNPs under various sample sizes and trait quantities by assigning 1,000 job instances on RSPH HPC. Notably, our computational workload configuration can complete the UKBB analysis in approximately 33.62 hours for 6,000,000 SNPs.

It should be noted that optimizing the HPC system with a more powerful processing configuration could significantly decrease computational time. Enhancements such as increasing CPU count and expanding storage and memory would contribute to this efficiency. Our evaluation of SCAMPI's computational performance on a single genotype, across various sample sizes and trait numbers, also utilized a MacBook Pro with an Apple M1 chip. This analysis, detailed in Supplemental Figure S10 (a)-(b), mirrors the one in Figure 4 and Figure S9, where SCAMPI processed a single SNP for four lipid-related traits among 300,000 participants in an average of 8.65 seconds. An HPC system powered with the M1 chip could presumably and feasibly complete our UKBB analysis, involving 300,000 samples and 4 traits, in just about 0.72 hours. Moreover, it will take 14.41 hours to analyze 6,000,000 SNPs.

Discussion

The observation that narrow-sense heritability estimates of complex traits are often considerably larger when estimated from close relatives than distant relatives points to a potential role of variants with interactive effects on such traits. In this work, we develop our method SCAMPI to help screen for such variants that can then be prioritized for subsequent interaction analyses using standard tools. By studying correlation patterns among multiple traits, we showed using simulated data that SCAMPI has improved power relative to univariate variance-based screening procedures. Like variance-based procedures, SCAMPI does not require the specification of the factor that interacts with the variant to influence the traits under study. This means that users do not need prior knowledge of potential interacting factors, which can often be overlooked,

unavailable, or difficult to collect. Furthermore, while SCAMPI produces an omnibus test to assess whether a SNP has an interactive effect on at least one of the traits under study, the method allows a user to identify the specific traits that are driving the signal by inspection of the individual cross-product p-values that are aggregated to form the omnibus test. The method, implemented in R code, is scalable to biobank-scale data and can handle many phenotypes.

While we developed SCAMPI with the intent of identifying variants harboring interaction effects with other genetic variants or environments, the method generally detects any variants with non-additive effects, which can also include dominance effects or parent-of-origin effects. To help delineate dominance effects from potential gene-gene or gene-environment effects, one can rerun SCAMPI regressing out the dominance effect of the variant in the DGLM model prior to analysis and observing whether the original interaction signal remains. For parent-of-origin testing, one can recode the SCAMPI regression framework to assess whether the trait correlation among heterozygotes is significantly different from the two homozygote categories.[82] We note that the appearance of a variant with a possible interaction effect can also arise if the variant is in linkage disequilibrium (LD) with a nearby variant that has a marginal effect on the traits under study.[32] In this situation, we suggest identifying such variants with marginal effects in LD with the test variant prior to analysis and regressing the effects of such variants out of the DGLM mean model prior to analysis using SCAMPI.

SCAMPI makes a few modeling assumptions that warrant further discussion. By implementing a DGLM model that assumes a Gaussian distribution to standardize traits, the SCAMPI framework inherently assumes the trait values under study follow a multivariate normal distribution. To meet this assumption in the main analysis, we transform the traits to normality using a non-parametric rank-based method, the Inverse Normal Transformation (INT), prior to

SCAMPI analysis. We also explored whether transforming the traits before residualizing on the main effects of genotype and confounders (which we refer to as Direct INT or D-INT) led to different inference from transforming after residualizing (which we refer to as Indirect INT or I-INT) [53] and found no marked difference in results (see Tables S5-S7). Rather than conducting a rank-based inverse normal transformation, we could also explore trait standardization on the original scale using a different form of a DGLM that assumes the trait outcome follows a gamma distribution. An additional SCAMPI assumption is that the sample size is large enough and the minor allele frequency of the tested variant common enough to enable p-value derivation of the cross-product regression test using asymptotic theory. For SCAMPI analysis of less-common variants in modest sample sizes, we recommend deriving the p-values of the cross-product regression tests using resampling procedures (which randomly shuffle genotypes across subjects) rather than relying on asymptotic theory to ensure valid inference.

Our SCAMPI framework complements a recent kernel-based method Latent Interaction Testing (LIT) for interaction testing that used kernel distance covariance techniques to test whether similarity of sample trait correlation patterns correlate with genotype similarity at a test SNP.[38] SCAMPI has practical features that LIT lacks, including the ability to directly assess which phenotypes among those modeled demonstrate interaction effects (as illustrated in Table 2 and Supplemental Table S3). Additionally, because SCAMPI is based on aggregating results across multiple cross-trait regression tests, it can handle missing data more efficiently than LIT (which requires complete information on all traits for inference). To illustrate, suppose we have a sample where N subjects possess information on two phenotypes while only half of these subjects further possess additional information on a third phenotype. For joint analysis of all 3 phenotypes, LIT only considers the N/2 subjects with complete trait data for inference. SCAMPI, on the other hand,

555 can incorporate the remaining N/2 subjects that have only information on phenotypes 1 and 2
556 within its cross-trait statistic. The flexible regression framework that forms the backbone of
557 SCAMPI also enables extensions to perform interaction screening for a variety of other study
558 designs used in genetic projects, including longitudinal and family-based designs. Moreover,
559 SCAMPI can be extended to meta-analysis settings where individual-level data cannot be shared
560 across studies. We will explore these SCAMPI extensions in future work.
561
562 SCAMPI R Package is available for installation on GitHub: https://github.com/epstein-
563 software/SCAMPI
564
565 **Funding Support**
566 This work was supported by NIH grants R01 AG071170 (AJB, SB, DJC, MPE) and R01
567 AG075827 (TSW and APW).
568
569
570
571
572
573
574
575
576
577

Reference

1. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42, 565-569.
2. Ni, G., Moser, G., Wray, N.R., and Lee, S.H. (2018). Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. Am J Hum Genet 102, 1185-1194.
3. Wainschtein, P., Jain, D., Zheng, Z., Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., Psaty, B.M., Kooperberg, C., et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat Genet 54, 263-273.
4. Elks, C.E., den Hoed, M., Zhao, J.H., Sharp, S.J., Wareham, N.J., Loos, R.J., and Ong, K.K. (2012). Variability in the heritability of body mass index: a systematic review and meta-regression. Front Endocrinol (Lausanne) 3, 29.
5. Ridge, P.G., Hoyt, K.B., Boehme, K., Mukherjee, S., Crane, P.K., Haines, J.L., Mayeux, R., Farrer, L.A., Pericak-Vance, M.A., Schellenberg, G.D., et al. (2016). Assessment of the genetic variance of late-onset Alzheimer's disease. Neurobiol Aging 41, 200.e213-200.e220.
6. Seshadri, S., Fitzpatrick, A.L., Ikram, M.A., DeStefano, A.L., Gudnason, V., Boada, M., Bis, J.C., Smith, A.V., Carassquillo, M.M., Lambert, J.C., et al. (2010). Genome-wide analysis of genetic loci associated with Alzheimer disease. Jama 303, 1832-1840.
7. Naj, A.C., and Schellenberg, G.D. (2017). Genomic variants, genes, and pathways of Alzheimer's disease: An overview. Am J Med Genet B Neuropsychiatr Genet 174, 5-26.
8. Naj, A.C., Jun, G., Beecham, G.W., Wang, L.S., Vardarajan, B.N., Buros, J., Gallins, P.J., Buxbaum, J.D., Jarvik, G.P., Crane, P.K., et al. (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet 43, 436-441.
9. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45, 1452-1458.
10. Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., van der Lee, S.J., Amlie-Wolf, A., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat Genet 51, 414-430.
11. Jones, L., Harold, D., and Williams, J. (2010). Genetic evidence for the involvement of lipid metabolism in Alzheimer's disease. Biochim Biophys Acta 1801, 754-761.
12. Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.C., Carrasquillo, M.M., Abraham, R., Hamshere, M.L., Pahwa, J.S., Moskvina, V., et al. (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet 43, 429-435.
13. de la Fuente, J., Grotzinger, A.D., Marioni, R.E., Nivard, M.G., and Tucker-Drob, E.M. (2021). Multivariate Modeling of Direct and Proxy GWAS Indicates Substantial Common Variant Heritability of Alzheimer’s Disease. medRxiv, 2021.2005.2006.21256747.

14. Keller, M.F., Ferrucci, L., Singleton, A.B., Tienari, P.J., Laaksovirta, H., Restagno, G., Chiò, A., Traynor, B.J., and Nalls, M.A. (2014). Genome-wide analysis of the heritability of amyotrophic lateral sclerosis. JAMA neurology 71, 1123-1134.
15. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am J Hum Genet 90, 7-24.
16. Røysamb, E., Moffitt, T.E., Caspi, A., Ystrøm, E., and Nes, R.B. (2023). Worldwide Well-Being: Simulated Twins Reveal Genetic and (Hidden) Environmental Influences. Perspect Psychol Sci 18, 1562-1574.
17. Kempthorne, O. (1955). The Theoretical Values of Correlations between Relatives in Random Mating Populations. Genetics 40, 153-167.
18. Robinson, M.R., English, G., Moser, G., Lloyd-Jones, L.R., Triplett, M.A., Zhu, Z., Nolte, I.M., van Vliet-Ostaptchouk, J.V., Snieder, H., Esko, T., et al. (2017). Genotype–covariate interaction effects and the heritability of adult body mass index. Nature Genetics 49, 1174-1181.
19. Binder, E.B., Bradley, R.G., Liu, W., Epstein, M.P., Deveau, T.C., Mercer, K.B., Tang, Y., Gillespie, C.F., Heim, C.M., Nemeroff, C.B., et al. (2008). Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults. Jama 299, 1291-1305.
20. Bradley, R.G., Binder, E.B., Epstein, M.P., Tang, Y., Nair, H.P., Liu, W., Gillespie, C.F., Berg, T., Evces, M., Newport, D.J., et al. (2008). Influence of child abuse on adult depression: moderation by the corticotropin-releasing hormone receptor gene. Arch Gen Psychiatry 65, 190-200.
21. Bailey, J.M., Colón-Rodríguez, A., and Atchison, W.D. (2017). Evaluating a Gene-Environment Interaction in Amyotrophic Lateral Sclerosis: Methylmercury Exposure and Mutated SOD1. Current Environmental Health Reports 4, 200-207.
22. Morahan, J.M., Yu, B., Trent, R.J., and Pamphlett, R. (2007). A gene–environment study of the paraoxonase 1 gene and pesticides in amyotrophic lateral sclerosis. NeuroToxicology 28, 532-540.
23. Dunn, A.R., O'Connell, K.M.S., and Kaczorowski, C.C. (2019). Gene-by-environment interactions in Alzheimer's disease and Parkinson's disease. Neuroscience & Biobehavioral Reviews 103, 73-80.
24. McAllister, K., Mechanic, L.E., Amos, C., Aschard, H., Blair, I.A., Chatterjee, N., Conti, D., Gauderman, W.J., Hsu, L., Hutter, C.M., et al. (2017). Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. American Journal of Epidemiology 186, 753-761.
25. Hutter, C.M., Mechanic, L.E., Chatterjee, N., Kraft, P., and Gillanders, E.M. (2013). Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. Genet Epidemiol 37, 643-657.
26. Spiegelman, D. (2010). Approaches to uncertainty in exposure assessment in environmental epidemiology. Annual review of public health 31, 149-163.
27. Aschard, H., Lutz, S., Maus, B., Duell, E.J., Fingerlin, T.E., Chatterjee, N., Kraft, P., and Van Steen, K. (2012). Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Human genetics 131, 1591-1613.
28. Lindström, S., Yen, Y.-C., Spiegelman, D., and Kraft, P. (2009). The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. Human heredity 68, 171-181.

29. Kraft, P., and Aschard, H. (2015). Finding the missing gene–environment interactions. European journal of epidemiology 30, 353-355.
30. Paré, G., Cook, N.R., Ridker, P.M., and Chasman, D.I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. PLoS genetics 6, e1000981.
31. Brown, M.B., and Forsythe, A.B. (1974). Robust Tests for the Equality of Variances. Journal of the American Statistical Association 69, 364-367.
32. Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K.E., Xue, A., Zhang, M., Powell, J.E., Goddard, M.E., Wray, N.R., et al. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. Science Advances 5, eaaw3538.
33. Westerman, K.E., Majarian, T.D., Giulianini, F., Jang, D.-K., Miao, J., Florez, J.C., Chen, H., Chasman, D.I., Udler, M.S., Manning, A.K., et al. (2022). Variance-quantitative trait loci enable systematic discovery of gene-environment interactions for cardiometabolic serum biomarkers. Nature Communications 13, 3993.
34. Zhu, X., Feng, T., Tayo, B.O., Liang, J., Young, J.H., Franceschini, N., Smith, J.A., Yanek, L.R., Sun, Y.V., Edwards, T.L., et al. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. Am J Hum Genet 96, 21-36.
35. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., Furlotte, N.A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet 50, 229-237.
36. O'Reilly, P.F., Hoggart, C.J., Pomyen, Y., Calboli, F.C., Elliott, P., Jarvelin, M.R., and Coin, L.J. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS One 7, e34861.
37. Broadaway, K.A., Cutler, D.J., Duncan, R., Moore, J.L., Ware, E.B., Jhun, M.A., Bielak, L.F., Zhao, W., Smith, J.A., Peyser, P.A., et al. (2016). A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. Am J Hum Genet 98, 525-540.
38. Bass, A.J., Bian, S., Wingo, A.P., Wingo, T.S., Cutler, D.J., and Epstein, M.P. (2024). Identifying latent genetic interactions in genome-wide association studies using multiple traits. Genome Medicine 16, 62.
39. Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E., and 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, 410-421.
40. Liu, Y., and Xie, J. (2020). Cauchy Combination Test: A Powerful Test With Analytic p-Value Calculation Under Arbitrary Dependency Structures. Journal of the American Statistical Association 115, 393-402.
41. Moore, Camille M., Jacobson, Sean A., and Fingerlin, Tasha E. (2020). Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. Human Heredity 84, 256-271.
42. Joo, J., Kwak, M., Chen, Z., and Zheng, G. (2010). Efficiency robust statistics for genetic linkage and association studies under genetic model uncertainty. Statistics in medicine 29, 158-180.
43. Zheng, G., Freidlin, B., and Gastwirth, J.L. (2006). Comparison of robust tests for genetic association using case-control studies. Lecture Notes-Monograph Series, 253-265.

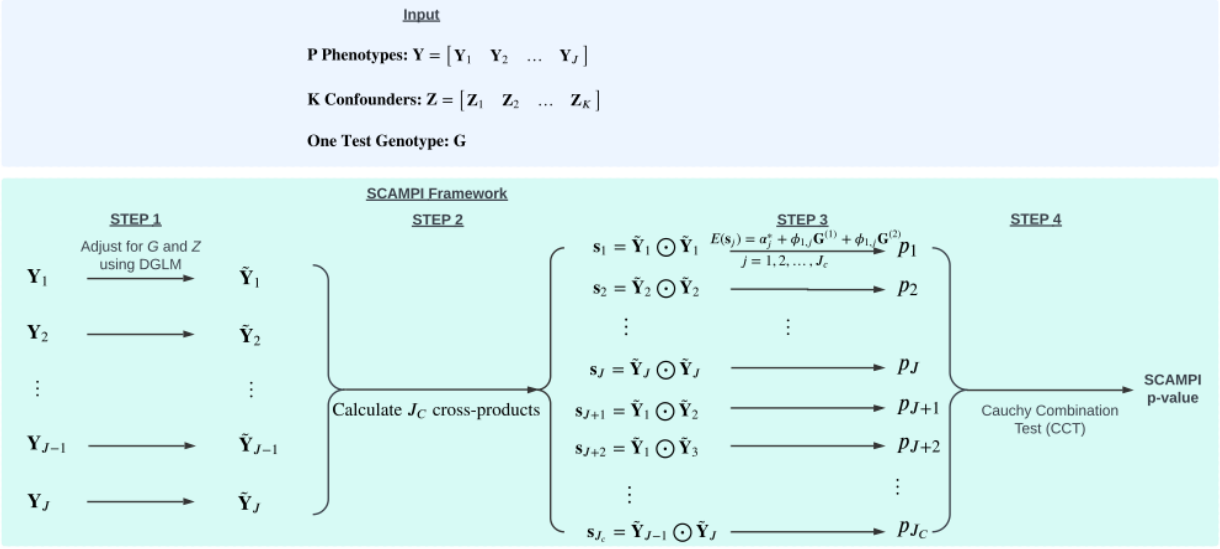
714 44. Joo, J., Kwak, M., Ahn, K., and Zheng, G. (2009). A Robust Genome-Wide Scan Statistic of
715 the Wellcome Trust Case–Control Consortium. Biometrics 65, 1115-1122.
716 45. Lea, A., Subramaniam, M., Ko, A., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I.,
717 Mononen, N., Raitakari, O.T., Ala-Korpela, M., et al. (2019). Genetic and environmental
718 perturbations lead to regulatory decoherence. eLife 8, e40538.
719 46. Musharoff, S., Park, D., Dahl, A., Galanter, J., Liu, X., Huntsman, S., Eng, C., Burchard,
720 E.G., Ayroles, J.F., and Zaitlen, N. (2018). Existence and implications of population
721 variance structure. bioRxiv, 439661.
722 47. Smyth, G.K. (1989). Generalized linear models with varying dispersion. Journal of the royal
723 statistical society series b-methodological 51, 47-60.
724 48. Murphy, M.D., Fernandes, S.B., Morota, G., and Lipka, A.E. (2022). Assessment of two
725 statistical approaches for variance genome-wide association studies in plants. Heredity, 1-
726 10.
727 49. Marderstein, A.R., Davenport, E.R., Kulm, S., Van Hout, C.V., Elemento, O., and Clark,
728 A.G. (2021). Leveraging phenotypic variability to identify genetic interactions in human
729 phenotypes. Am J Hum Genet 108, 49-67.
730 50. Collister, J.A., Liu, X., and Clifton, L. (2022). Calculating Polygenic Risk Scores (PRS) in
731 UK Biobank: A Practical Guide for Epidemiologists. Front Genet 13, 818574.
732 51. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A.,
733 Vukcevic, D., Delaneau, O., and O'Connell, J. (2018). The UK Biobank resource with
734 deep phenotyping and genomic data. Nature 562, 203-209.
735 52. Wigmore, E.M., Clarke, T.-K., Howard, D., Adams, M., Hall, L., Zeng, Y., Gibson, J.,
736 Davies, G., Fernandez-Pujals, A., and Thomson, P.A. (2017). Do regional brain volumes
737 and major depressive disorder share genetic architecture? A study of Generation Scotland
738 (n= 19 762), UK Biobank (n= 24 048) and the English Longitudinal Study of Ageing (n=
739 5766). Translational psychiatry 7, e1205-e1205.
740 53. McCaw, Z.R., Lane, J.M., Saxena, R., Redline, S., and Lin, X. (2020). Operating
741 characteristics of the rank-based inverse normal transformation for quantitative trait
742 analysis in genome-wide association studies. Biometrics 76, 1262-1272.
743 54. Scuteri, A., Sanna, S., Chen, W.-M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R.,
744 Orrú, M., and Usala, G. (2007). Genome-wide association scan shows genetic variants in
745 the FTO gene are associated with obesity-related traits. PLoS genetics 3, e115.
746 55. Barber, M.J., Mangravite, L.M., Hyde, C.L., Chasman, D.I., Smith, J.D., McCarty, C.A., Li,
747 X., Wilke, R.A., Rieder, M.J., and Williams, P.T. (2010). Genome-wide association of
748 lipid-lowering response to statins in combined study populations. PloS one 5, e9763.
749 56. Cade, B.E., Chen, H., Stilp, A.M., Gleason, K.J., Sofer, T., Ancoli-Israel, S., Arens, R., Bell,
750 G.I., Below, J.E., and Bjonnes, A.C. (2016). Genetic associations with obstructive sleep
751 apnea traits in Hispanic/Latino Americans. American journal of respiratory and critical
752 care medicine 194, 886-897.
753 57. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K.,
754 Aslibekyan, S., et al. (2020). Dynamic incorporation of multiple in silico functional
755 annotations empowers rare variant association analysis of large whole-genome
756 sequencing studies at scale. Nat Genet 52, 969-983.
757 58. Li, X., Quick, C., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Selvaraj, M.S., Sun, R., Dey,
758 R., Arnett, D.K., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of

rare variant associations in large whole genome sequencing studies. Nat Genet 55, 154-164.

59. Li, Z., Li, X., Zhou, H., Gaynor, S.M., Selvaraj, M.S., Arapoglou, T., Quick, C., Liu, Y., Chen, H., Sun, R., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. Nat Methods 19, 1599-1611.
60. Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Baker, J., Malangone, C., Lopez, I., Miranda, A., Cruz-Castillo, C., Fumis, L., et al. (2022). The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Research 51, D1353-D1359.
61. Fuior, E.V., and Gafencu, A.V. (2019). Apolipoprotein C1: Its Pleiotropic Effects in Lipid Metabolism and Beyond. Int J Mol Sci 20.
62. Casula, M., Colpani, O., Xie, S., Catapano, A.L., and Baragetti, A. (2021). HDL in Atherosclerotic Cardiovascular Disease: In Search of a Role. Cells 10.
63. Dullaart, R.P., Perton, F., van der Klauw, M.M., Hillege, H.L., Sluiter, W.J., and Group, P.S. (2010). High plasma lecithin: cholesterol acyltransferase activity does not predict low incidence of cardiovascular events: possible attenuation of cardioprotection associated with high HDL cholesterol. Atherosclerosis 208, 537-542.
64. Mabuchi, H., Nohara, A., and Inazu, A. (2014). Cholesteryl ester transfer protein (CETP) deficiency and CETP inhibitors. Molecules and cells 37, 777.
65. Rousset, X., Vaisman, B., Amar, M., Sethi, A.A., and Remaley, A.T. (2009). Lecithin: cholesterol acyltransferase: from biochemistry to role in cardiovascular disease. Current opinion in endocrinology, diabetes, and obesity 16, 163.
66. Deeb, S.S., Zambon, A., Carr, M.C., Ayyobi, A.F., and Brunzell, J.D. (2003). Hepatic lipase and dyslipidemia: interactions among genetic variants, obesity, gender, and diet. Journal of Lipid Research 44, 1279-1286.
67. Tan, J., Che, Y., Liu, Y., Hu, J., Wang, W., Hu, L., Zhou, Q., Wang, H., and Li, J. (2021). CELSR2 deficiency suppresses lipid accumulation in hepatocyte by impairing the UPR and elevating ROS level. The FASEB Journal 35, e21908.
68. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-713.
69. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nat Genet 45, 1274-1283.
70. Barton, A.R., Sherman, M.A., Mukamel, R.E., and Loh, P.-R. (2021). Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. Nat Genet 53, 1260-1269.
71. Prins, B.P., Kuchenbaecker, K.B., Bao, Y., Smart, M., Zabaneh, D., Fatemifar, G., Luan, J.a., Wareham, N.J., Scott, R.A., Perry, J.R.B., et al. (2017). Genome-wide analysis of health-related biomarkers in the UK Household Longitudinal Study reveals novel associations. In Scientific reports. p 11008.
72. Gorden, A., Yang, R., Yerges-Armstrong, L.M., Ryan, K.A., Speliotes, E., Borecki, I.B., Harris, T.B., Chu, X., Wood, G.C., Still, C.D., et al. (2013). Genetic Variation at NCAN Locus Is Associated with Inflammation and Fibrosis in Non-Alcoholic Fatty Liver Disease in Morbid Obesity. Human Heredity 75, 34-43.

