Supplementary Materials: Testing for genetic associations in arbitrarily structured populations

Minsun Song, Wei Hao, and John D. Storey

Contents

SUPPLEMENTARY NOTE

This supplementary note assumes the reader is familiar with the details and mathematical notation from Online Methods.

1 Logistic Factor Analysis (LFA)

When forming a latent variable model of structure, where the goal is to make minimal assumptions about the underlying structure, there are benefits to modeling $\mathrm{logit}(\pi_{ij})$ in terms of a latent variable model instead of π_{ij} directly [1]. The quantity $\mathrm{logit}(\pi_{ij}) = \log(\pi_{ij}/(1-\pi_{ij}))$ is called the "natural parameter" of the distribution of x_{ij} when we assume Hardy-Weinberg equilibrium so that $x_{ij} \sim$ Binomial$(2, \pi_{ij})$. The quantity $\mathrm{logit}(\pi_{ij})$ occurs as a linear term in the log-likelihood of the data, and it is the target parameter in logistic regression because of its straightforward mathematical properties. This viewpoint also facilitates calculating the distribution of x_{ij} given the structure, which is the essential challenge in accounting for structure in the proposed association testing framework.

In the association testing framework we have developed, it turns out that developing a latent variable model and estimate of the $\mathrm{logit}(\pi_{ij})$ is particularly appropriate. The approach is called "logistic factor analysis" (LFA). Let $\mathbf{L}$ be an $m \times n$ matrix with (i,j) element equal to $\mathrm{logit}(\pi_{ij})$. Consider the following parameterization:

$$\mathbf{L} = \mathbf{A}\mathbf{H}, \tag{4}$$

where $\mathbf{A}$ is an $m \times d$ matrix, $\mathbf{H}$ is a $d \times n$ matrix, and $d \ll n$. The columns of $\mathbf{H}$ are independent, and column j captures the structure information for individual j. That is, $\Pr(x_{ij}|\mathbf{h}^j, \boldsymbol{z}_j) = \Pr(x_{ij}|\mathbf{h}^j)$ where $\mathbf{h}^j$ is column j of $\mathbf{H}$. Row i of $\mathbf{A}$ determines how SNP i is affected by structure. We have shown in ref. [1] that this model performs well in estimating structure resulting from discrete subpopulations, admixed populations, the Balding-Nichols model [2], the Pritchard-Stephens-Donnelly model [3], and models of spatially oriented structure.

In practice, $\mathbf{H}$ will be unknown, so it must be estimated. We have developed a method called logistic factor analysis (LFA) that we have shown to estimate $\mathbf{H}$ well [1]. Specifically, the LFA estimate $\widehat{\mathbf{H}}$ has been shown to span the same space as the true $\mathbf{H}$ at a high level of accuracy, which implies that replacing $\mathbf{H}$ with $\widehat{\mathbf{H}}$ in the above equations yields nearly identical results. The accuracy of $\widehat{\mathbf{H}}$ in estimating $\mathbf{H}$ has been demonstrated even when the individual-specific allele frequencies are not directly constructed from model (4), $\mathbf{L} = \mathbf{A}\mathbf{H}$.

2 Proposed Association Testing Framework

We have derived a statistical hypothesis test of association that is equivalent to testing whether $\beta_i = 0$ for each SNP i in the trait models (1) and (2) (defined in Online Methods), and whose null distribution does not depend on structure or the non-genetic effects correlated with structure, making it immune to spurious associations due to structure. Specifically, the test allows for general levels of complexity in structure because the test is based on adjusting for structure according to individual-specific allele frequencies.

A Model of Genetic Variation Given the Trait and Structure. As a first step, we have proved a theorem (see below) that shows that $\beta_i = 0$ in models (1) and (2) implies that $b_i = 0$ in the following model:

$$\begin{aligned} x_{ij}|y_j, \boldsymbol{z}_j &\sim \text{Binomial}\left(2, \text{logit}^{-1}(a_i + b_i y_j + \text{logit}(\pi_{ij}))\right), \\ \text{logit}\left(\frac{\text{E}[x_{ij}|y_j, \boldsymbol{z}_j]}{2}\right) &= a_i + b_i y_j + \text{logit}(\pi_{ij}) \end{aligned} \tag{5}$$

for all $j = 1, 2, \ldots, n$. This establishes a model that can be used to test for associations in place of models (1) and (2).