805 73. Li, X.Y., Liu, Z., Li, L., Wang, H.J., and Wang, H. (2022). TM6SF2 rs58542926 is related to
806 hepatic steatosis, fibrosis and serum lipids both in adults and children: A meta-analysis.
807 Front Endocrinol (Lausanne) 13, 1026901.
808 74. Ke, P., Xu, M., Feng, J., Tian, Q., He, Y., Lu, K., and Lu, Z. (2023). Association between
809 weight change and risk of liver fibrosis in adults with type 2 diabetes. J Glob Health 13,
810 04138.
811 75. Hamułka, J., Górnicka, M., Sulich, A., and Frąckiewicz, J. (2019). Weight loss program is
812 associated with decrease α-tocopherol status in obese adults. Clinical Nutrition 38, 1861-
813 1870.
814 76. Miao, L., Yin, R.X., Pan, S.L., Yang, S., Yang, D.Z., and Lin, W.X. (2018). BCL3-PVRL2-
815 TOMM40 SNPs, gene-gene and gene-environment interactions on dyslipidemia. Sci Rep
816 8, 6189.
817 77. Urbina, E.M., and Daniels, S.R. (2008). Chapter 14 - Hyperlipidemia. In Adolescent
818 Medicine, G.B. Slap, ed. (Philadelphia, Mosby), pp 90-96.
819 78. Zhou, X., Chen, Y., Mok, K.Y., Kwok, T.C.Y., Mok, V.C.T., Guo, Q., Ip, F.C., Chen, Y.,
820 Mullapudi, N., Giusti-Rodríguez, P., et al. (2019). Non-coding variability at the APOE
821 locus contributes to the Alzheimer's risk. Nat Commun 10, 3310.
822 79. Jia, L., Li, F., Wei, C., Zhu, M., Qu, Q., Qin, W., Tang, Y., Shen, L., Wang, Y., Shen, L., et
823 al. (2020). Prediction of Alzheimer's disease using multi-variants from a Chinese
824 genome-wide association study. Brain 144, 924-937.
825 80. Button, E.B., Robert, J., Caffrey, T.M., Fan, J., Zhao, W., and Wellington, C.L. (2019). HDL
826 from an Alzheimer's disease perspective. Curr Opin Lipidol 30, 224-234.
827 81. Chasman, D.I., Kozlowski, P., Zee, R.Y., Kwiatkowski, D.J., and Ridker, P.M. (2006).
828 Qualitative and quantitative effects of APOE genetic variation on plasma C-reactive
829 protein, LDL-cholesterol, and apoE protein. Genes & Immunity 7, 211-219.
830 82. Head, S.T., Leslie, E.J., Cutler, D.J., and Epstein, M.P. (2023). POIROT: a powerful test for
831 parent-of-origin effects in unrelated samples leveraging multiple phenotypes.
832 Bioinformatics 39.
833
834
835
836
837
838
839
840
841

842 **Figure 1**



843

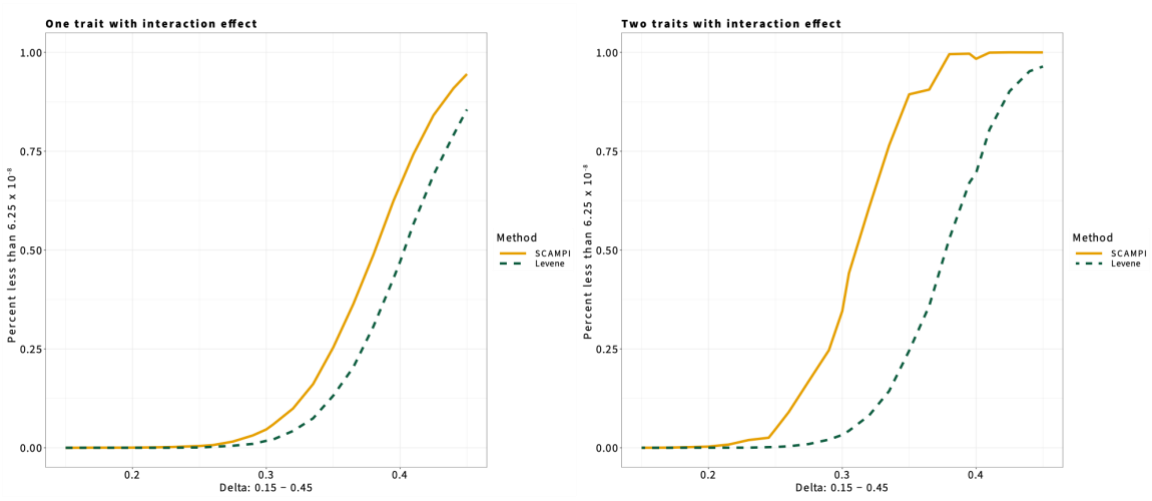
844 **Figure 1. Illustration of the SCAMPI framework.**

845 The SCAMPI framework involves a four-step process that consists of 1) adjustment of phenotypes
846 for genotype and confounders, 2) calculation of cross products from adjusted phenotypes, 3)
847 derivation of p-values from regression tests of cross products on test genotype, and 4) aggregating
848 all p-values using the Cauchy Combination Test (CCT) to derive the final SCAMPI p-value to
849 determine overall significance.

850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865

Figure 2

(a) (b)



(c) (d)

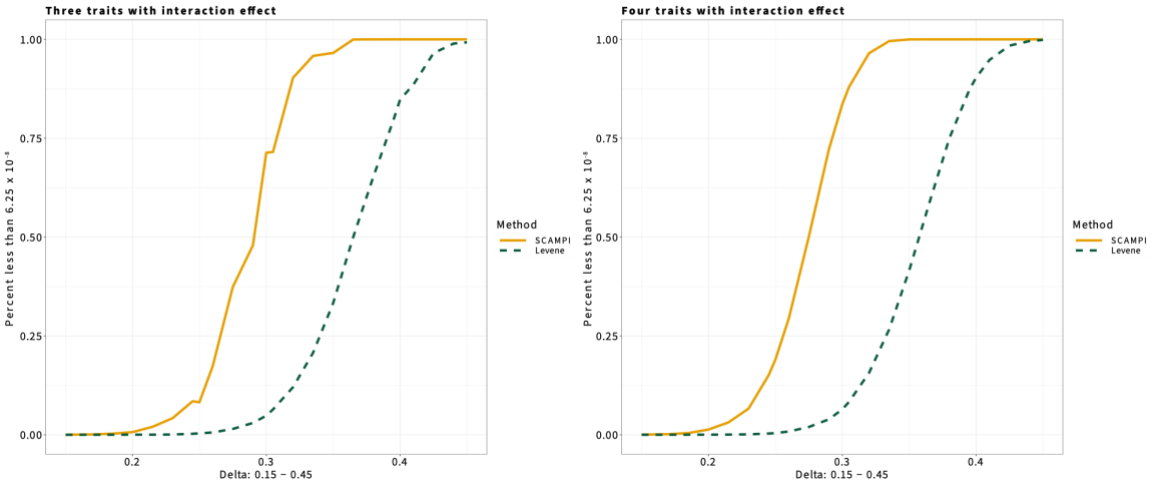
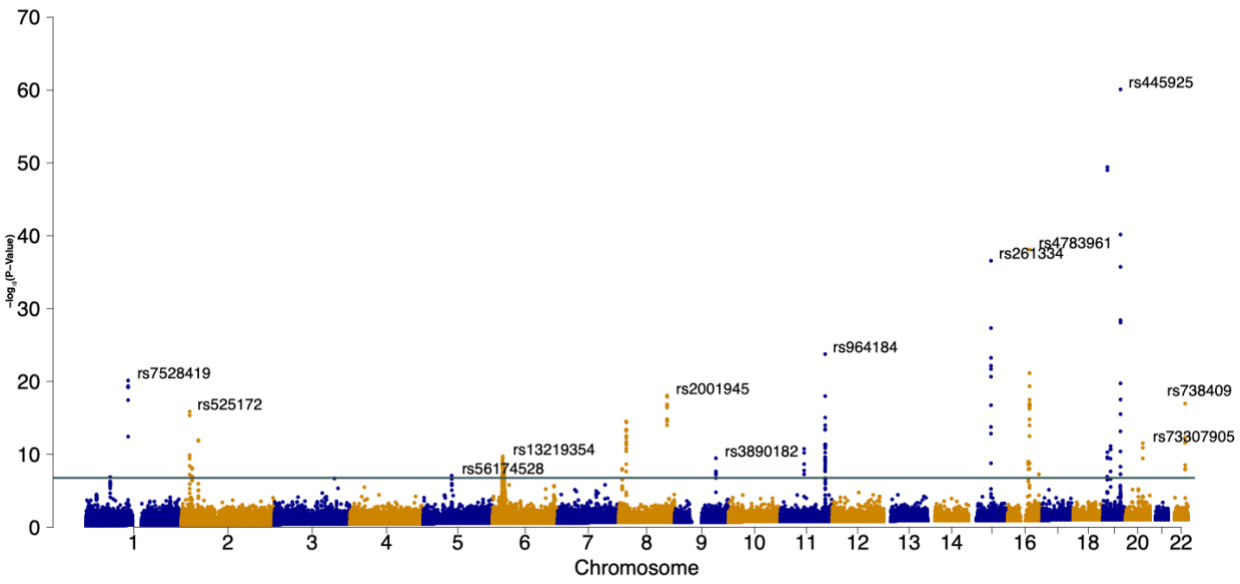


Figure 2. UKBB-inspired power simulation of SCAMPI for four traits

Power of SCAMPI at $\alpha = 6.25 \times 10^{-8}$ for four traits at a sample size of 20,000 with MAF=0.05 and $\gamma = 0.05$. The correlation among the four traits is inspired by correlation among lipid traits considered in our applied UKBB analysis. Yellow solid line represents power of SCAMPI, while dashed green line denotes power of the benchmark Multivariate Levene's test. Sub-figures (a) – (d) examines power when interaction effects exist for one trait, two traits, three traits and four traits, respectively. We analyzed 10,000 replicates under each model.

878 **Figure 3**



879

880 **Figure 3. Genome-wide results on lipid traits in UKBB using SCAMPI**

881 SCAMPI results for detecting latent interaction effects on high-density lipoprotein cholesterol
882 (HDL-C), low-density lipoproteins cholesterol (LDL-C), triglycerides (TGs), and Body Mass
883 Index (BMI). After SNP QC, 288,910 SNPs are included in the analysis with their MAF ≥ 0.05.
884 SCAMPI successfully identified 210 SNPs from 68 genes and intergenic regions at the pre-
885 specified study-wide significance ($\alpha = 1.67 \times 10^{-7}$) represented by the green solid horizontal
886 line. The SNP with the smallest SCAMPI p-value is rs445925 located on *ApoC1*.

887

888

889

890

891

892

Figure 4

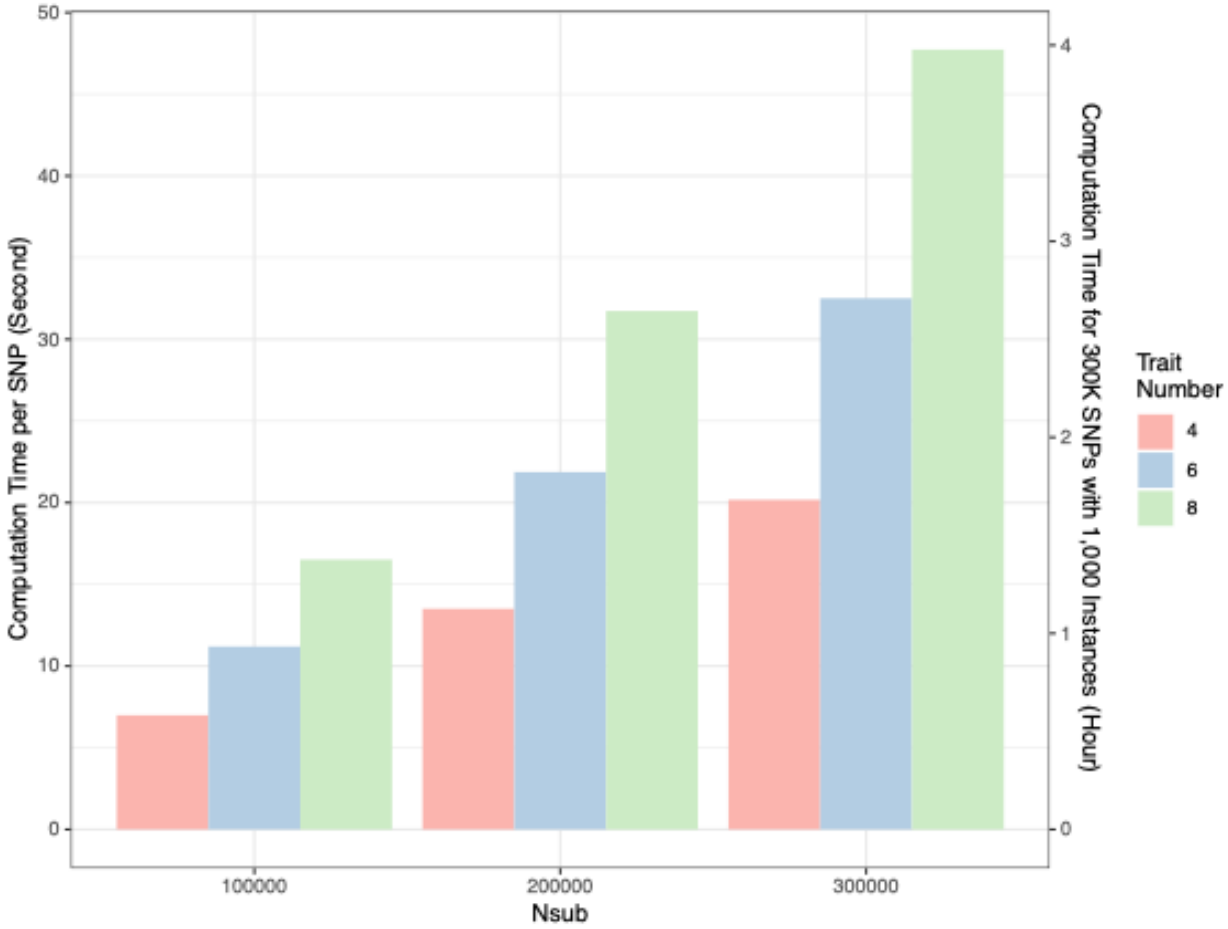


Figure 4. Computational performance of SCAMPI

Computational run time of SCAMPI for different sample sizes and number of traits using High-Performance Computing (HPC) cluster hosted by Emory University Rollins School of Public Health (RSPH). Computational run time is based on average of 1,000 simulations for different scenarios with varying trait number and sample size. The first y-axis in Figure 4 displays the time in seconds to complete SCAMPI for one SNP at different configurations. The second y-axis shows the hours required to complete analyses assuming 1,000 job instances on a high-performance cluster.

904 **Table 1**

N	MAF	Σ	J	LEQ 0.01	LEQ 0.001
3.00E+05	0.05	0.01	2	9.56E-03	1.01E-03
			4	1.00E-02	1.18E-03
			8	1.05E-02	1.13E-03
3.00E+05	0.05	0.25	2	9.96E-03	1.09E-03
			4	1.09E-02	9.30E-04
			8	1.12E-02	1.01E-03
3.00E+05	0.05	0.5	2	1.05E-02	1.16E-03
			4	1.16E-02	1.02E-03
			8	1.26E-02	1.23E-03
3.00E+05	0.25	0.01	2	9.51E-03	1.17E-03
			4	1.03E-02	1.13E-03
			8	1.00E-02	1.14E-03
3.00E+05	0.25	0.25	2	1.01E-02	1.12E-03
			4	1.09E-02	1.06E-03
			8	1.09E-02	9.40E-04
3.00E+05	0.25	0.5	2	1.03E-02	9.80E-04
			4	1.09E-02	1.03E-03
			8	1.28E-02	1.09E-03

905

906 **Table 1. Nominal rate of empirical type 1 error rates for SCAMPI.**

907 The empirical type I error rates for SCAMPI at nominal rates α of 10^{-2} and 10^{-3} in 100,000
908 simulations with 300,000 observations. The result is presented across a range of conditions
909 including varying Minor Allele Frequencies (MAF), numbers of phenotypes (J), and covariance
910 of the phenotypes Σ when γ is simulated from a uniform distribution between 0 and 0.3. The value
911 presented in the 'LEQ 0.01' and 'LEQ 0.001' columns reflect the pre-specified nominal error rates.

912

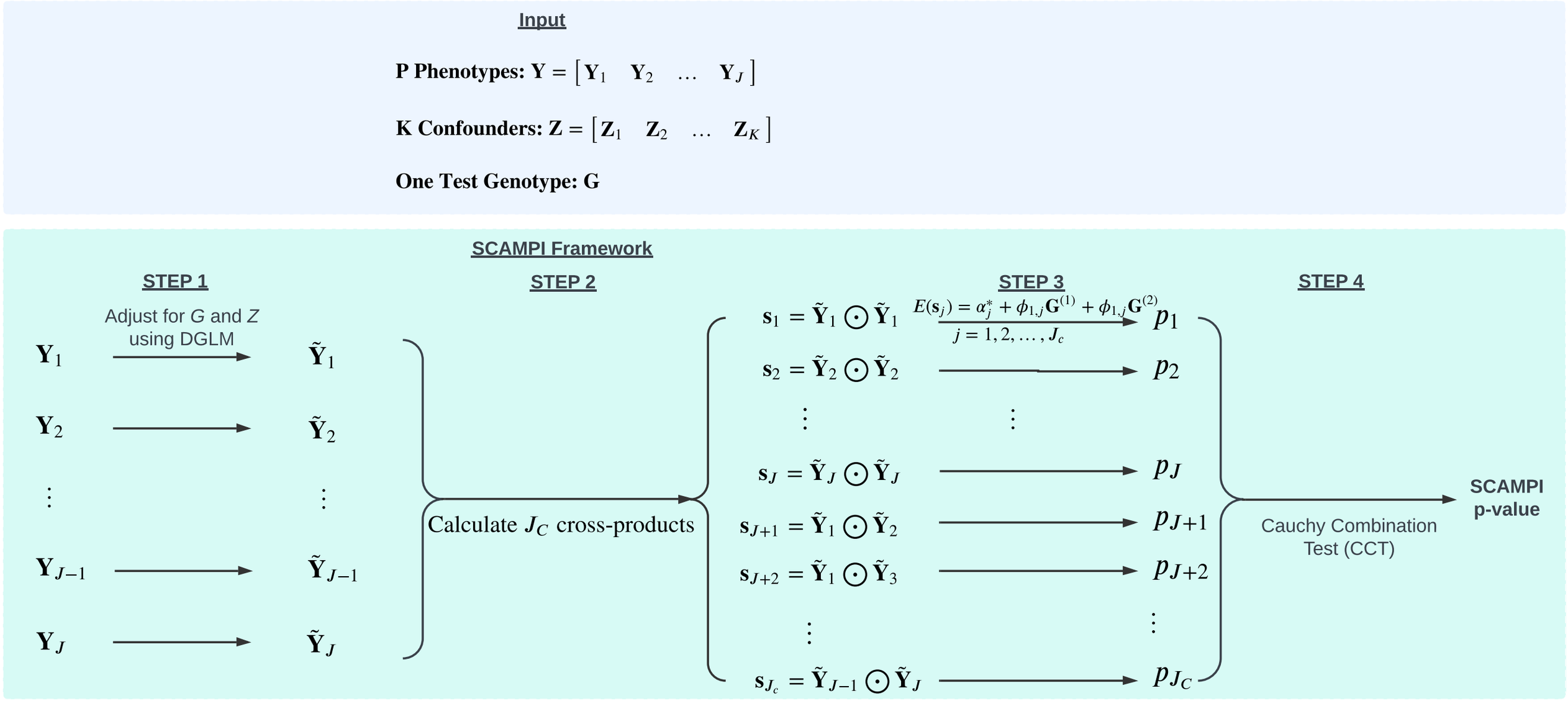
Table 2

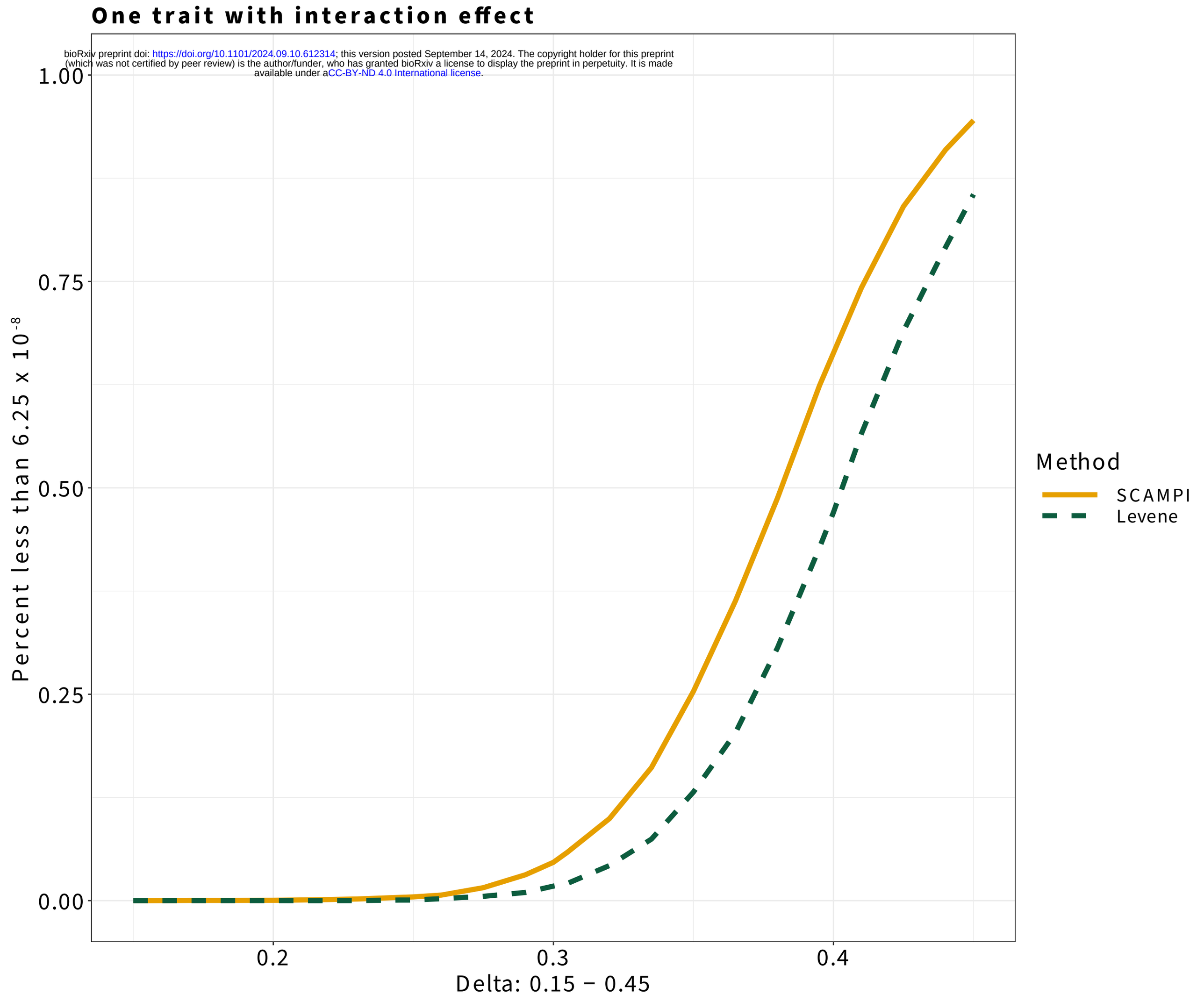
Chr	Pos	Alt	Ref	RS #	Gene	SCAMPI P-value	Significant Variance/Correlation Components	PheWAS
1	**109817192**	**G**	**A**	rs7528419	*CELSR2*	7.33E-21	Corr(TRIG, LDL), Var(LDL)	TRIG, LDL
2	**21382976**	**G**	**T**	rs525172	Intergenic	1.33E-16	Corr(TRIG, HDL), Var(LDL)	TRIG, LDL
5	**74400516**	**C**	**G**	rs56174528	*ANKRD31*	8.26E-08	Var(LDL)	LDL
6	**27185664**	**C**	**T**	rs13219354	*PRSS16*	1.83E-10	Var(BMI)	BMI
8	**126477978**	**C**	**G**	rs2001945	(*TRIB1*)	8.34E-19	Var(LDL)	LDL
9	**107647655**	**A**	**G**	rs3890182	*ABCA1*	3.34E-10	Var(HDL)	HDL
11	**116648917**	**C**	**G**	rs964184	*ZPR1*	1.73E-24	Corr(TRIG, LDL), Corr(TRIG, HDL), Corr(LDL, HDL), Corr(LDL, BMI), Var(Trig), Var(LDL)	TRIG, LDL, HDL
15	**58726744**	**C**	**G**	rs261334	*LIPC; LIPC-AS1*	2.63E-37	Corr(HDL, BMI), Var(TRIG)	TRIG, HDL
16	**56994894**	**A**	**G**	rs4783961	*CETP*	7.61E-39	Var(HDL)	HDL
19	**45415640**	**A**	**G**	rs445925	*APOC1*	8.10E-61	Corr(TRIG, LDL), Corr(TRIG, HDL), Corr(LDL, HDL), Corr(LDL, BMI), Var(Trig), Var(LDL), Var(HDL)	TRIG, LDL, HDL
20	**44545773**	**C**	**A**	rs73307905	(*PLTP*)	2.90E-12	Var(HDL)	HDL
22	**44324727**	**G**	**C**	rs738409	*PNPLA3*	1.07E-17	Corr(TRIG, BMI), Var(TRIG)	BMI

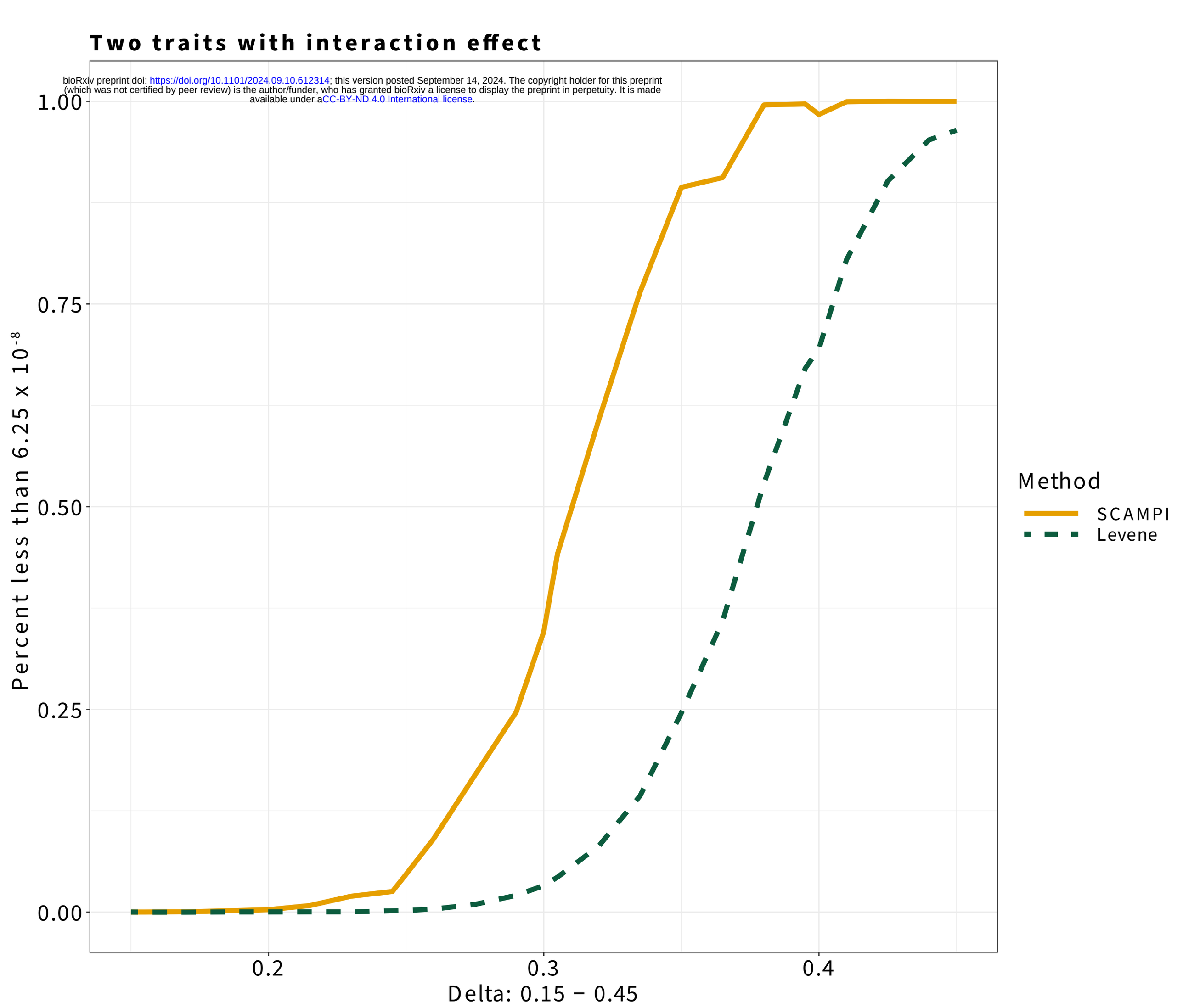
Table 2. The lead SNPs, identified by SCAMPI within each chromosome, implies interaction effects for the four lipid traits in UKBB

SCAMPI identified 12 lead SNPs from 12 chromosomes. The position of the SNPs is based on the Genome Reference Consortium Human Build 37 (GRCh37). Column "Gene" indicates the gene where the SNP locates. Column "SCAMPI P-value" shows the SCAMPI p-value. Column "Significant Variance/Correlation Components" indicates the variance or the correlation components of the four lipids

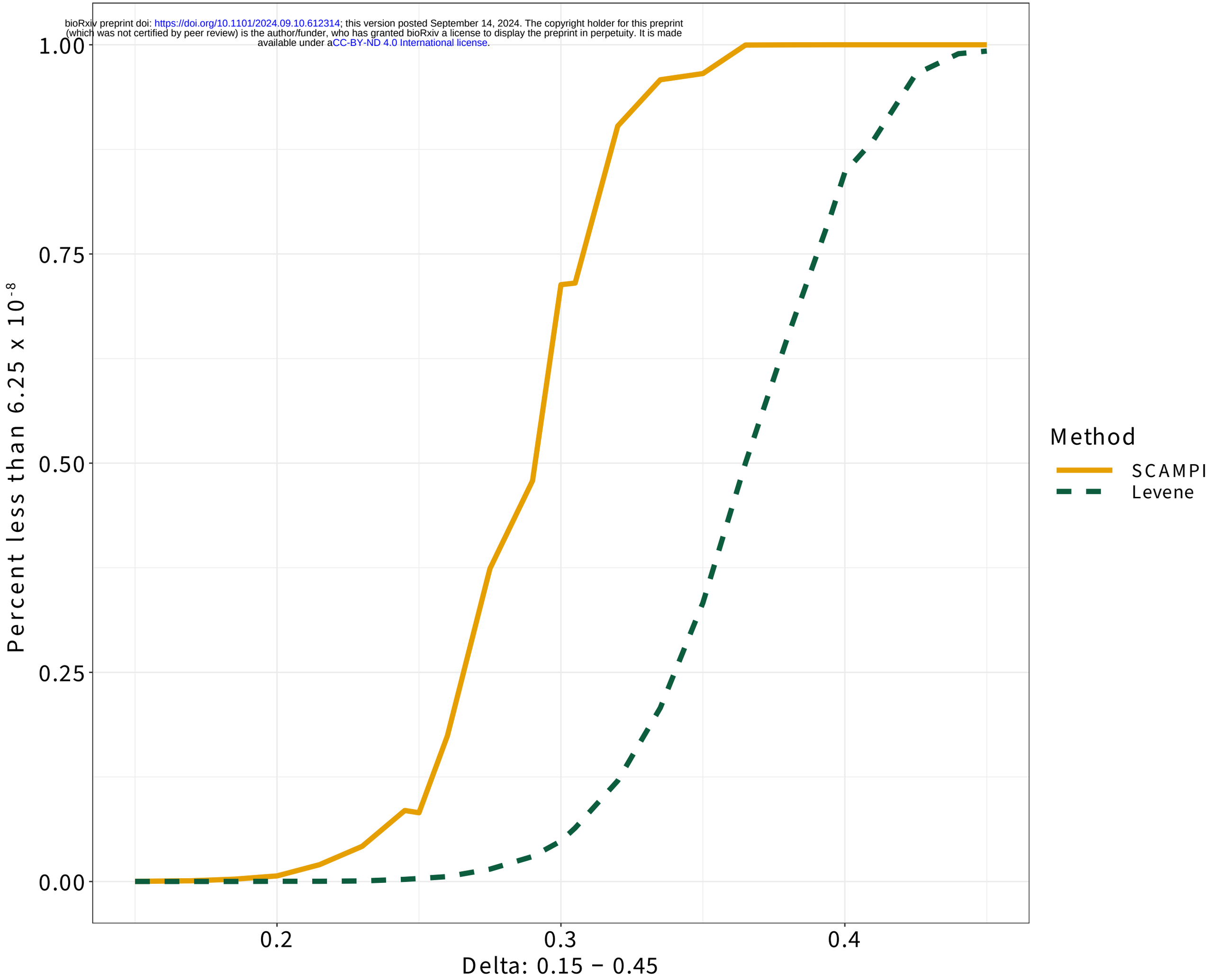
that are significantly associated with the corresponding SNP at the pre-specified study-wide significance ($\alpha = 1.67 \times 10^{-7}$). Column "PheWAS" lists the traits involved in the significant variance and correlation components as noted in column "Significant Variance/Correlation Components", and these traits are also identified to be significant in PheWAS results, which is cross-referenced based on the GWAS Catalog or UK Biobank from the Open Targets Platform.

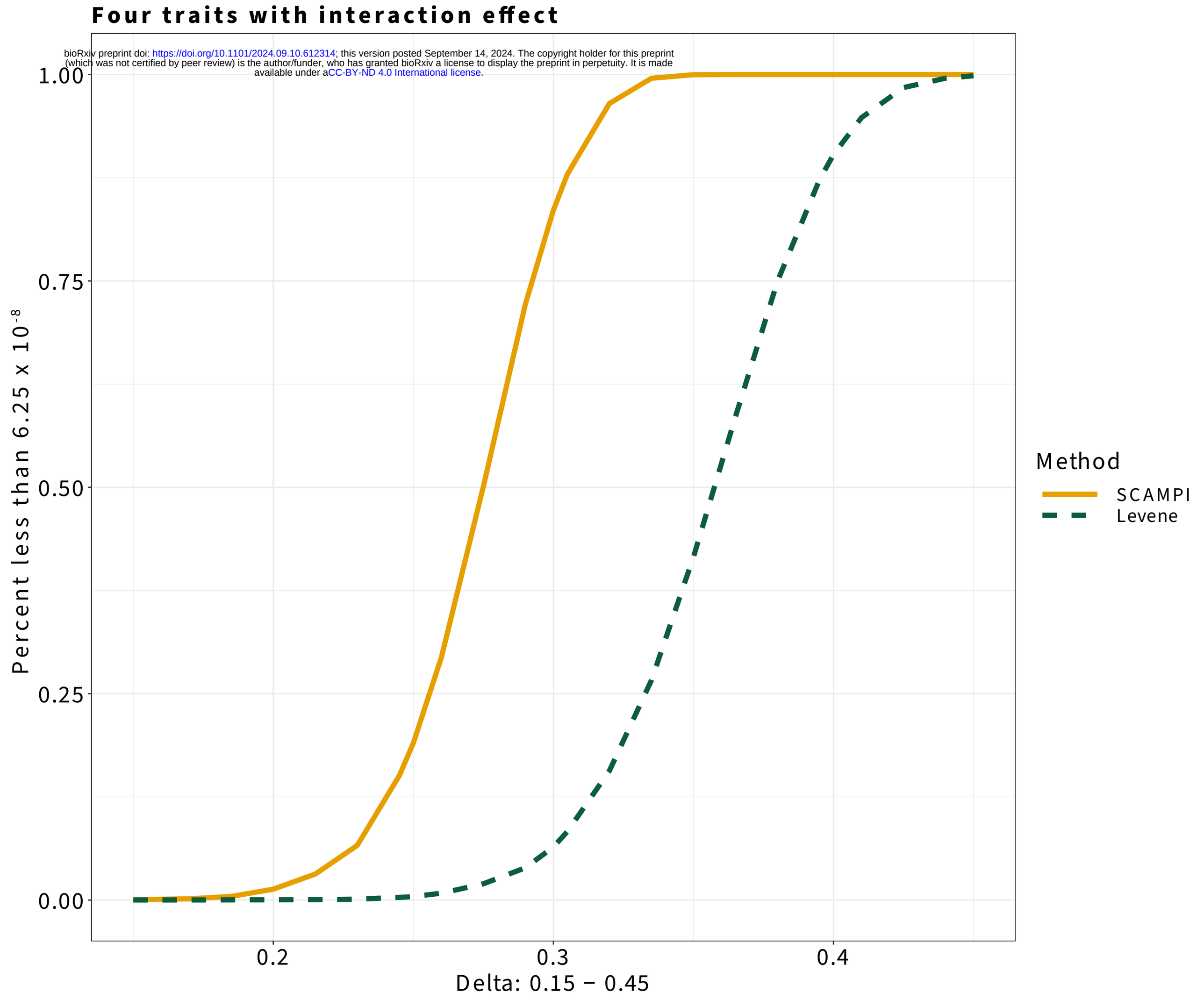

Input
P Phenotypes: $\mathbf{Y} = [\mathbf{Y}_1 \quad \mathbf{Y}_2 \quad \dots \quad \mathbf{Y}_J]$
K Confounders: $\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{Z}_2 \quad \dots \quad \mathbf{Z}_K]$
One Test Genotype: G
SCAMPI Framework
STEP 1
Adjust for G and Z using DGLM
$\mathbf{Y}_1$
$\mathbf{Y}_2$
$\mathbf{Y}_{J-1}$
$\mathbf{Y}_J$
$\tilde{\mathbf{Y}}_1$
$\tilde{\mathbf{Y}}_2$
$\tilde{\mathbf{Y}}_{J-1}$
$\tilde{\mathbf{Y}}_J$
STEP 2
Calculate J_C cross-products
STEP 3
$E(\mathbf{s}_j) = \alpha_j^* + \phi_{1,j}\mathbf{G}^{(1)} + \phi_{1,j}\mathbf{G}^{(2)}$
$j = 1, 2, \dots, J_c$
$\mathbf{s}_1 = \tilde{\mathbf{Y}}_1 \odot \tilde{\mathbf{Y}}_1$
$\mathbf{s}_2 = \tilde{\mathbf{Y}}_2 \odot \tilde{\mathbf{Y}}_2$
$\mathbf{s}_J = \tilde{\mathbf{Y}}_J \odot \tilde{\mathbf{Y}}_J$
$\mathbf{s}_{J+1} = \tilde{\mathbf{Y}}_1 \odot \tilde{\mathbf{Y}}_2$
$\mathbf{s}_{J+2} = \tilde{\mathbf{Y}}_1 \odot \tilde{\mathbf{Y}}_3$
$\mathbf{s}_{J_c} = \tilde{\mathbf{Y}}_{J-1} \odot \tilde{\mathbf{Y}}_J$
p_1
p_2
p_J
p_{J+1}
p_{J+2}
p_{J_C}
STEP 4
Cauchy Combination Test (CCT)
SCAMPI p-value

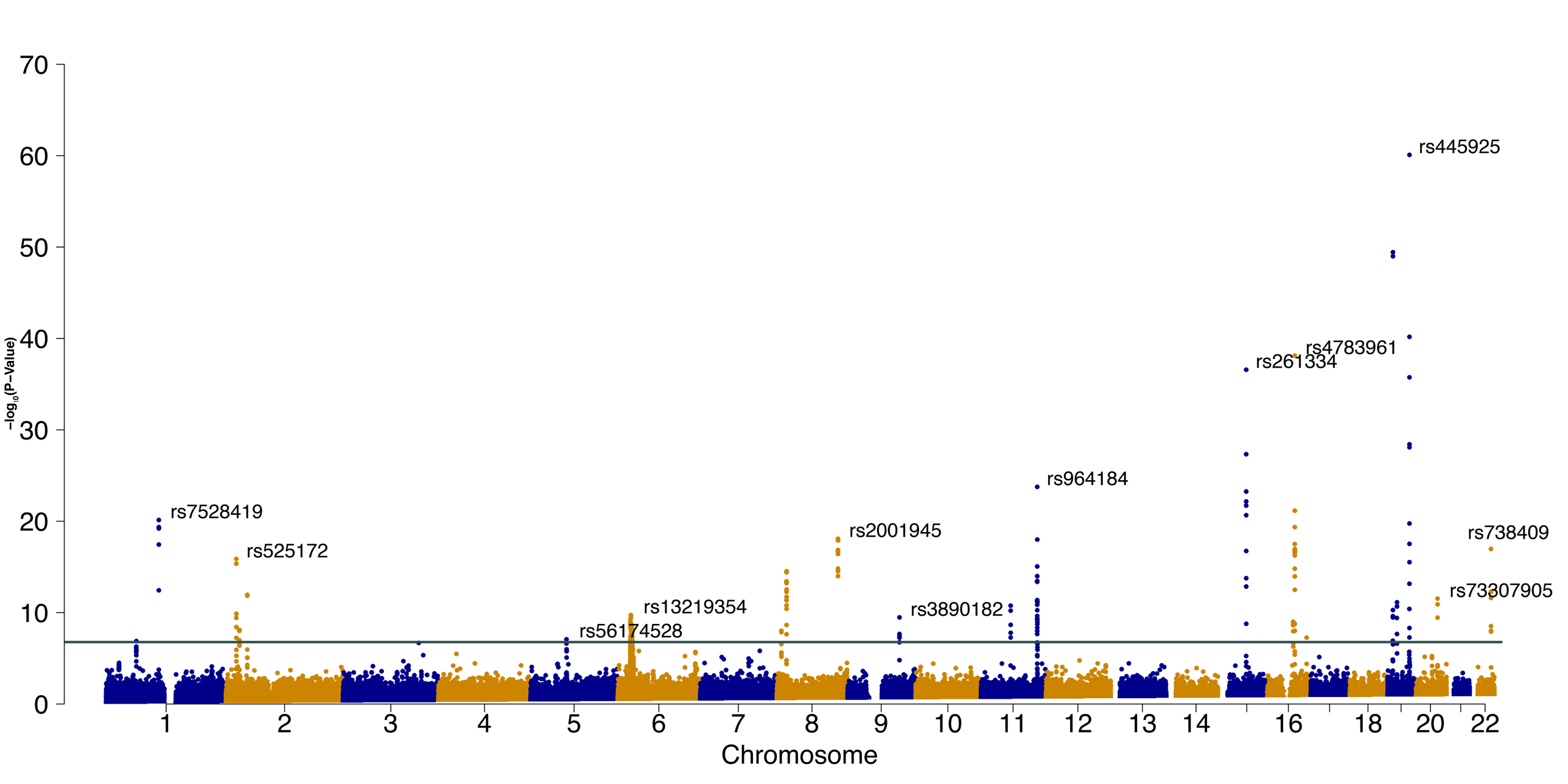


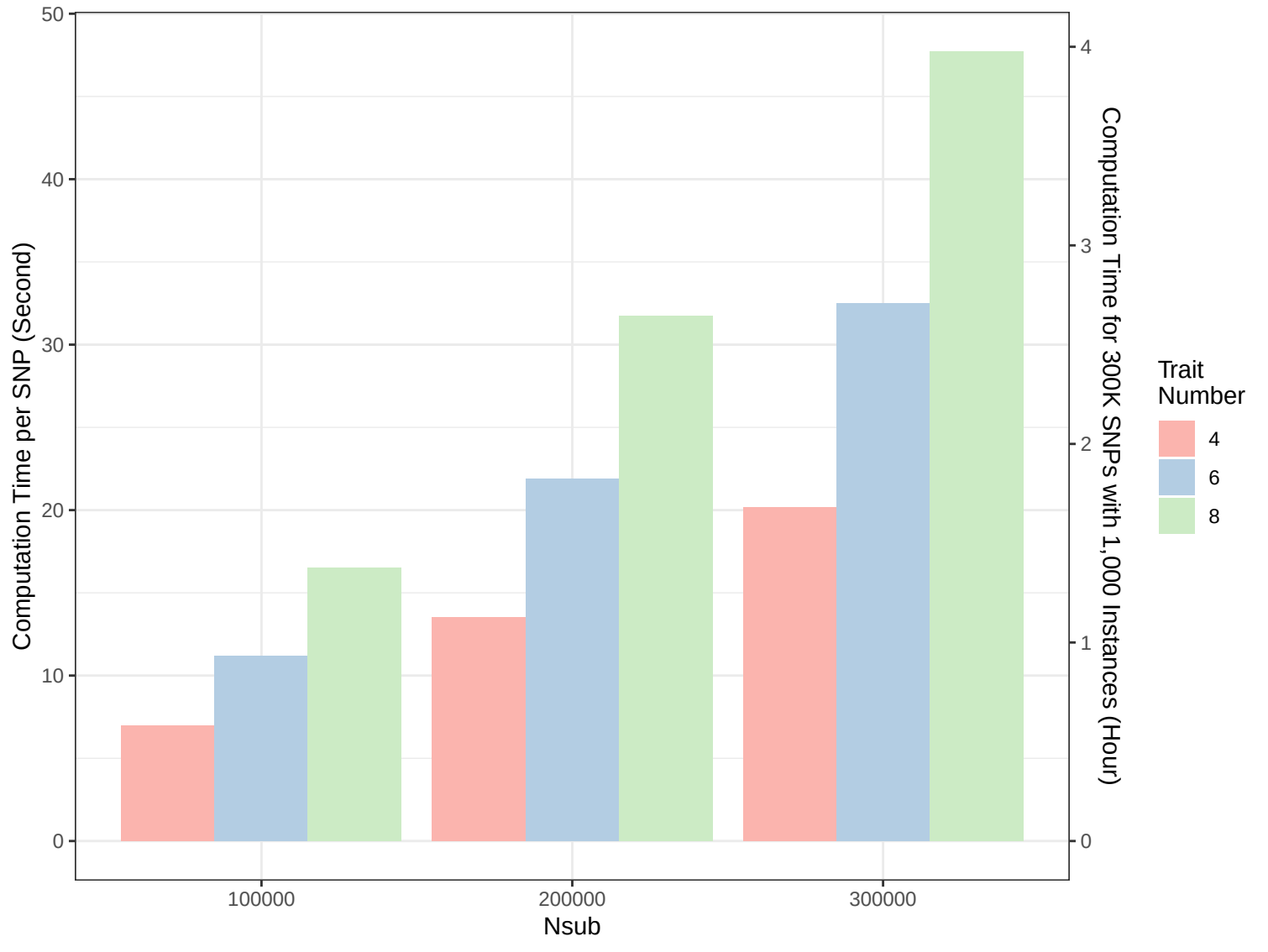


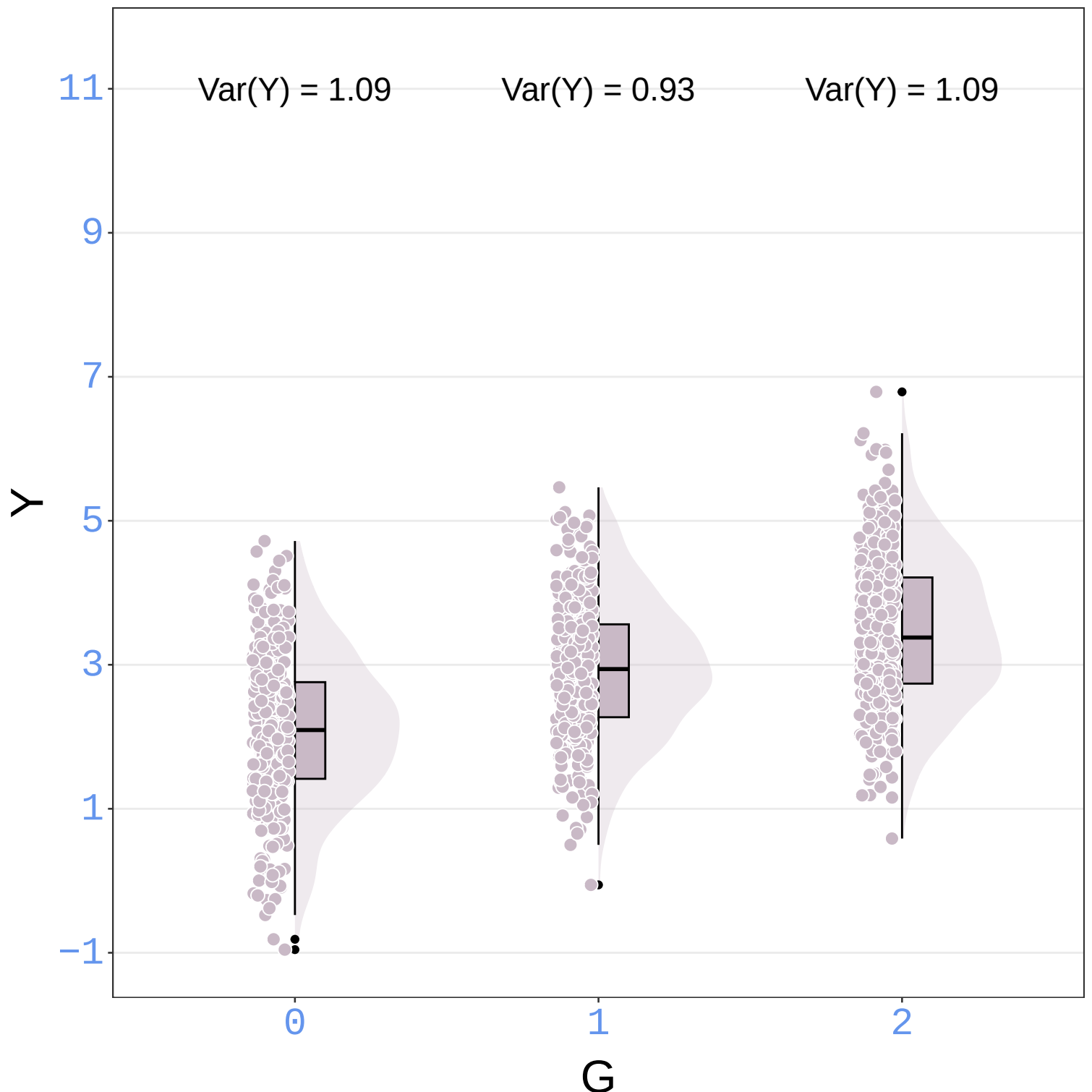
Three traits with interaction effect

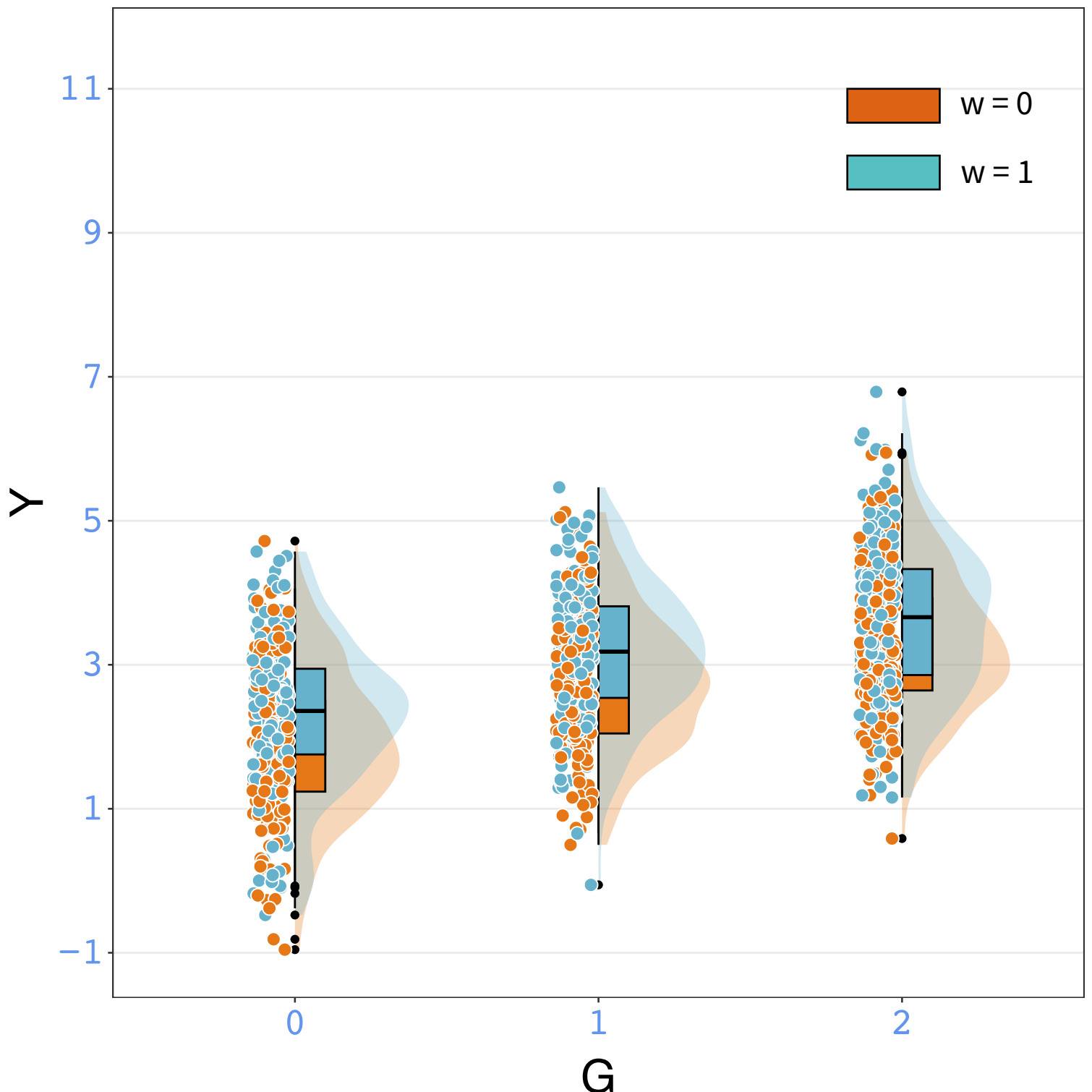


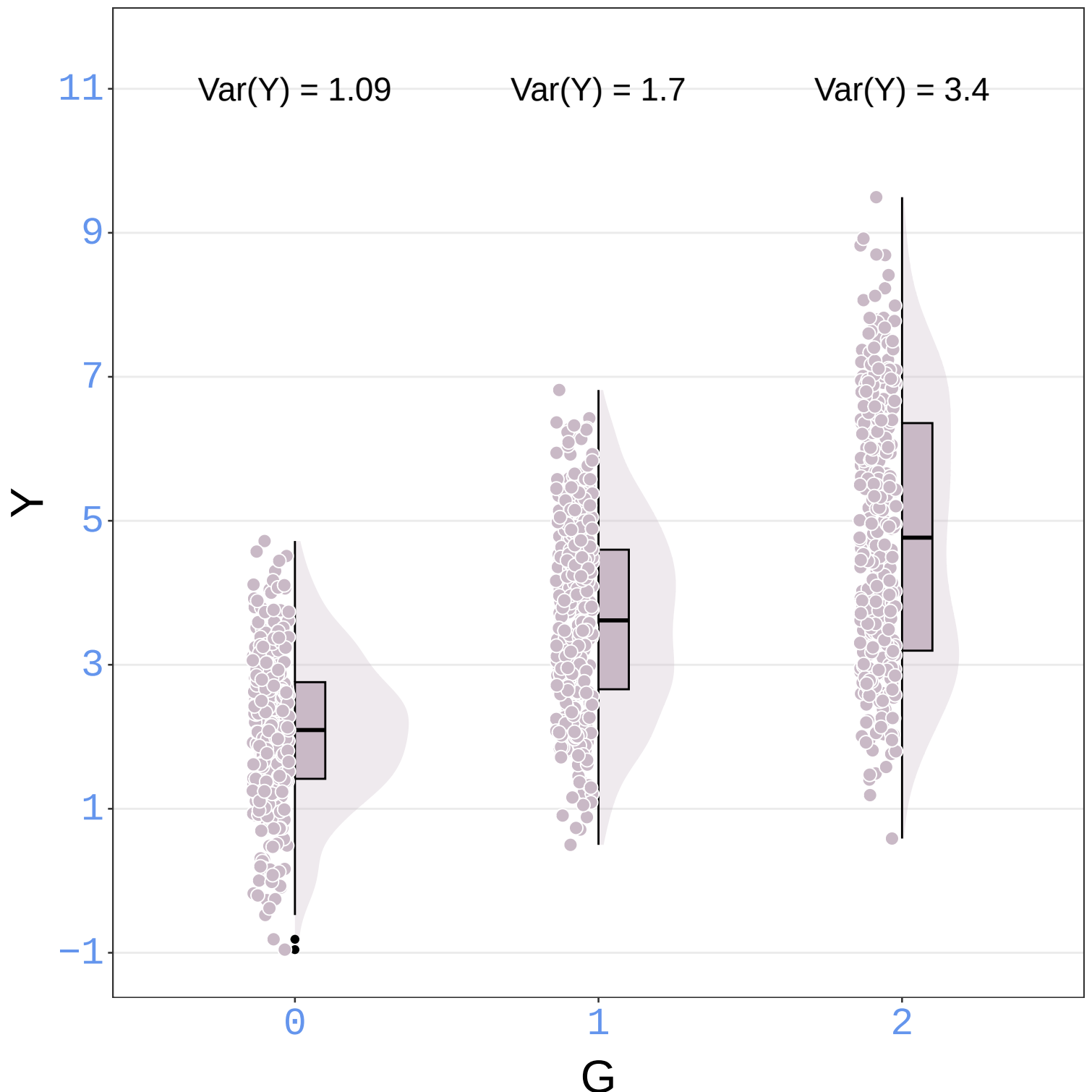


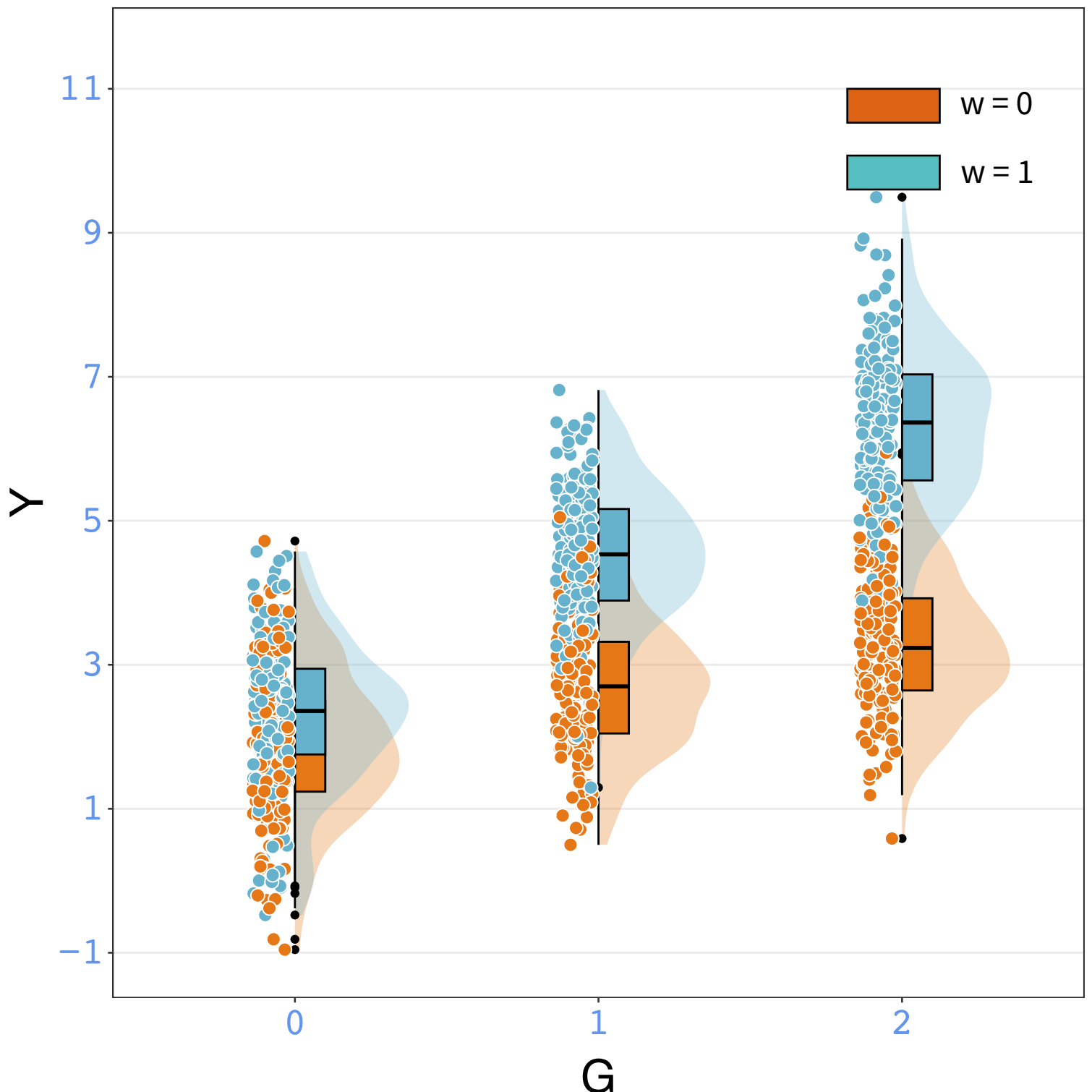

70
60
50
40
30
20
10
0
−log10(P−Value)
rs445925
rs4783961
rs261334
rs964184
rs7528419
rs525172
rs2001945
rs738409
rs73307905
rs13219354
rs3890182
rs56174528
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
18
20
22
Chromosome


Computation Time per SNP (Second)
Computation Time for 300K SNPs with 1,000 Instances (Hour)
Nsub
100000
200000
300000
0
10
20
30
40
50
1
2
3
4
Trait Number
4
6
8


Var(Y) = 1.09
Var(Y) = 0.93
Var(Y) = 1.09
11
9
7
5
3
1
−1
Y
0
1
2
G

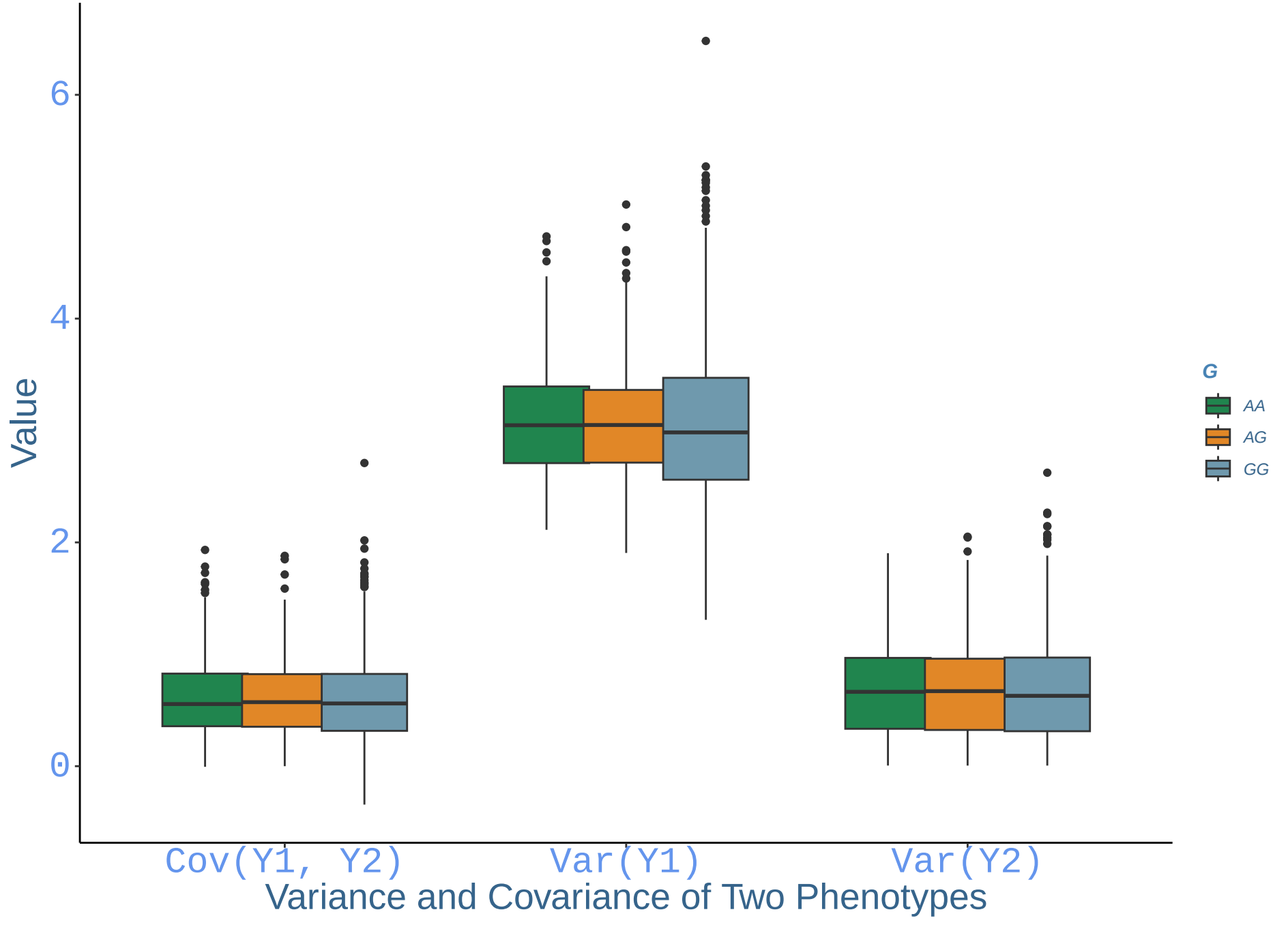

11
9
7
5
3
1
−1
Y
w = 0
w = 1
0
1
2
G


Var(Y) = 1.09
Var(Y) = 1.7
Var(Y) = 3.4
11
9
7
5
3
1
−1
Y
0
1
2
G


11
9
7
5
3
1
−1
Y
0
1
2
G
w = 0
w = 1

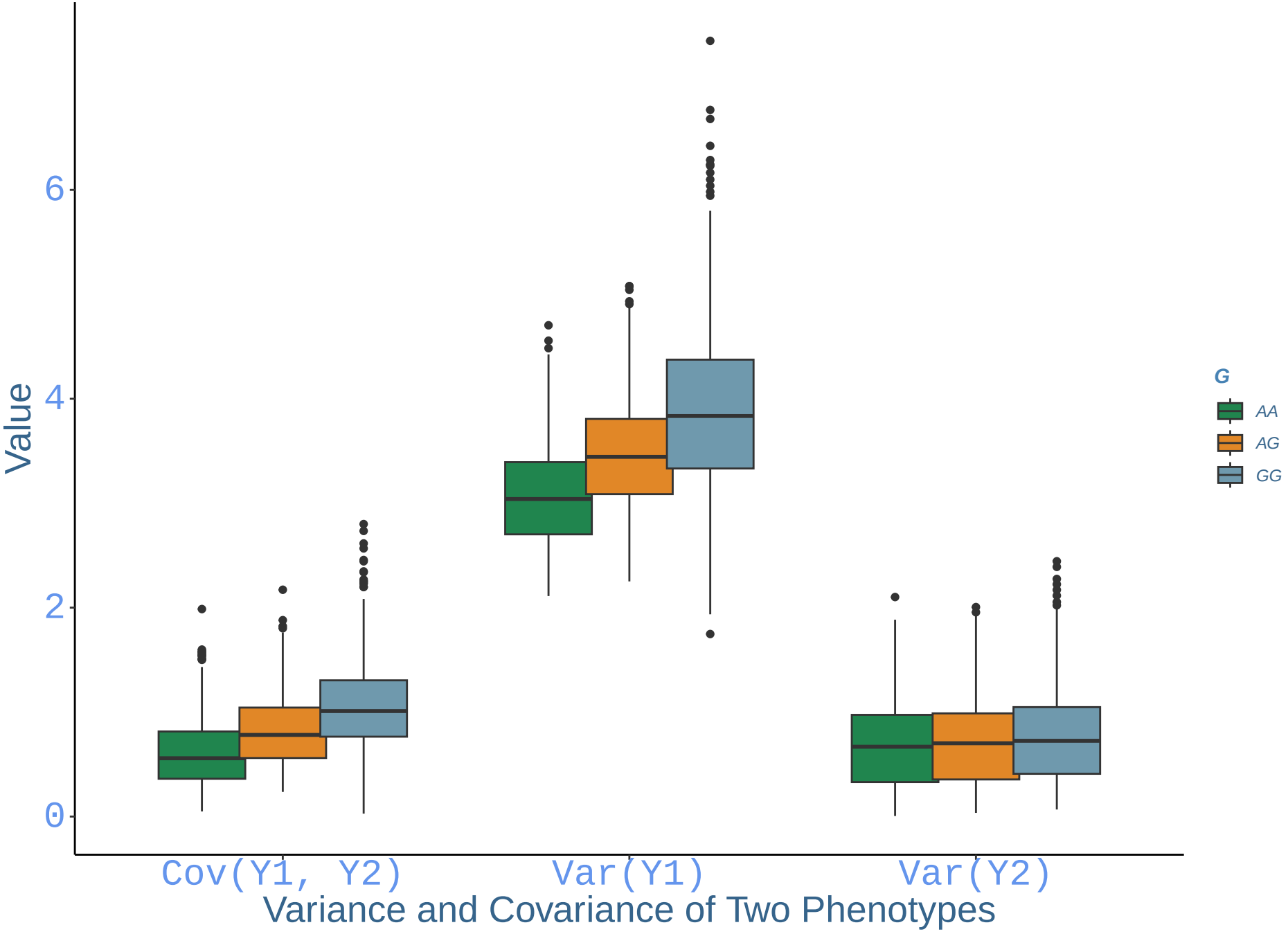
Variance and Covariance Distribution of Two Phenotypes Across Genotype Categories in 1,000 Simulated Datasets Without Interaction Effects

The differential covariance of the two phenotypes adds power while studying the interaction effects

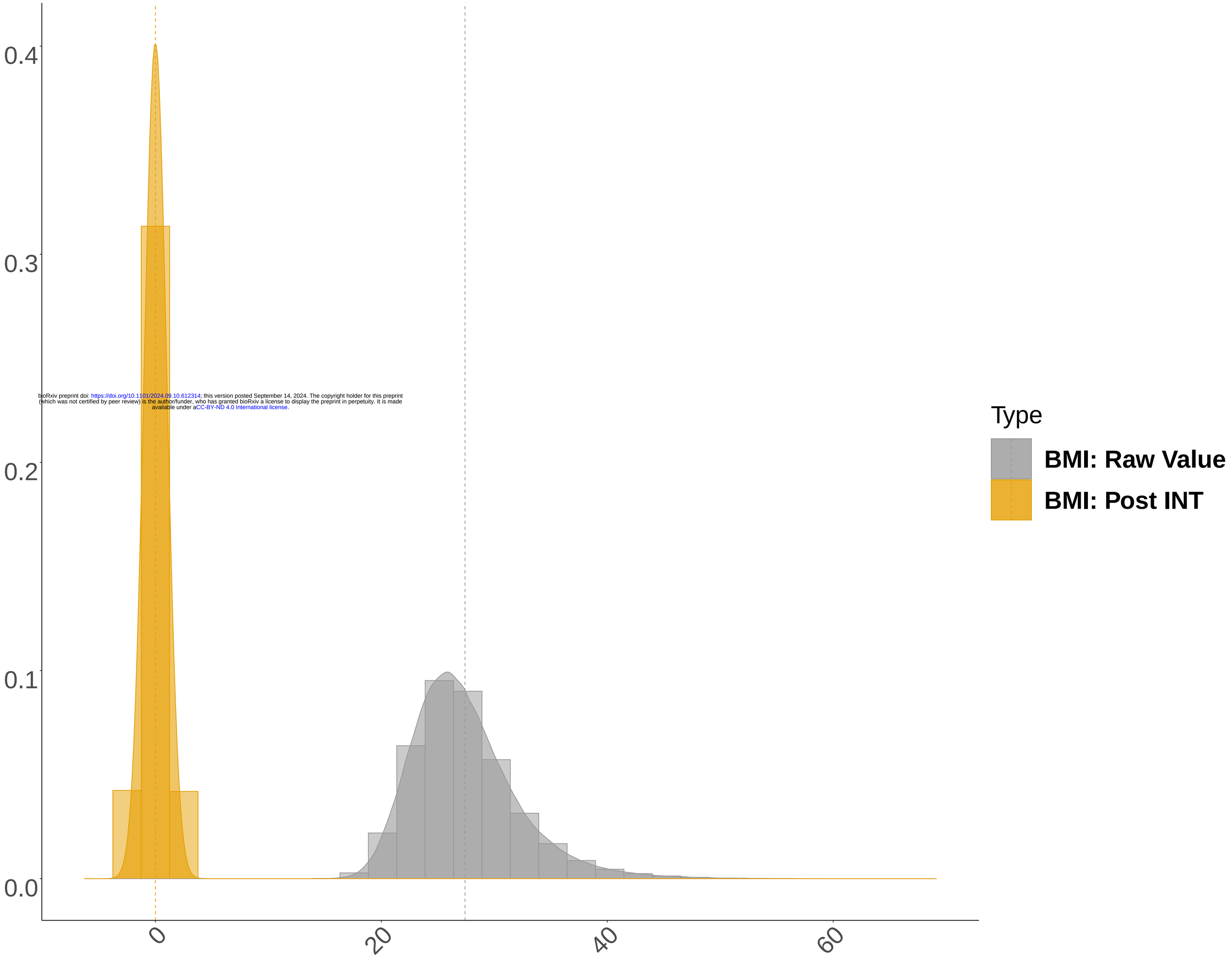


Variance and Covariance Distribution of Two Phenotypes Across Genotype Categories in 1,000 Simulated Datasets With Interaction Effects

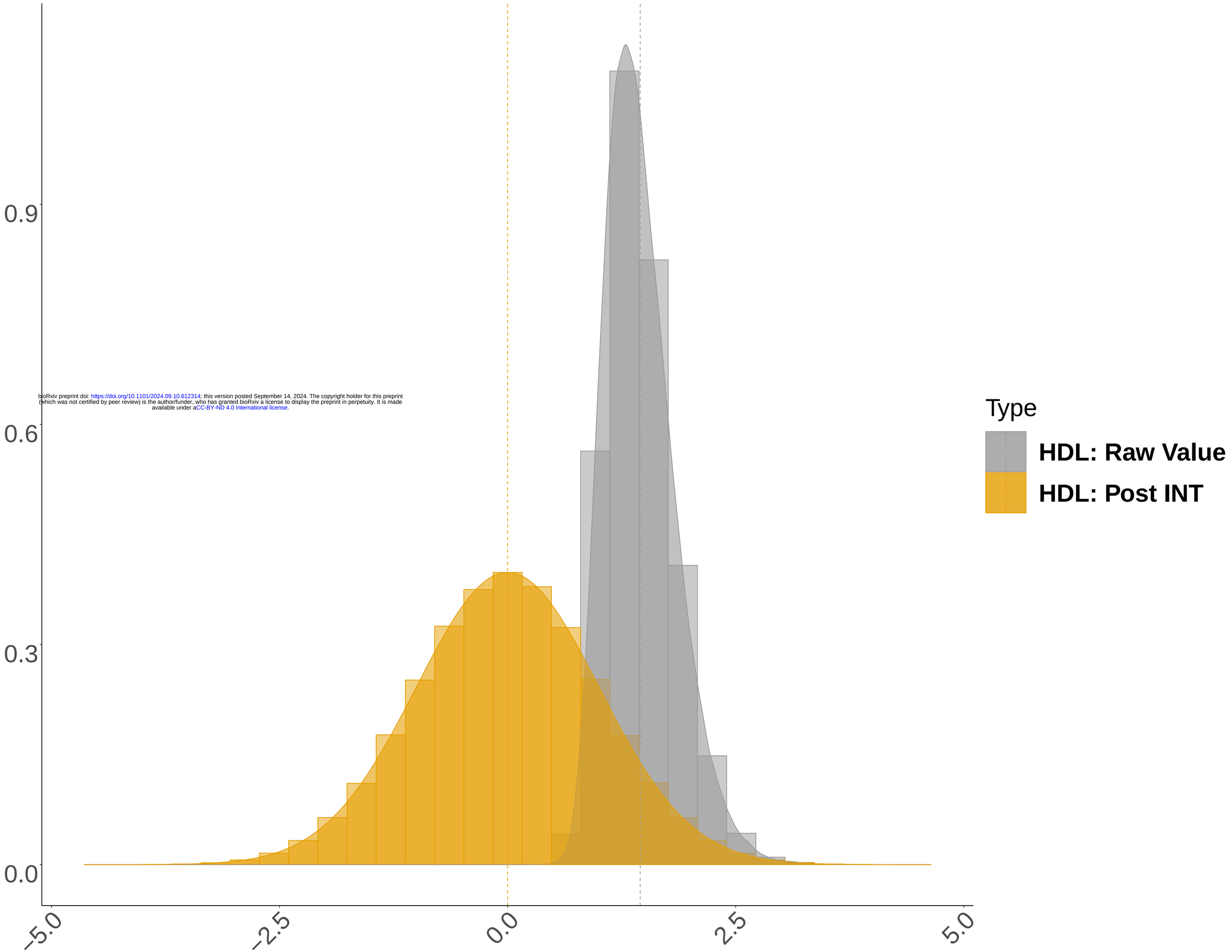
The differential covariance of the two phenotypes adds power while studying the interaction effects



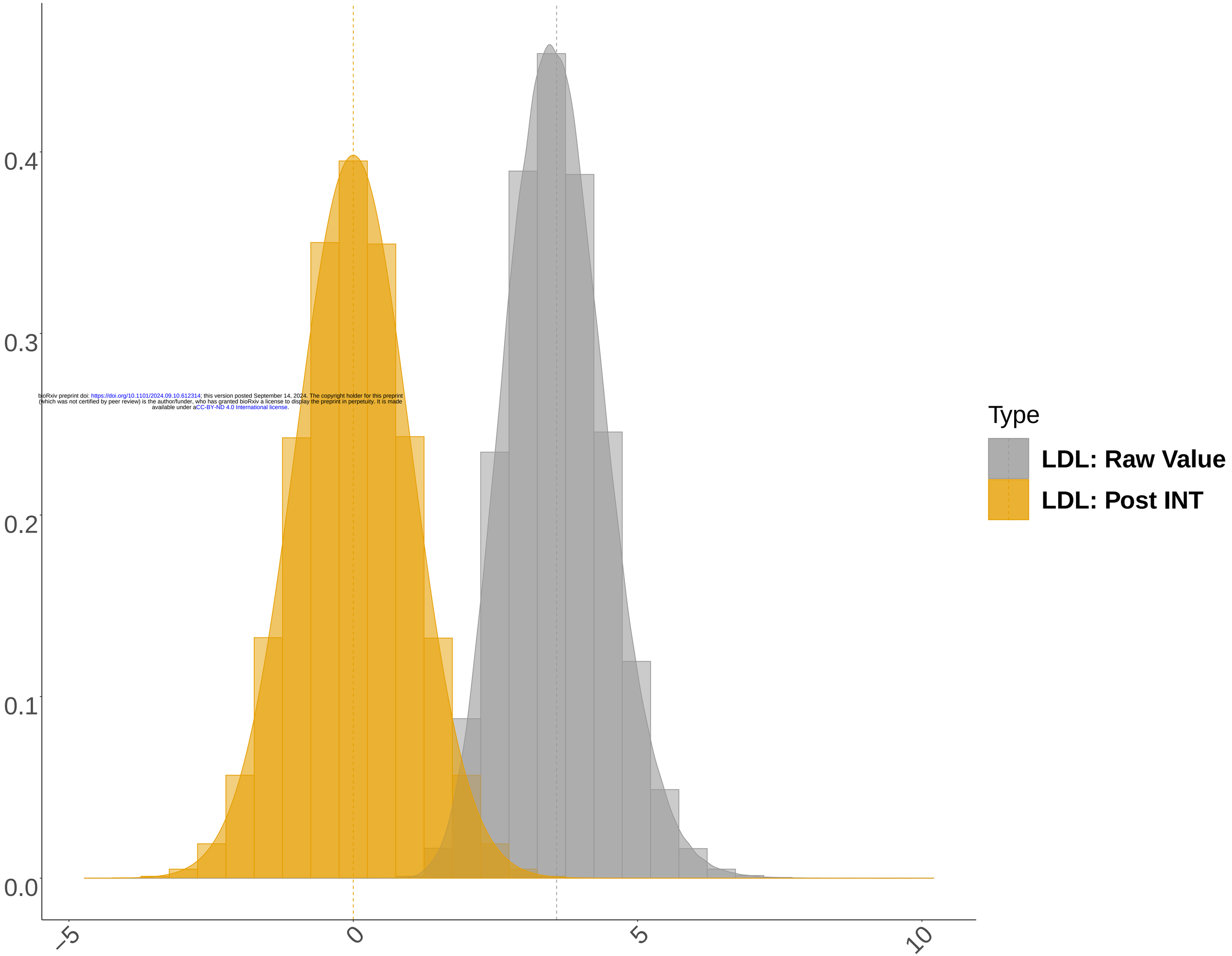
BMI: before and after inverse normal transformation (INT), N = 288,709



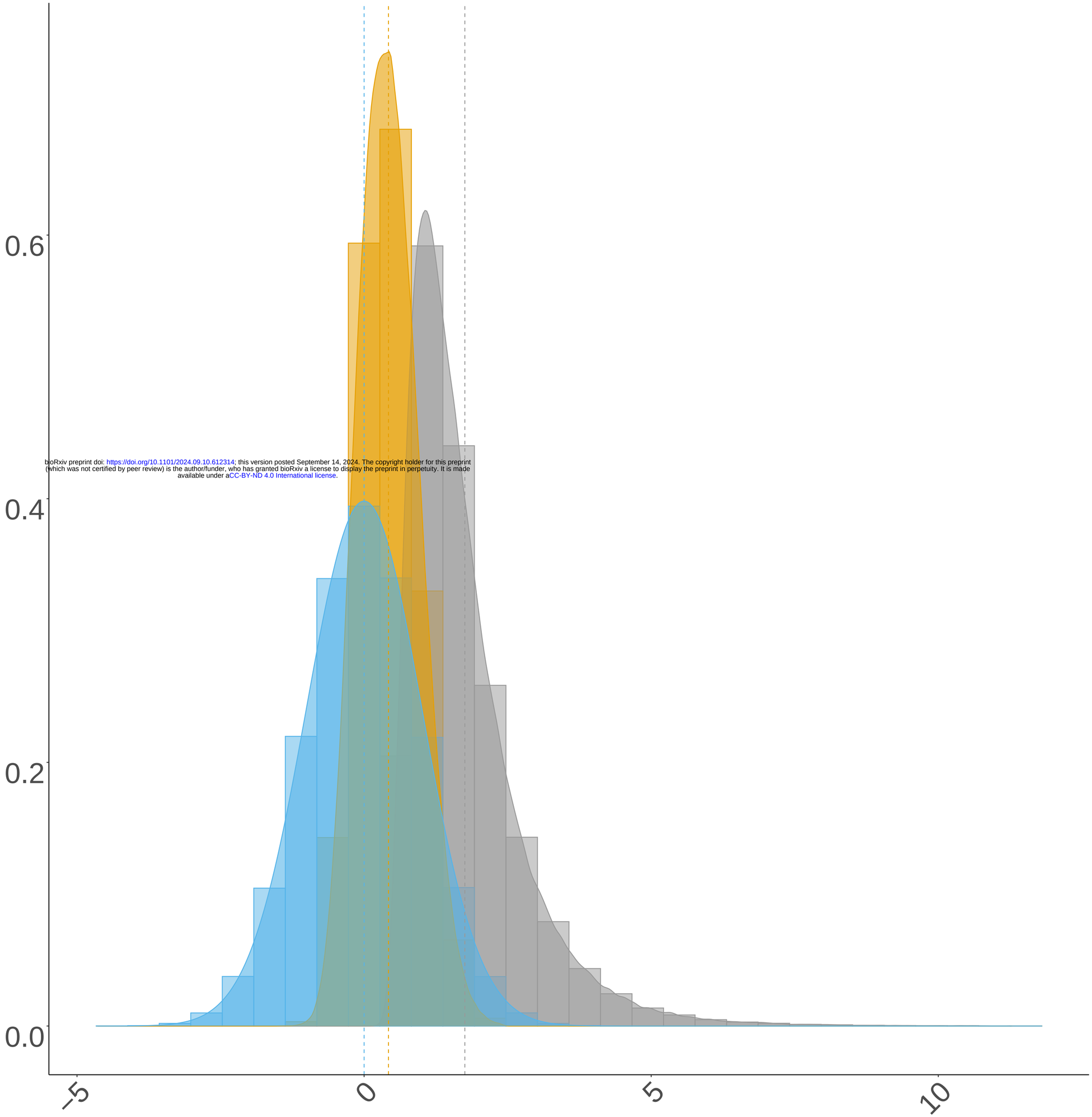
HDL: before and after inverse normal transformation (INT), N = 288,709



LDL: before and after inverse normal transformation (INT), N = 288,709



Natural Log Transformed Triglycerides: before and after inverse normal transformation (INT), N = 288,709



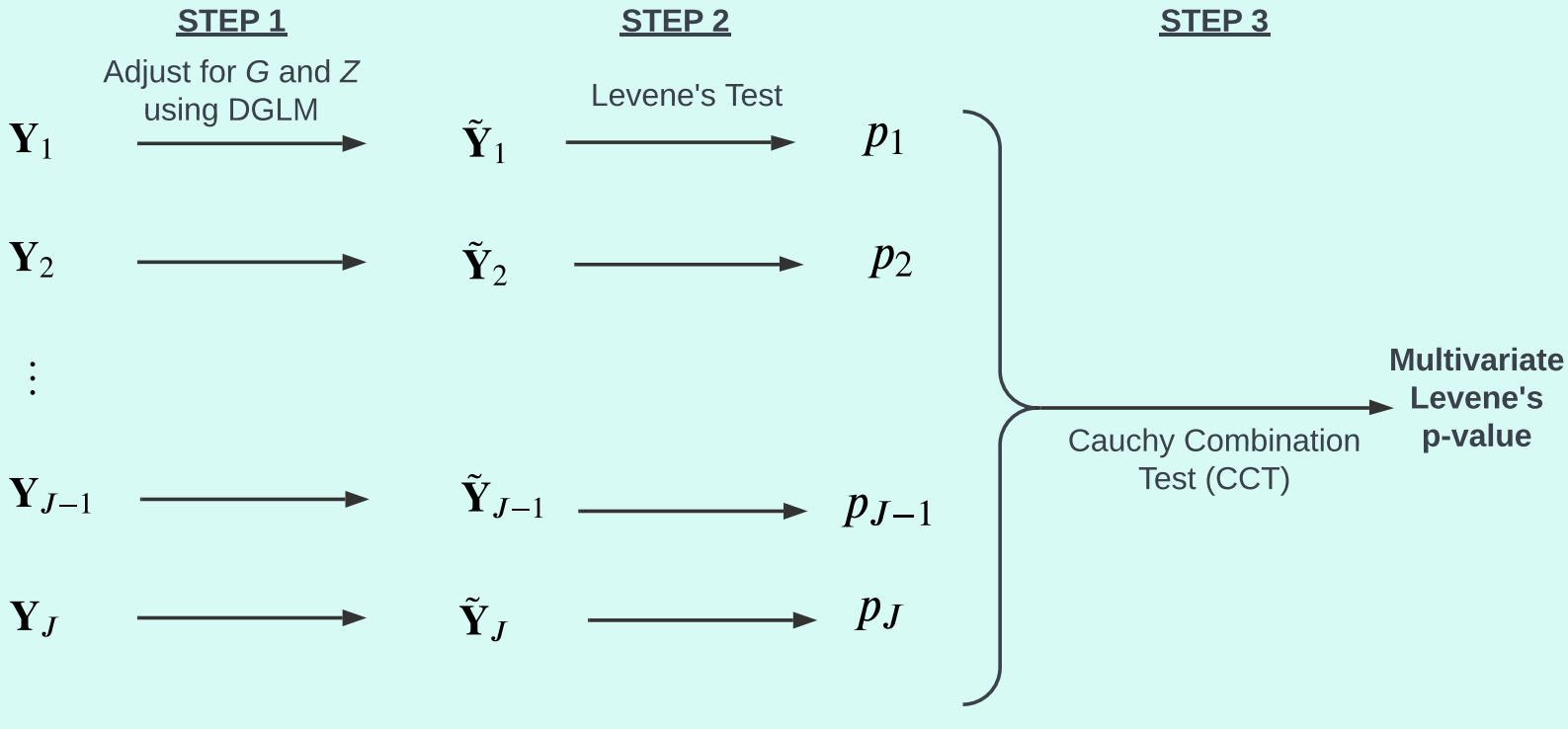
Input

P Phenotypes: $\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 & \dots & \mathbf{Y}_J \end{bmatrix}$

K Confounders: $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 & \dots & \mathbf{Z}_K \end{bmatrix}$

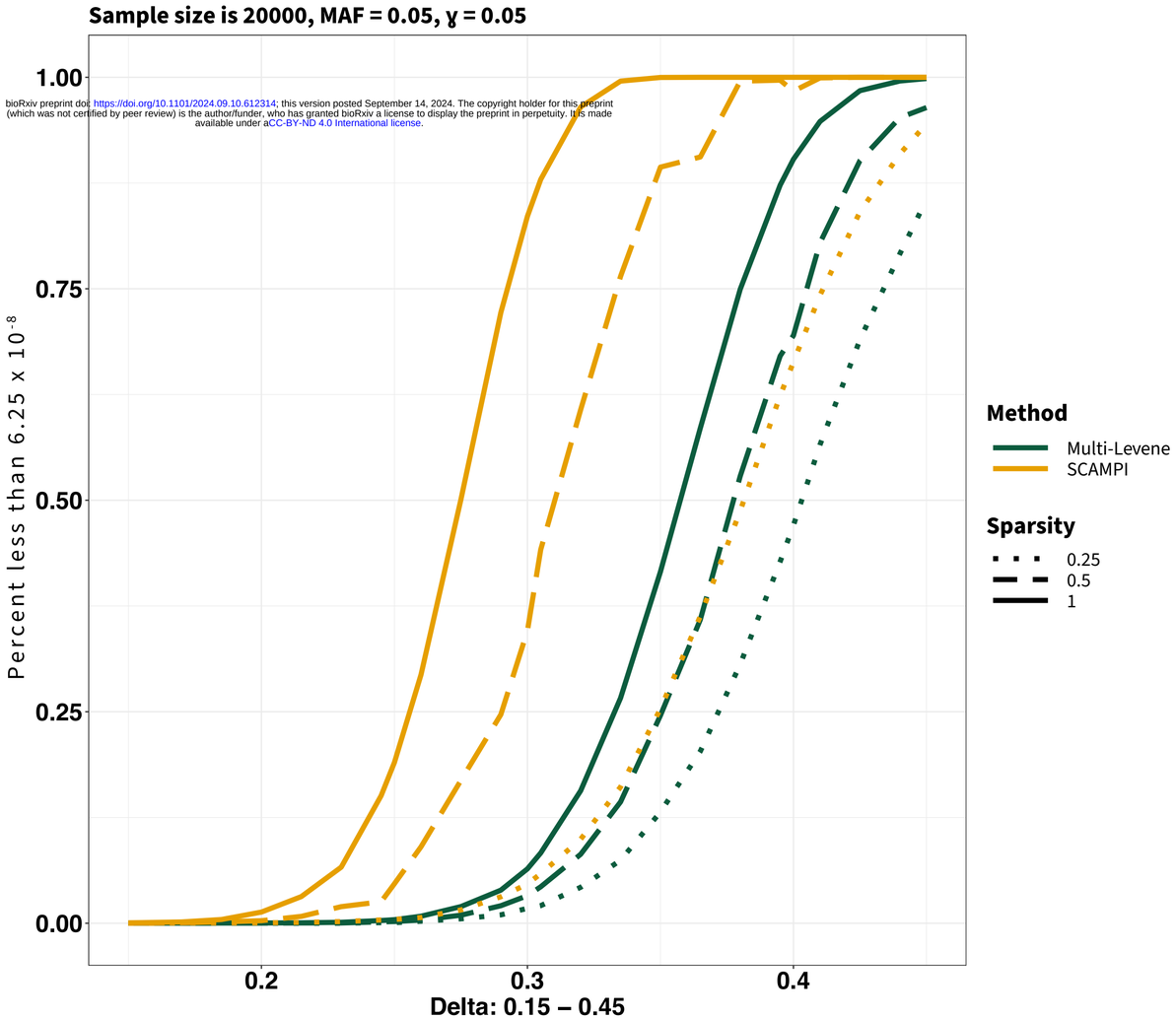
One Test Genotype: $\mathbf{G}$

Multivariate Levene's Test Framework

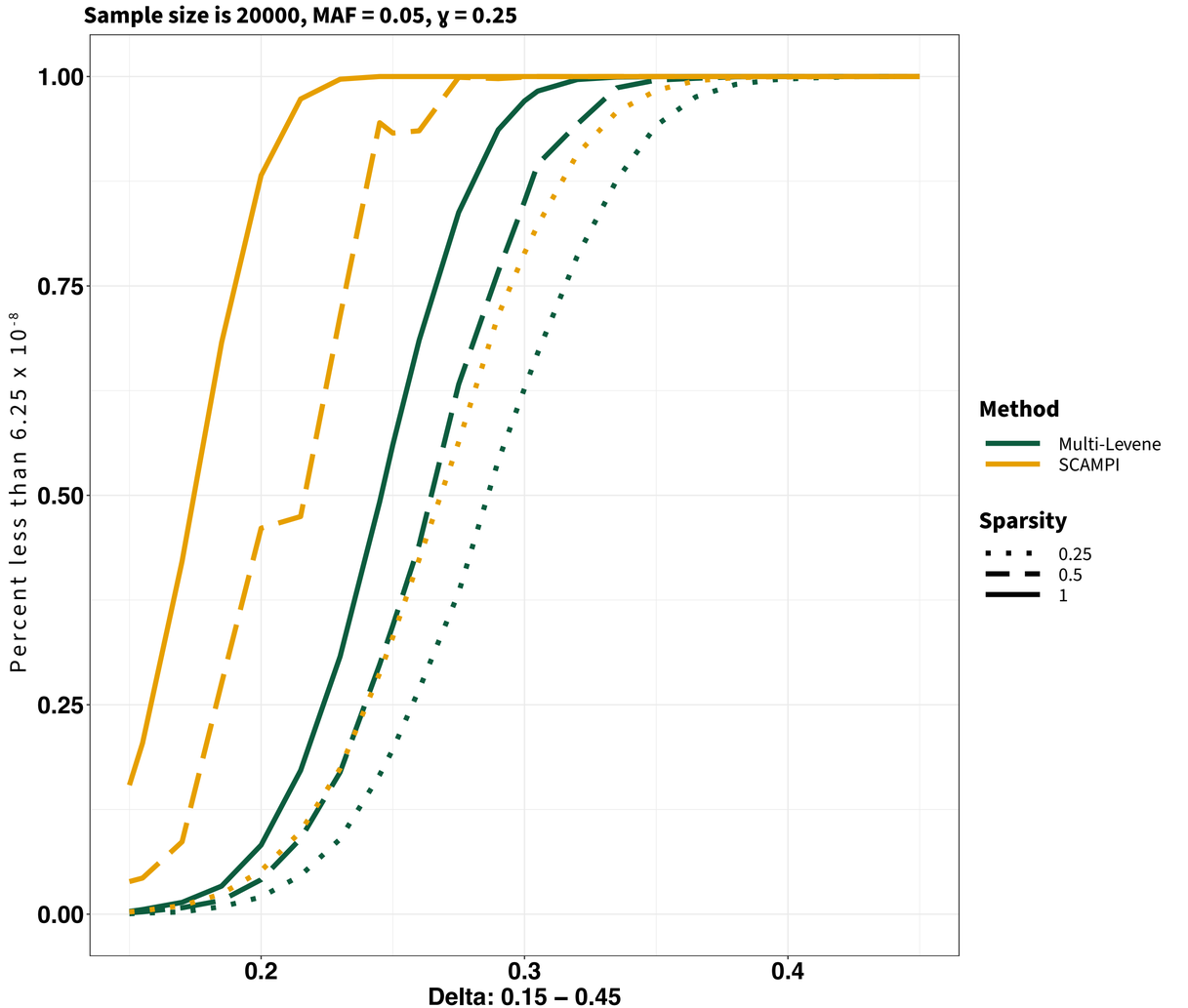


J = 4

ɣ = 0.05



ɣ = 0.25

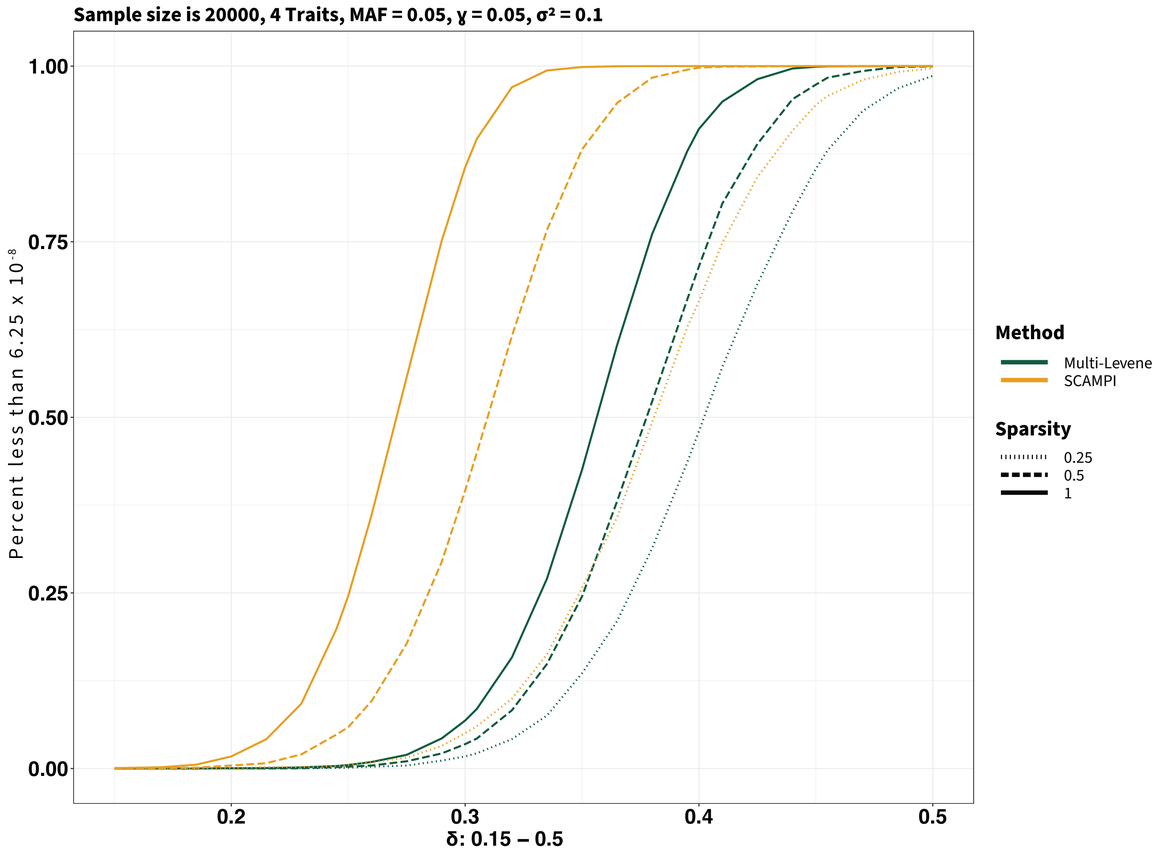


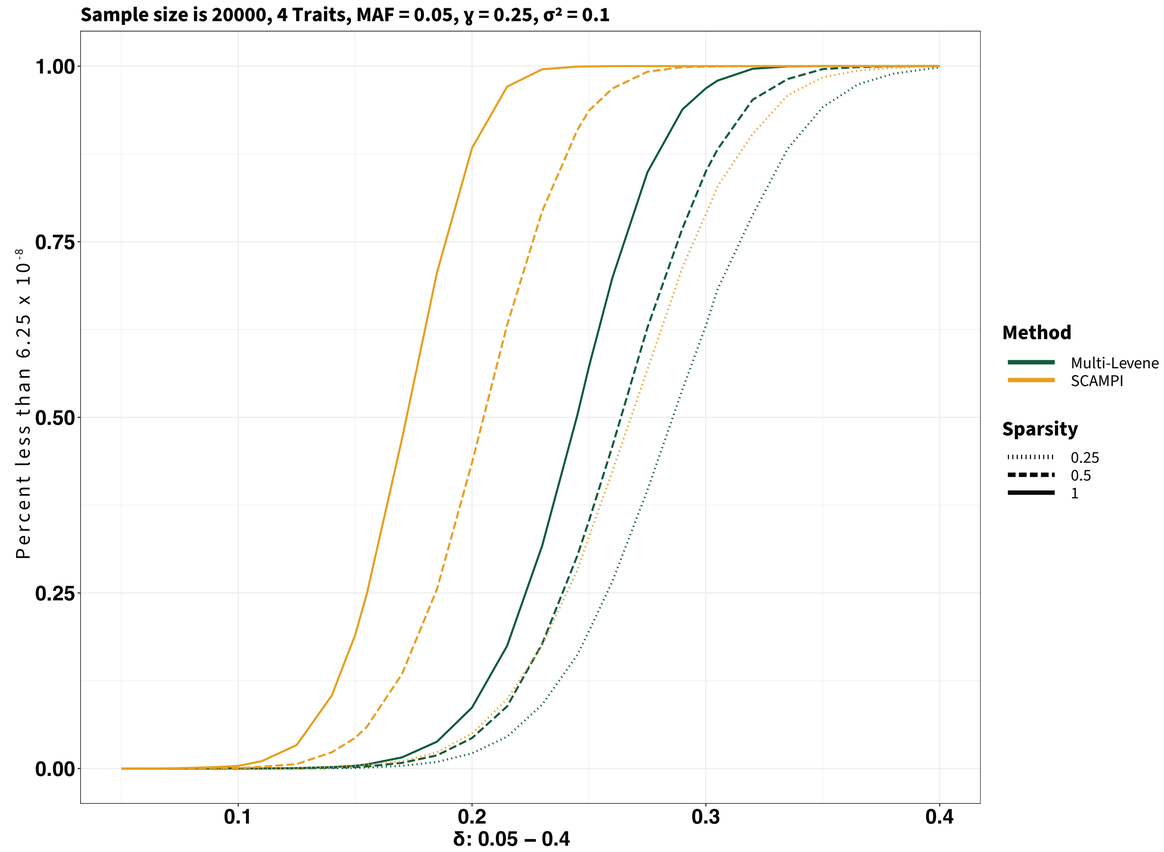
J = 4

ɣ = 0.05

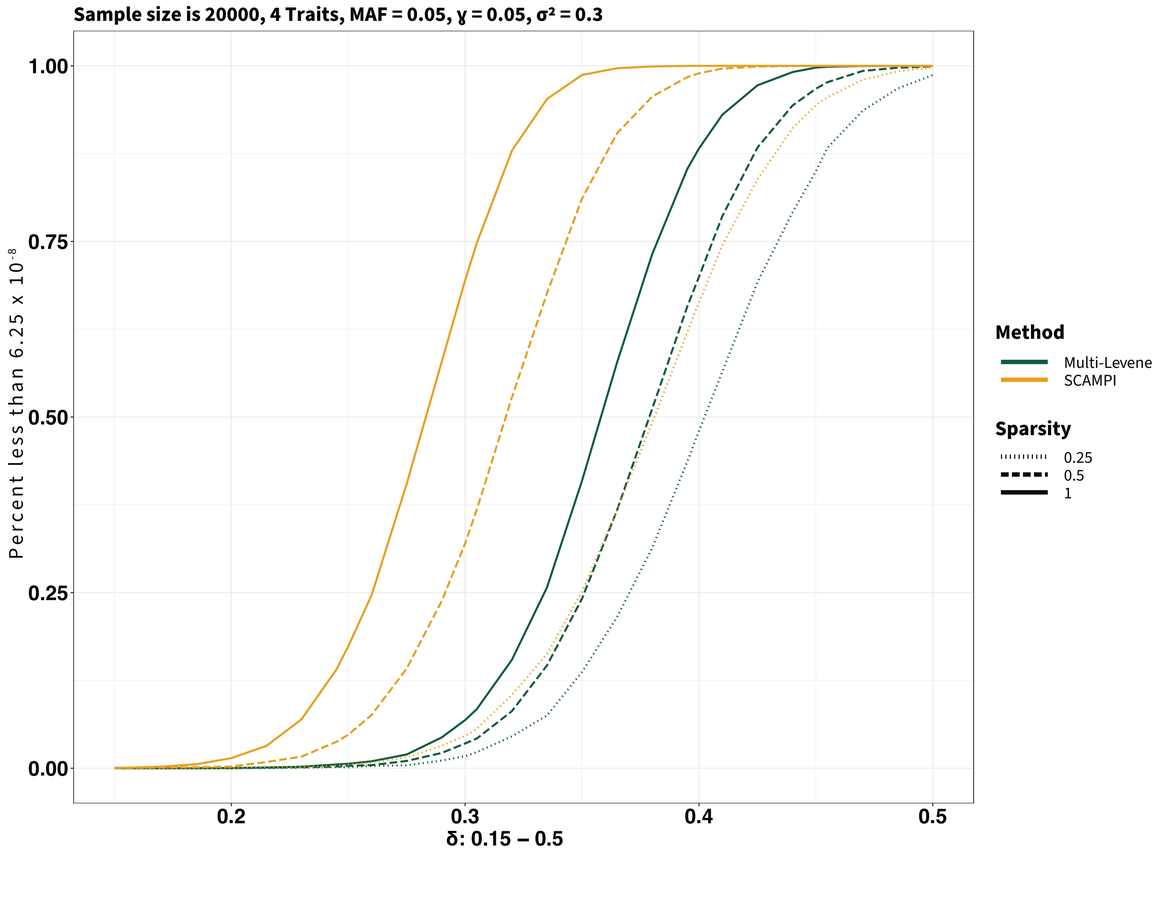
ɣ = 0.25

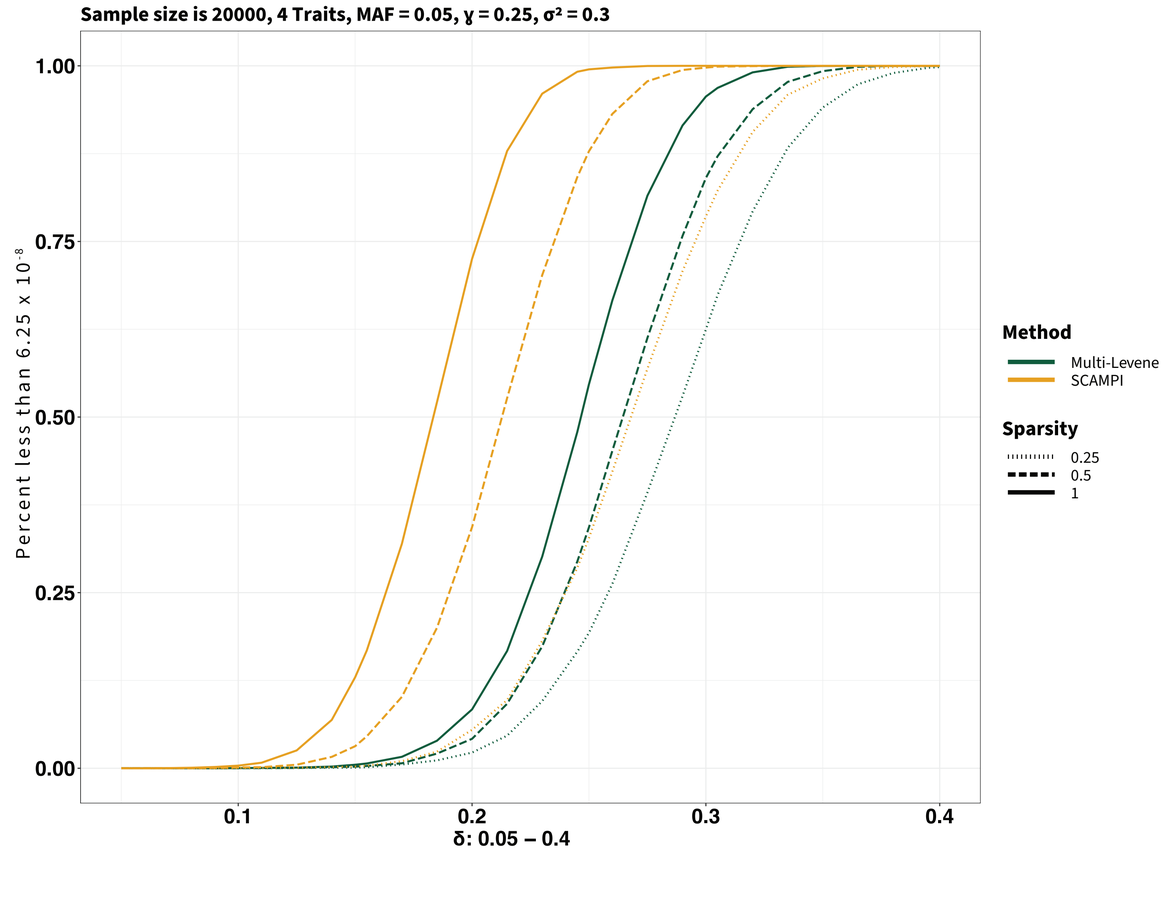
σ² = 0.1



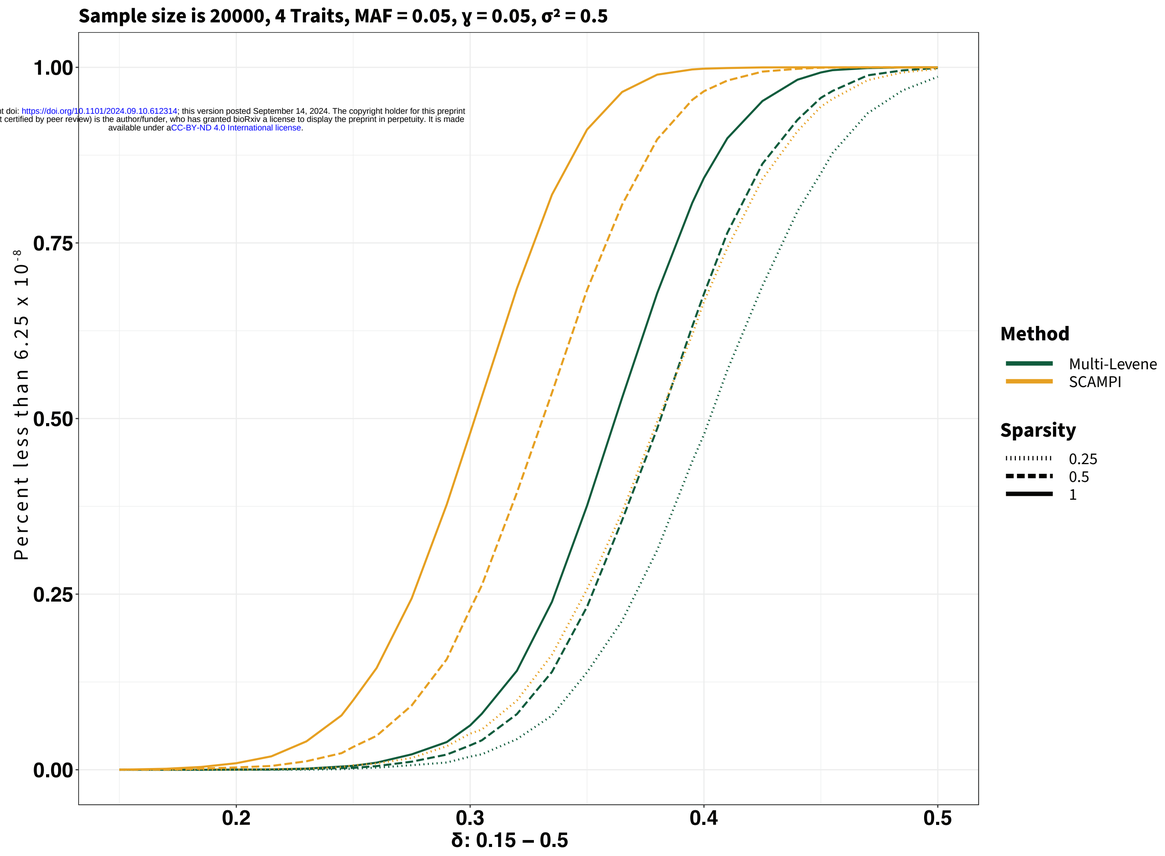


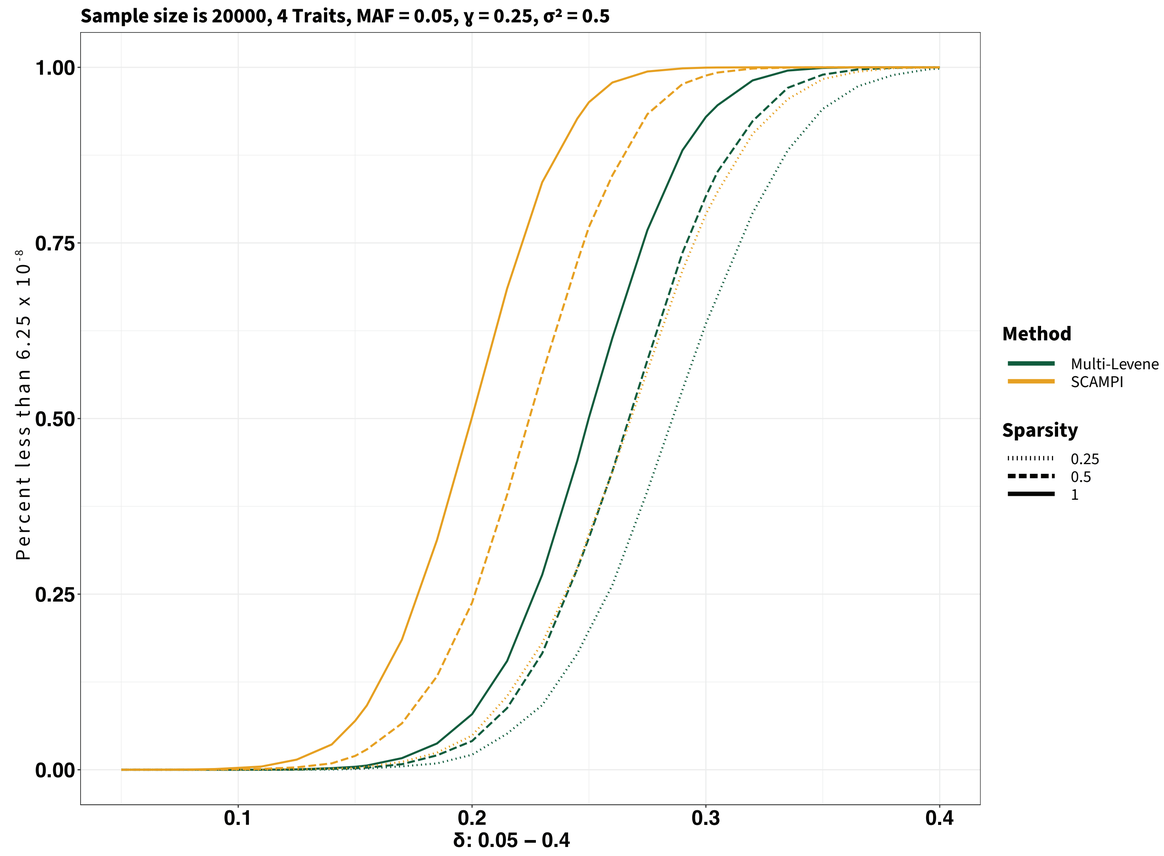
σ² = 0.3





σ² = 0.5



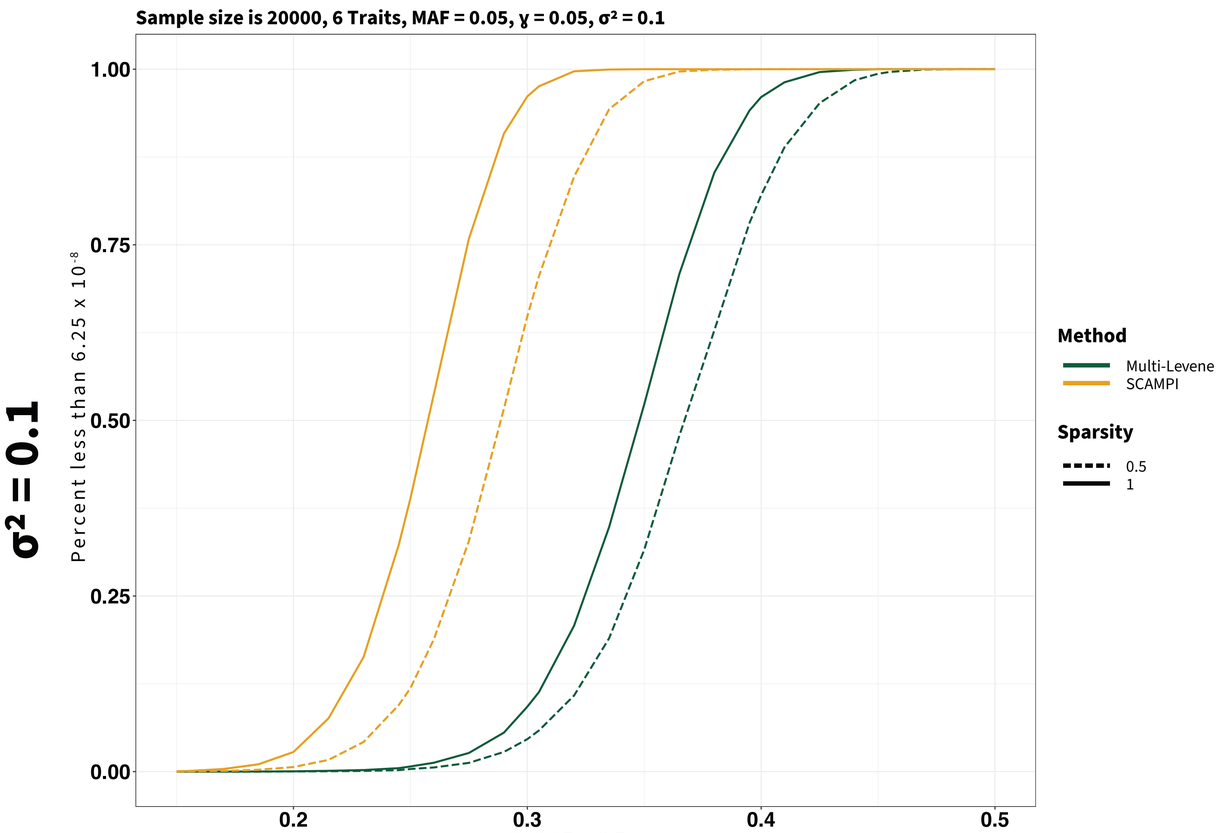


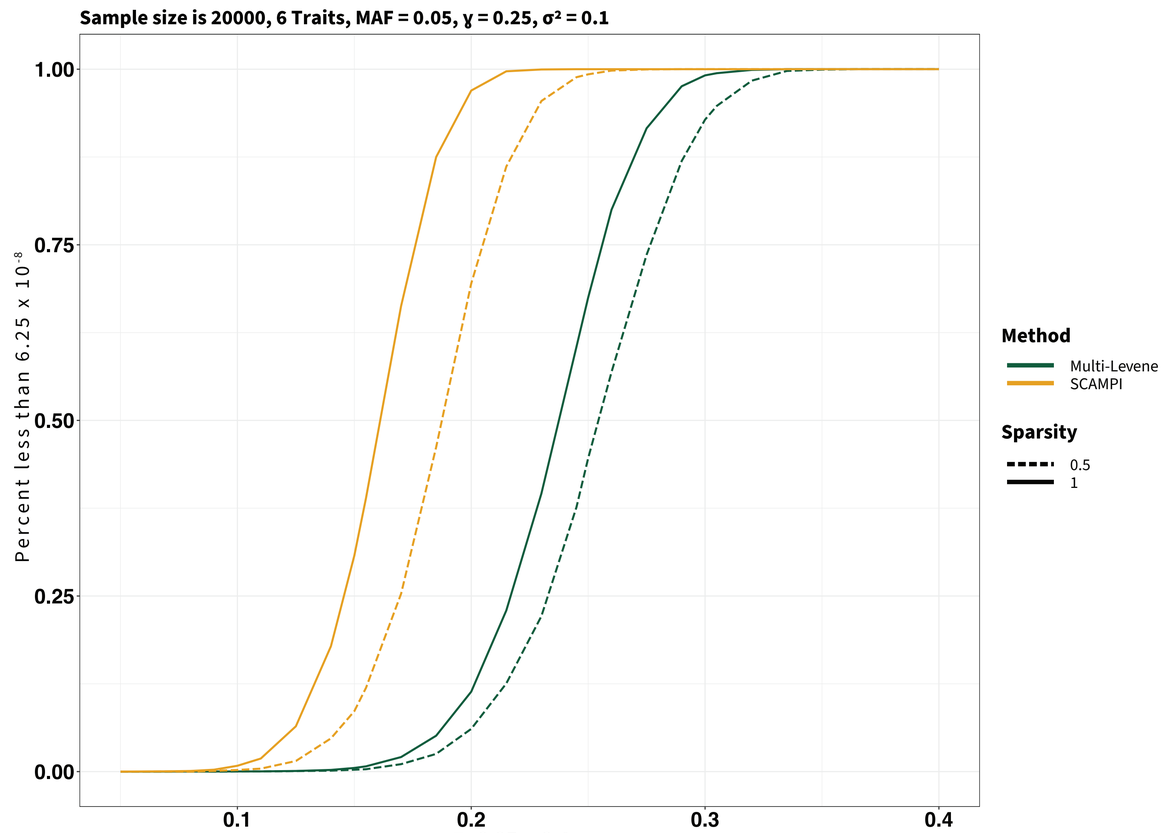
J = 6

ɣ = 0.05

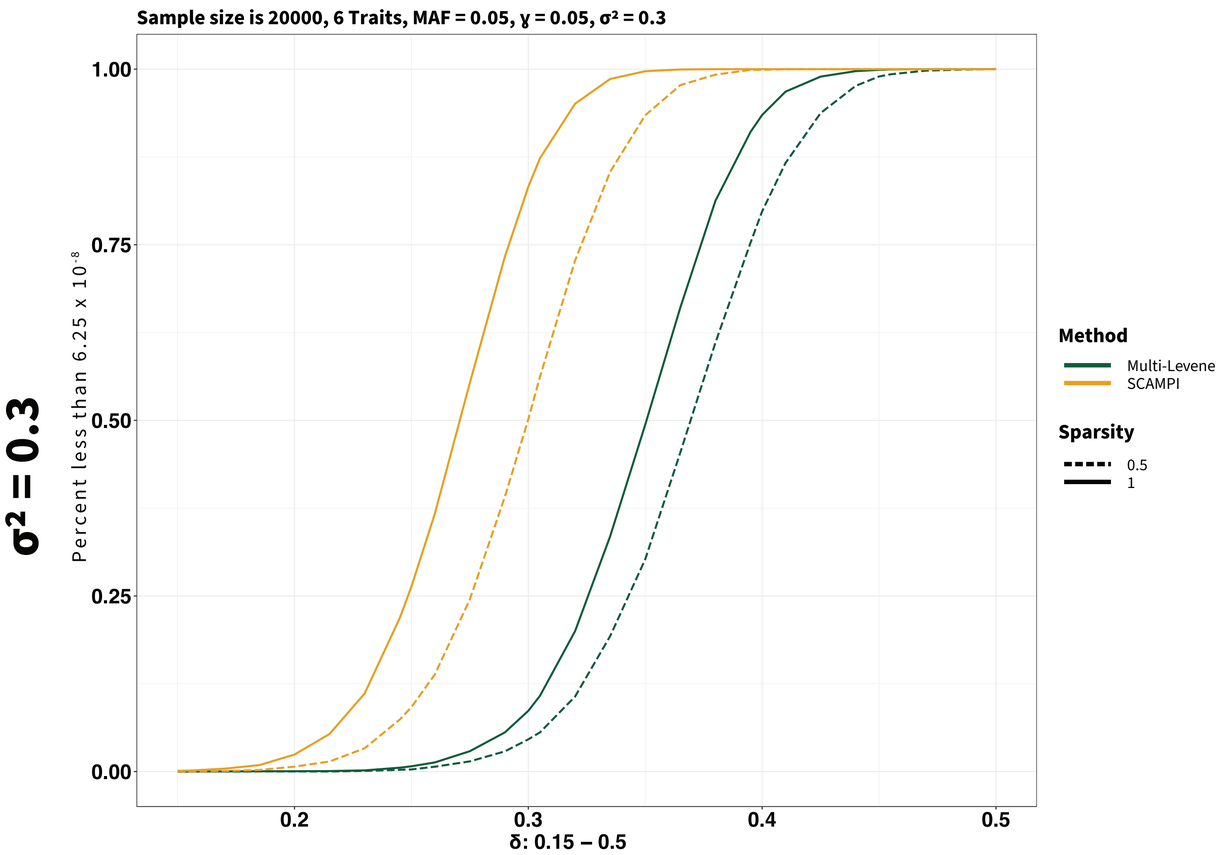
ɣ = 0.25

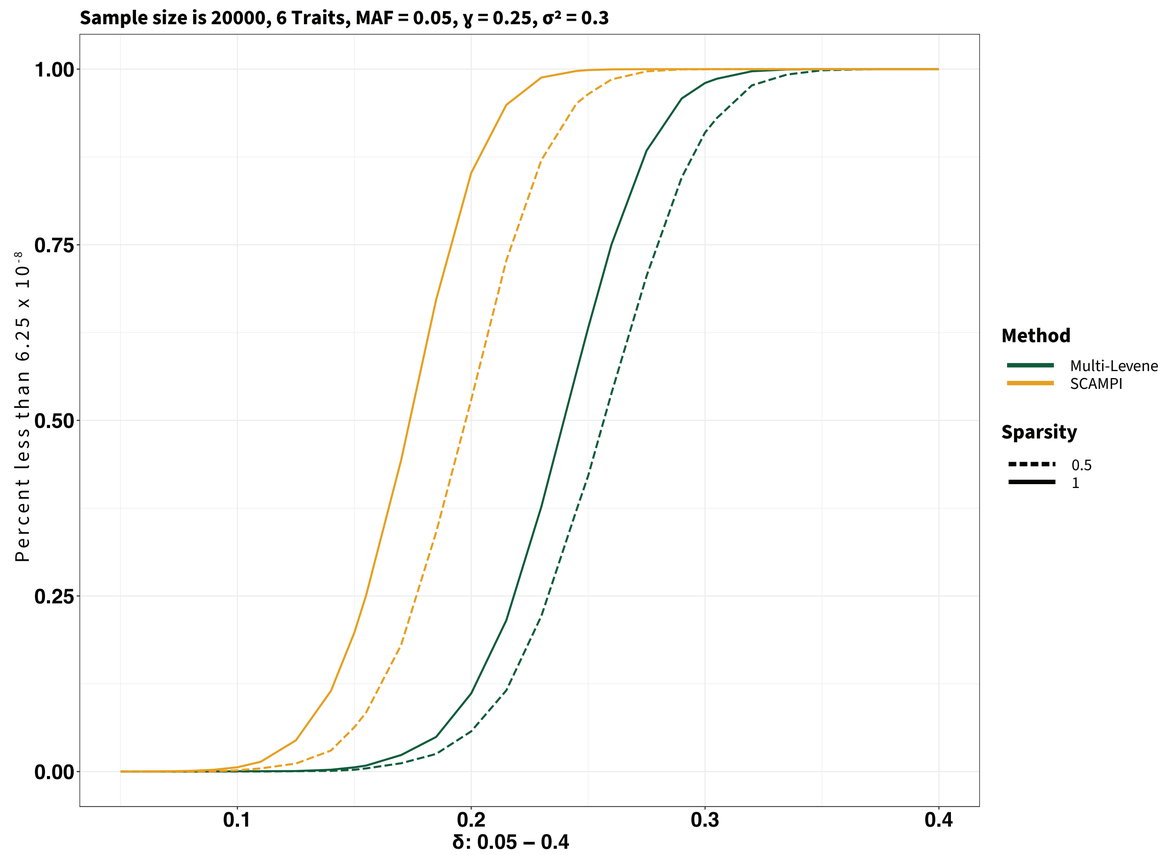
σ² = 0.1



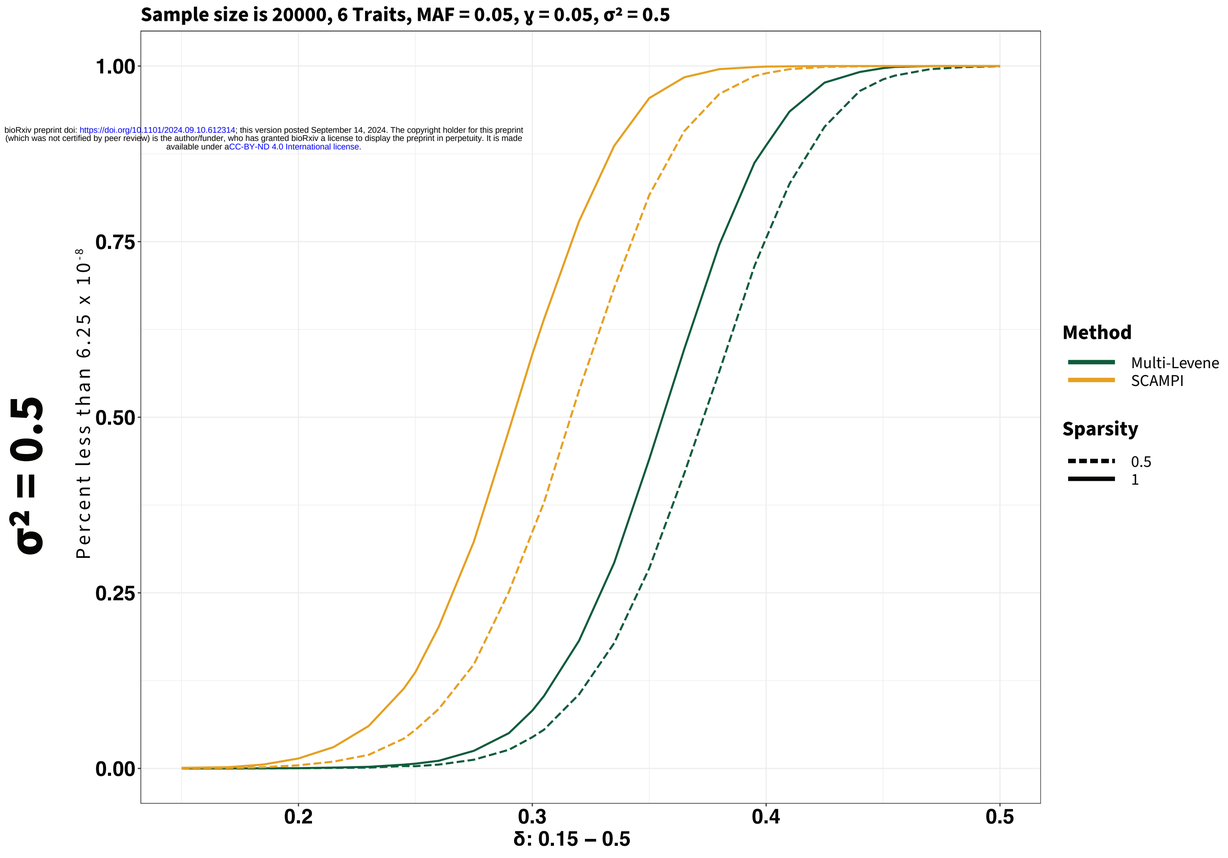


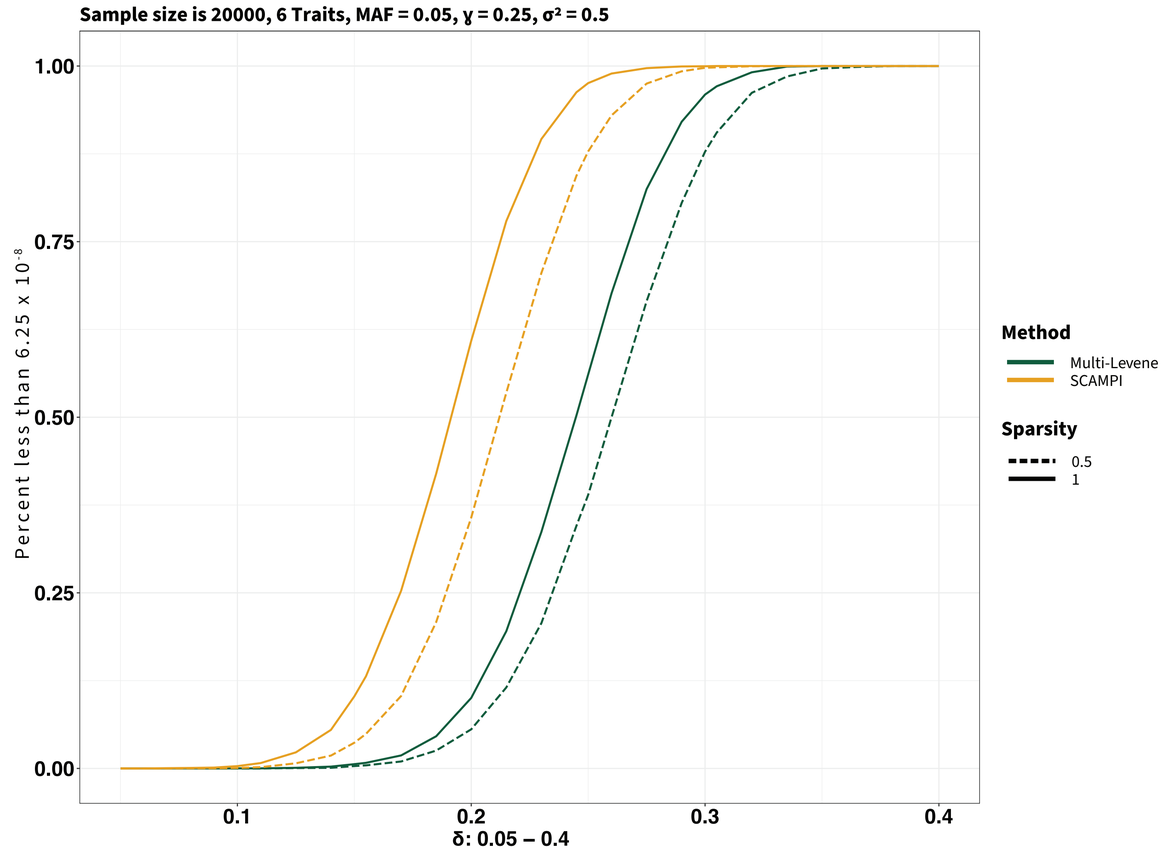
σ² = 0.3





σ² = 0.5



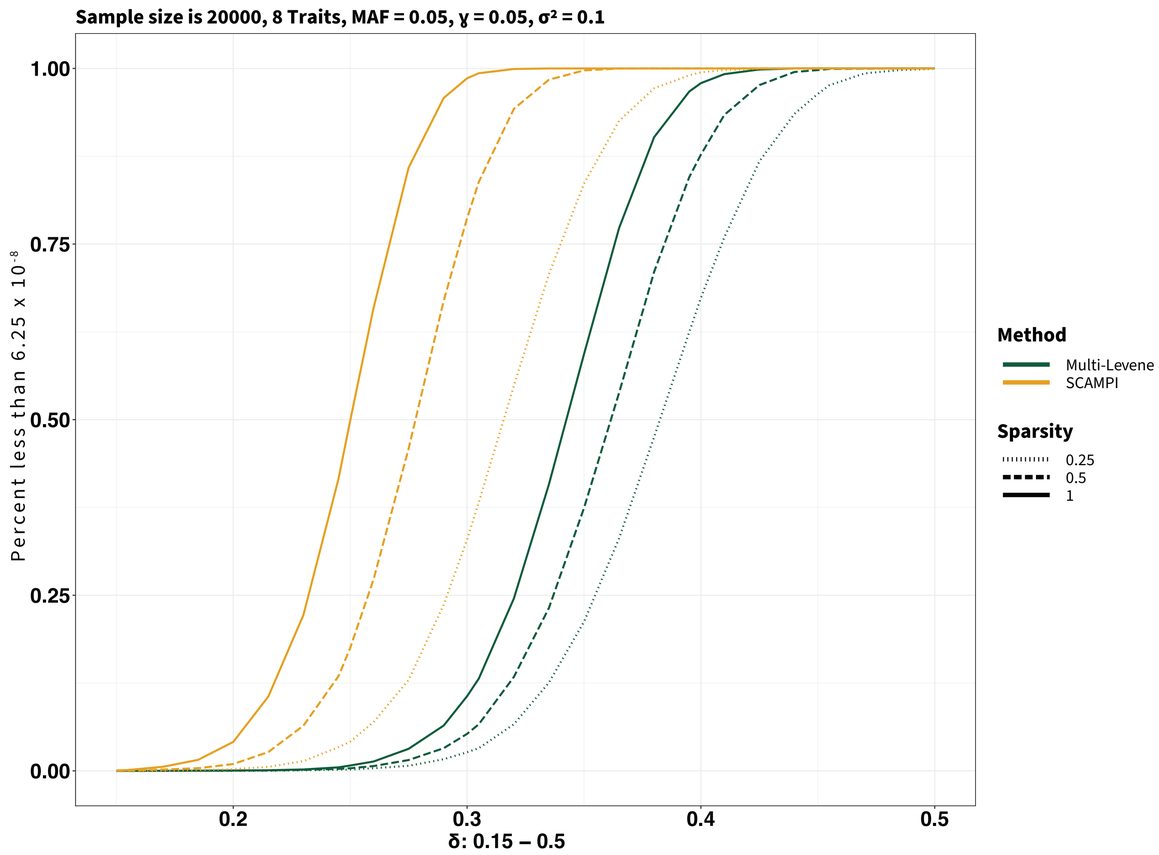


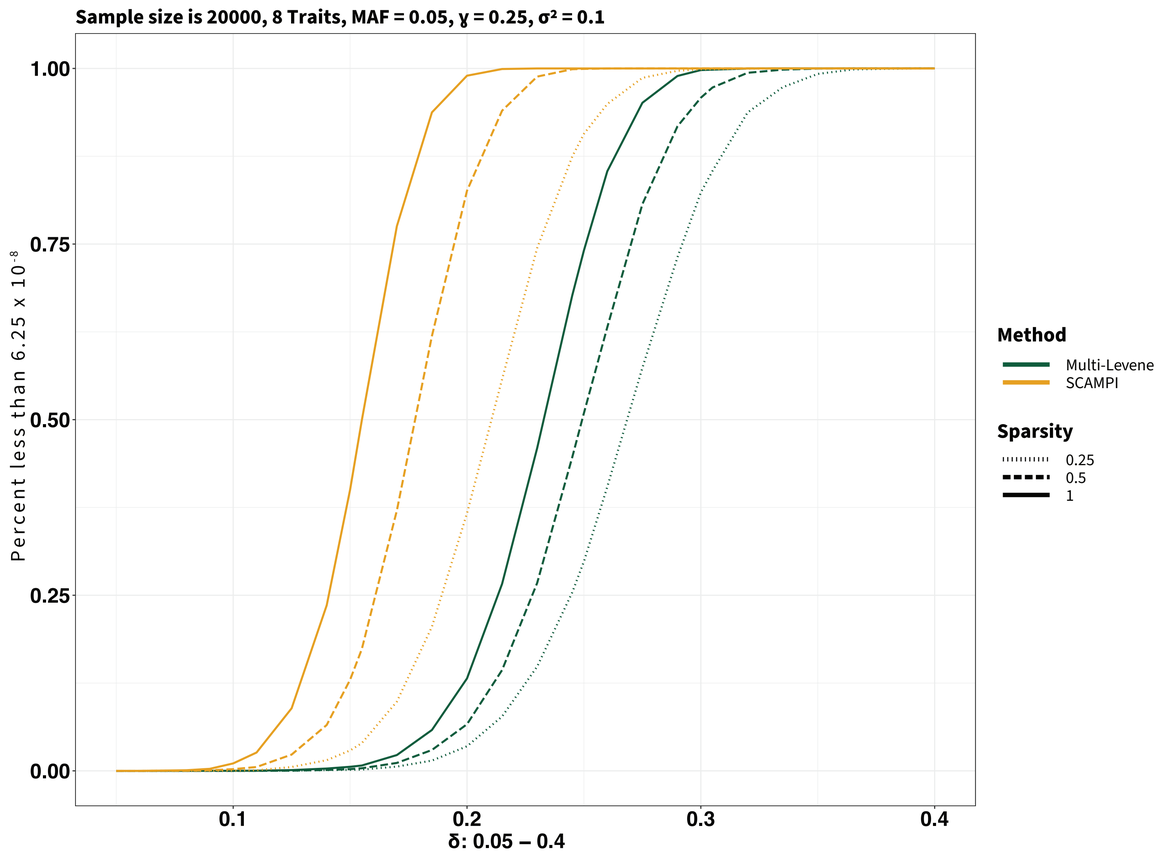
J = 8

ɣ = 0.05

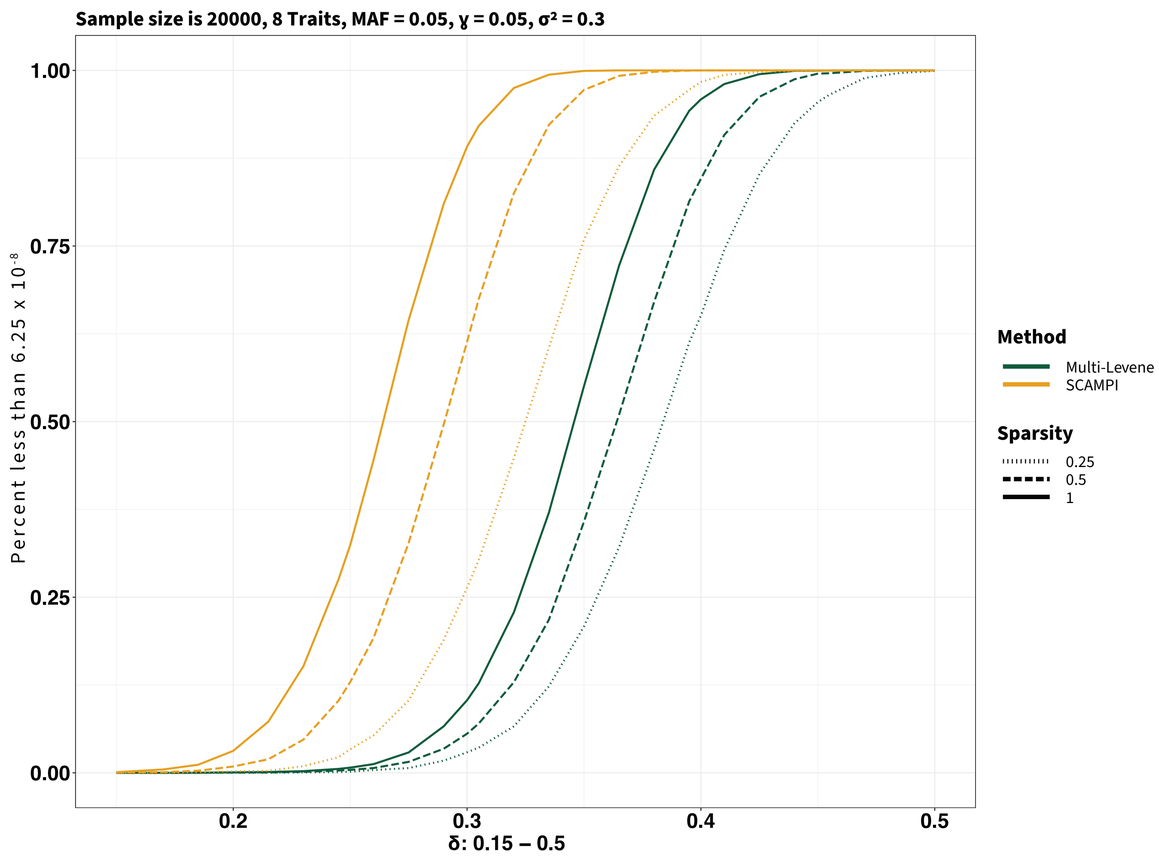
ɣ = 0.25

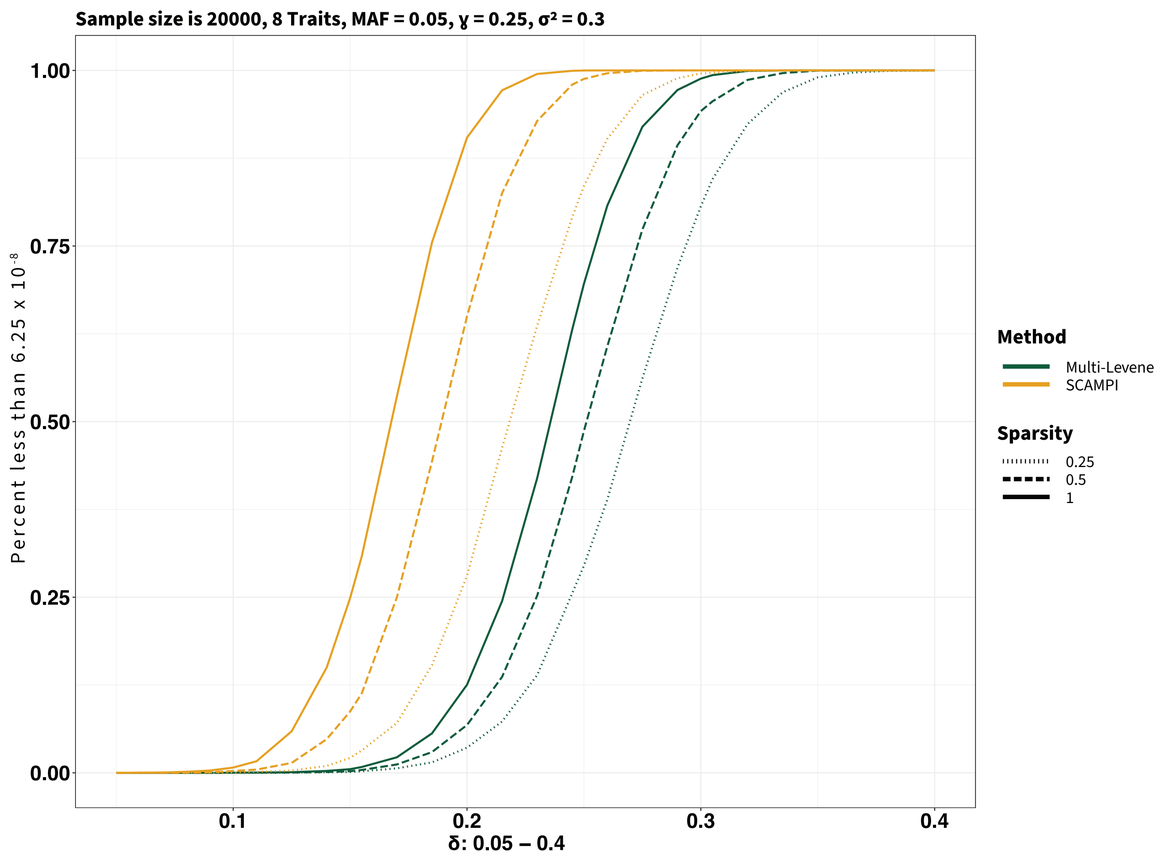
σ² = 0.1



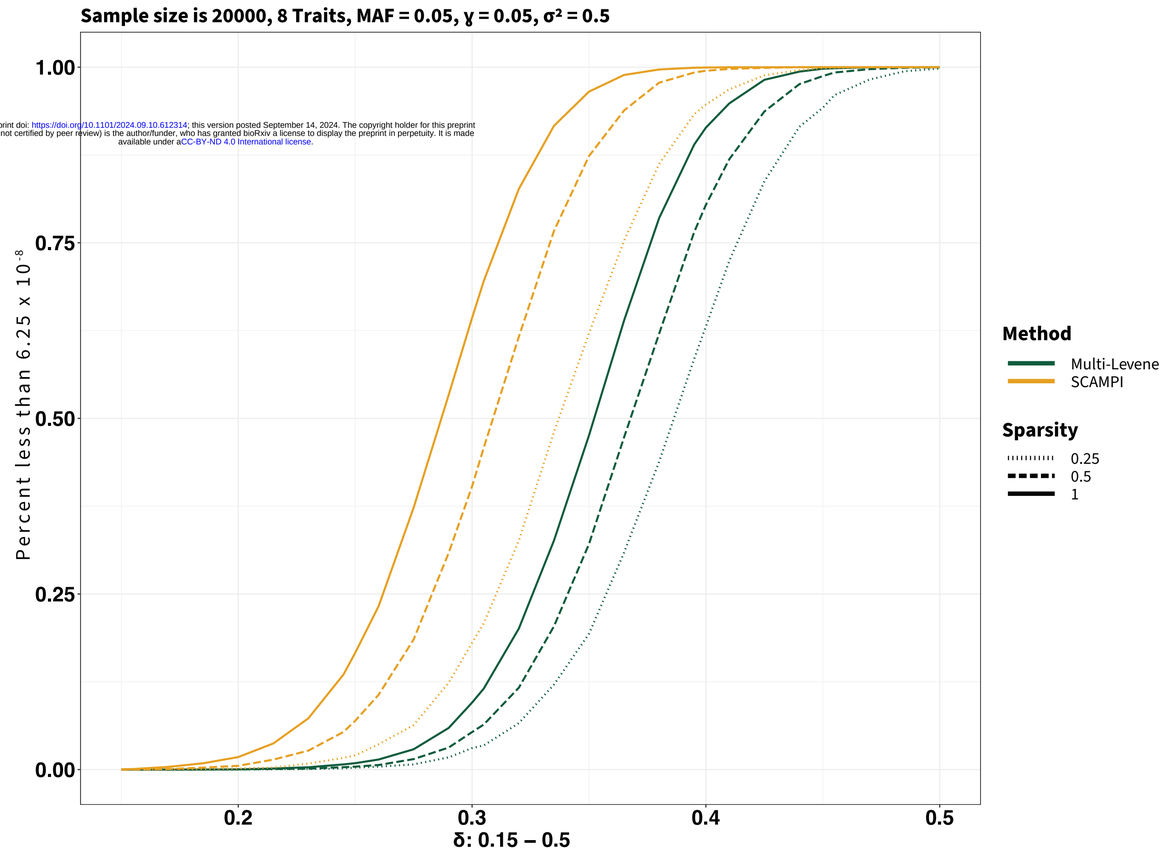


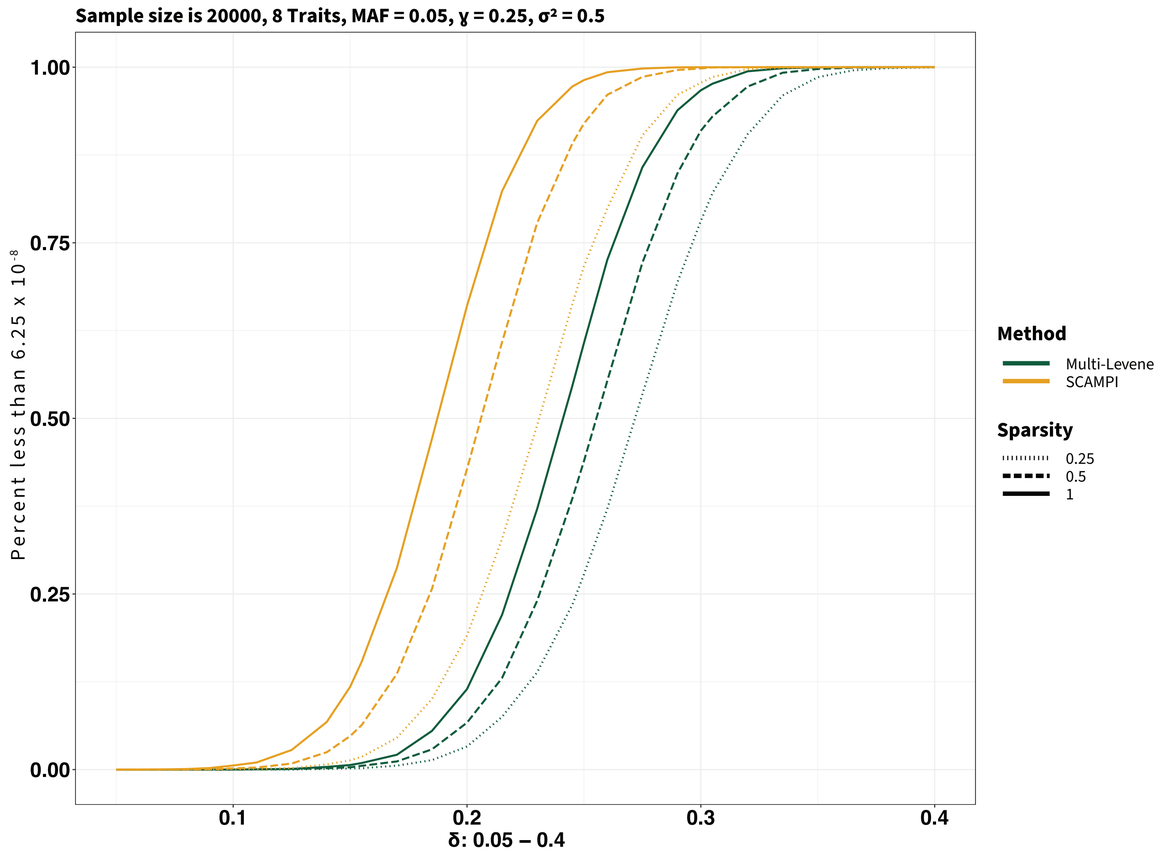
σ² = 0.3



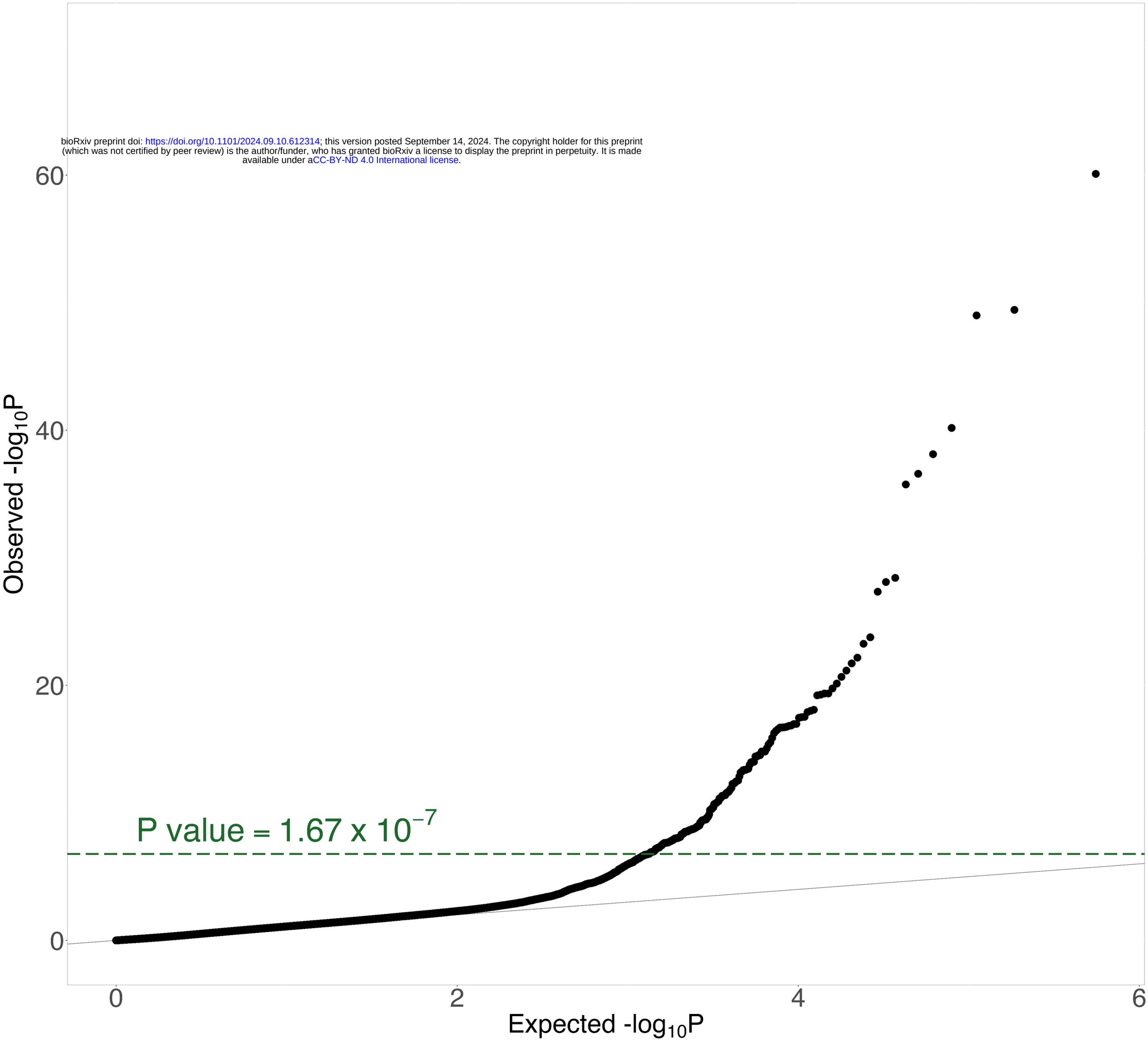


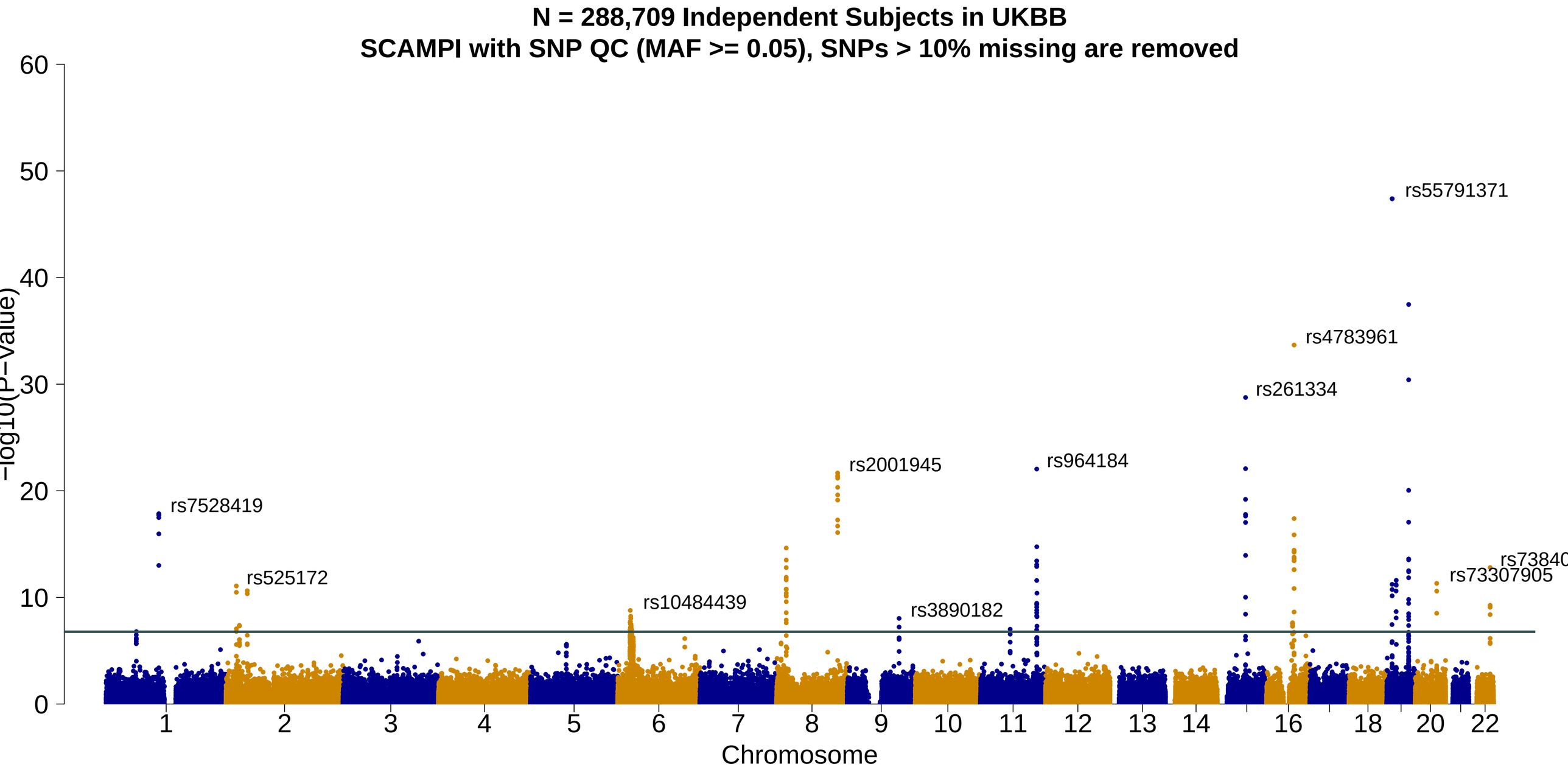
σ² = 0.5



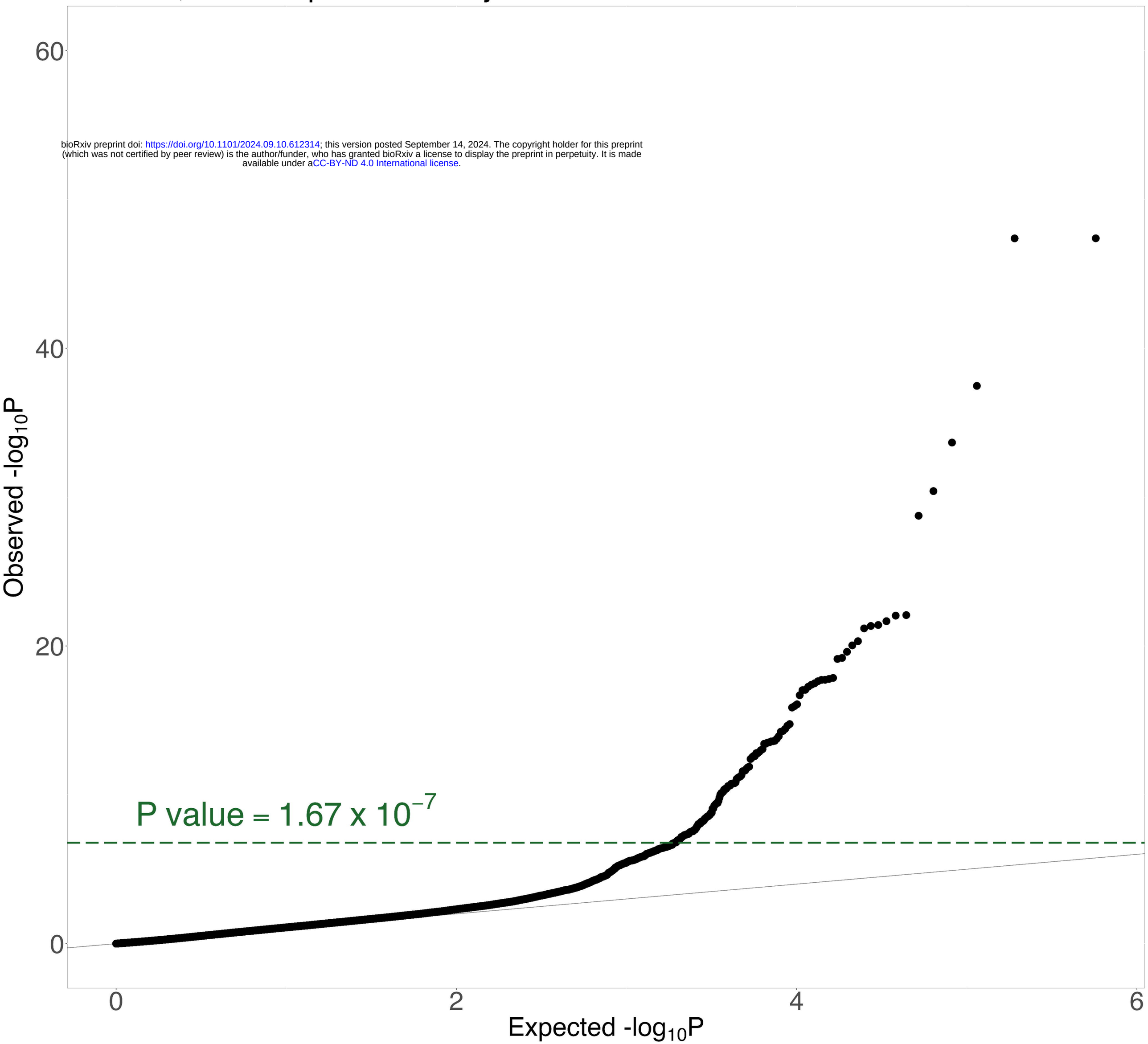


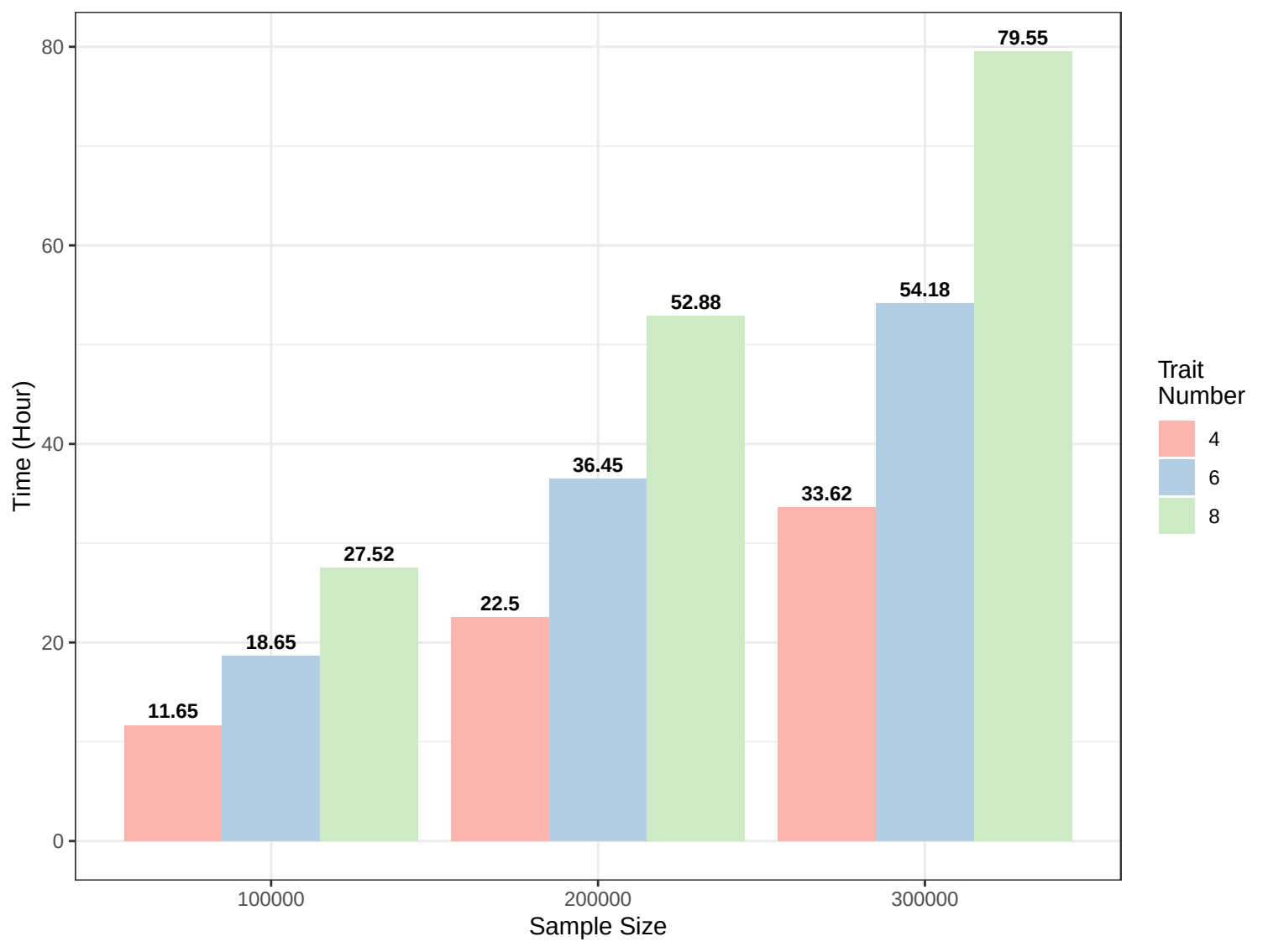
SCAMPI with SNP QC (MAF >= 0.05), SNPs > 10% missing are removed N = 288,709 Independent Subjects in UKBB

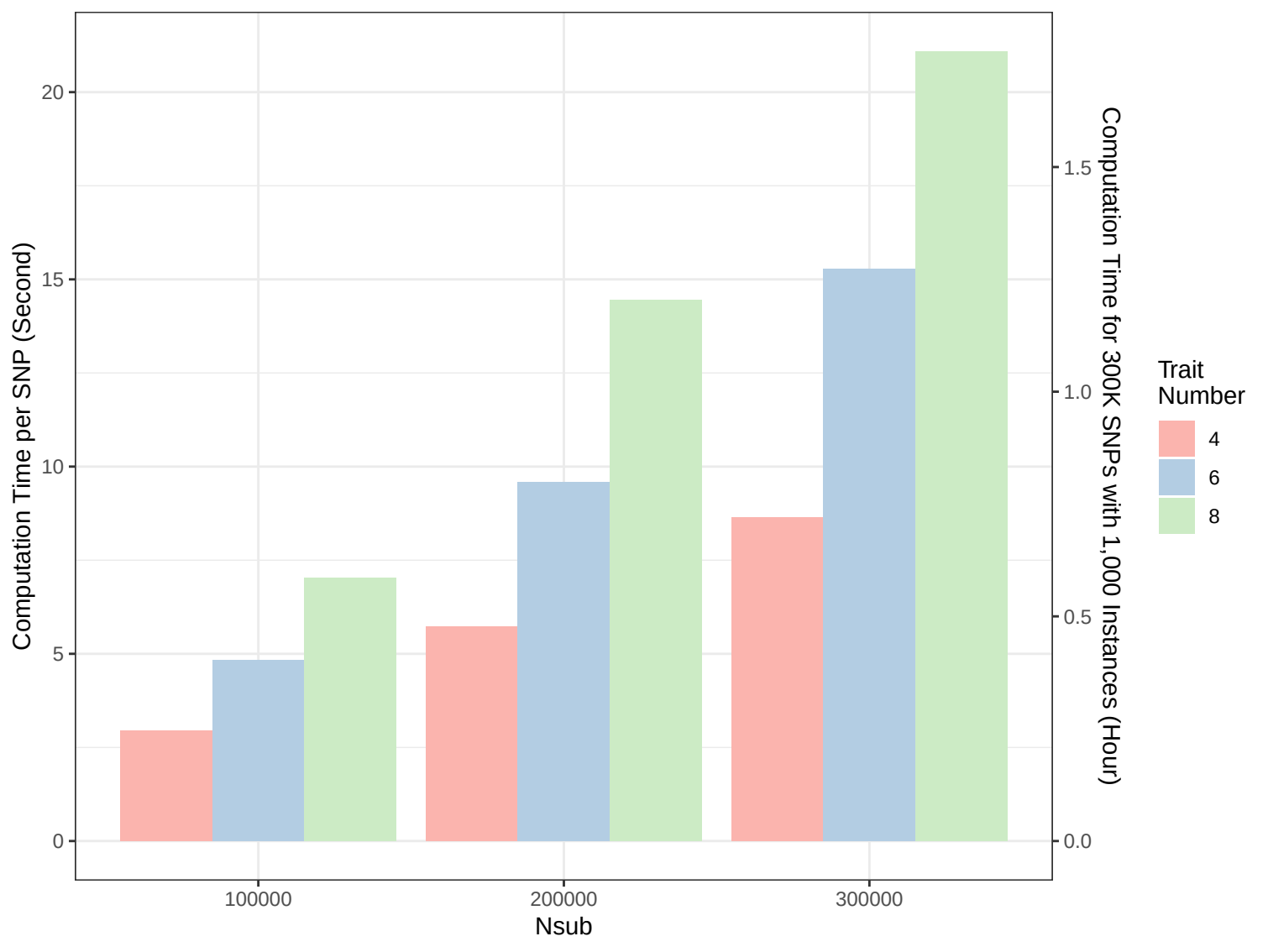


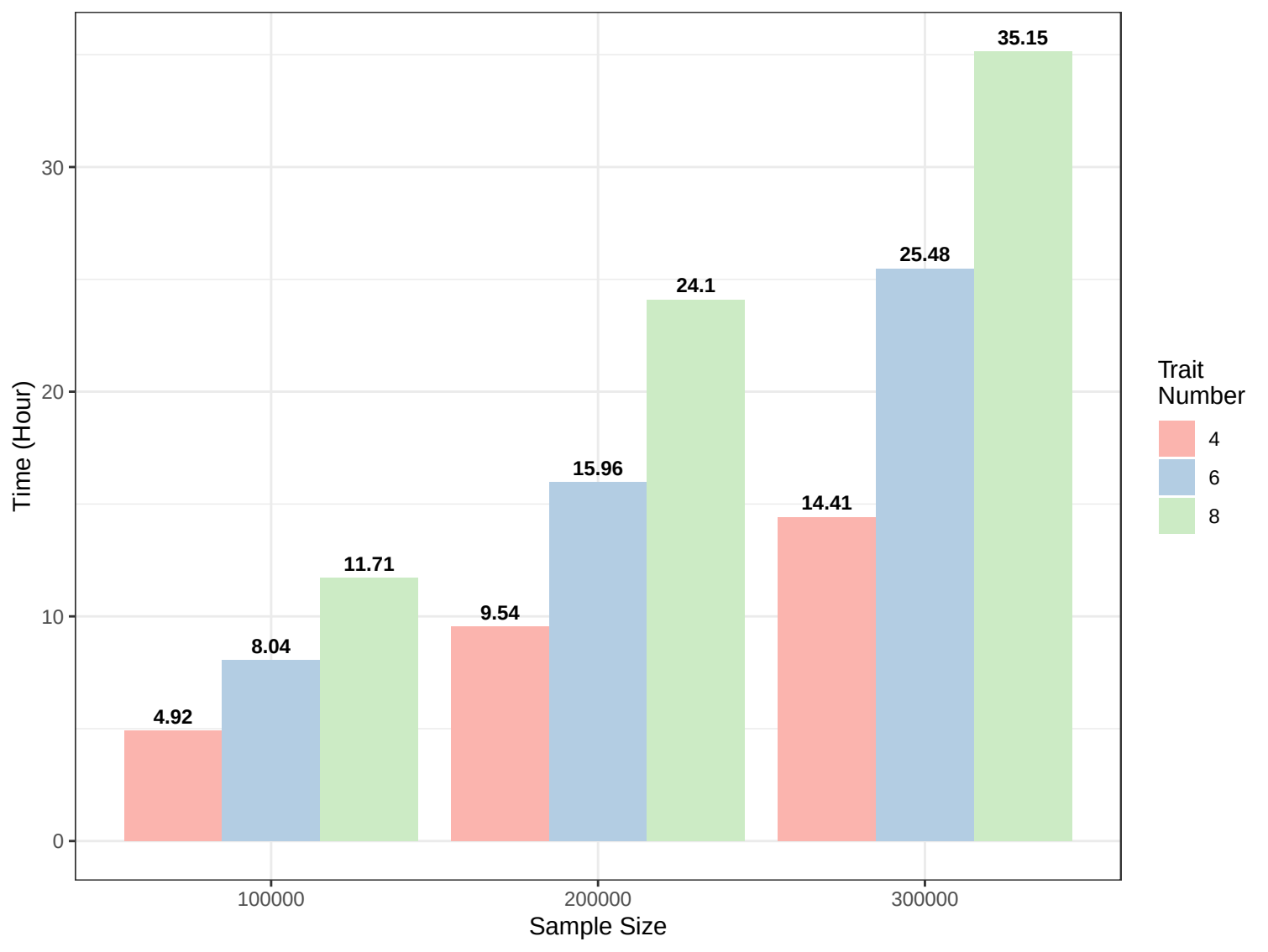

N = 288,709 Independent Subjects in UKBB
SCAMPI with SNP QC (MAF >= 0.05), SNPs > 10% missing are removed
−log10(P−value)
Chromosome
rs7528419
rs525172
rs10484439
rs2001945
rs3890182
rs964184
rs261334
rs4783961
rs55791371
rs73307905
rs73840

SCAMPI with SNP QC (MAF >= 0.05),
SNPs > 10% missing are removed
N = 241,167 Independent Subjects in UKBB




80
60
40
20
0
Time (Hour)
11.65
18.65
27.52
22.5
36.45
52.88
33.62
54.18
79.55
100000
200000
300000
Sample Size
Trait Number
4
6
8


Computation Time per SNP (Second)
20
15
10
5
0
Computation Time for 300K SNPs with 1,000 Instances (Hour)
1.5
1.0
0.5
0.0
100000
200000
300000
Nsub
Trait Number
4
6
8


Time (Hour)
0
10
20
30
4.92
8.04
11.71
9.54
15.96
24.1
14.41
25.48
35.15
100000
200000
300000
Sample Size
Trait Number
4
6
8