There are a few important details to note. First, the variables λ_j, σ_j^2, and $(x_{kj})_{k \neq i}$ do not appear in the model. This is important because it is impossible to estimate λ_j and σ_j^2 in the typical setting, and we will also typically not know the polygenic $\sum_{k \neq i} \beta_k x_{kj}$ component of the model. Second, the genotype variation is being modeled in terms of the trait variation, instead of the other way around. It is initially counter-intuitive because almost all association tests involve modeling the trait in terms of the SNP genotypes. As explained in more detail below, this reversal is crucial for adjusting the probability distribution of x_{ij} according to structure, and for eliminating the need to estimate λ_j, σ_j^2, and $(\beta_k)_{k \neq i}$.

We call our proposed test the "genotype conditional association test" (GCAT). The model we propose to utilize is sometimes called an inverse regression model because we utilize $\text{E}[x|y]$ rather than $\text{E}[y|x]$.

Proposed Test Conditional on Individual-Specific Allele Frequencies. As a second step, we have derived a test-statistic to test whether $b_i = 0$ in model (3) (defined in Online Methods) whose null distribution is immune to structure. The log-likelihood function of the parameters given individual j is

$$\ell(a_i, b_i | x_{ij}, y_j, \pi_{ij}) \propto \log\left(\Pr(x_{ij}|y_j, a_i, b_i, \pi_{ij})\right)$$

where the probability on the right-hand-side is calculated according to model (3). The log-likelihood of all n individuals is

$$\ell(a_i, b_i|\mathbf{x}_i, \mathbf{y}, \boldsymbol{\pi}_i) = \sum_{j=1}^{n} \ell(a_i, b_i|x_{ij}, y_j, \pi_{ij}) \propto \log \left[\prod_{j=1}^{n} \Pr(x_{ij}, y_j|a_i, b_i, \pi_{ij}) \right],$$

where $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{i,n})$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. The test statistic we utilize is a generalized likelihood ratio test statistic [4]:

$$T(\mathbf{x}_i, \mathbf{y}, \boldsymbol{\pi}_i) = 2 \left[\max_{a_i, b_i} \ell(a_i, b_i|\mathbf{x}_i, \mathbf{y}, \boldsymbol{\pi}_i) - \max_{a_i} \ell(a_i, b_i = 0|\mathbf{x}_i, \mathbf{y}, \boldsymbol{\pi}_i) \right]. \tag{6}$$

The log-likelihood is maximized by performing a logistic regression of all n observed genotypes for SNP i on the right hand side of model (3). We have proven a theorem below that shows that when $\beta_i = 0$ in models (1) or (2), the null distribution of this test statistic is χ^2_1, regardless of the values of π_{ij}, $(x_{kj})_{k \neq i}$, $(\beta_{kj})_{k \neq i}$, λ_j, and σ^2_j for $j = 1, 2, \ldots, n$ in models (1) and (2).

Proposed Test In Terms of LFA Model. As a third step, we have extended the above results to the case where the individual-specific allele frequencies are unknown and must be estimated. This requires a model of the individual-specific allele frequencies, and we utilize model (4) so that $\text{logit}(\pi_{ij}) = \sum_{k=1}^{d} a_{ik} h_{kj}$. First, assume that $\mathbf{H}$ from model (4) is known. We have proved that $\beta_i = 0$ in models (1) and (2) implies $b_i = 0$ in the following model:

$$\begin{aligned} x_{ij}|y_j, \boldsymbol{z}_j &\sim \text{Binomial} \left(2, \text{logit}^{-1} \left(\sum_{k=1}^{d} a_{ik} h_{kj} + b_i y_j \right) \right), \\ \text{logit} \left(\frac{\text{E}\left[x_{ij}|y_j, \boldsymbol{z}_j \right]}{2} \right) &= \sum_{k=1}^{d} a_{ik} h_{kj} + b_i y_j \end{aligned} \tag{7}$$

for all $j = 1, 2, \ldots, n$, where $\mathbf{h}^j$ is column j of $\mathbf{H}$ and it is noted that without loss of generality we let $h_{dj} = 1$ making a_{id} an intercept term. The test-statistic used to test for an association between SNP i and the trait is the following generalized likelihood ratio test statistic:

$$T(\mathbf{x}_i, \mathbf{y}, \mathbf{H}) = 2 \left[\max_{\boldsymbol{a}_i, b_i} \ell(\boldsymbol{a}_i, b_i|\mathbf{x}_i, \mathbf{y}, \mathbf{H}) - \max_{\boldsymbol{a}_i} \ell(a_i, b_i = 0|\mathbf{x}_i, \mathbf{y}, \mathbf{H}) \right], \tag{8}$$

where $\boldsymbol{a}_i = (a_{i1}, a_{i2}, \ldots, a_{i,d})$. The log-likelihoods in this test statistic are maximized by performing a logistic regression of all n observed genotypes for SNP i on the right hand side of model (7) on all n individuals. As the previous case, we have proven a theorem below that shows that when $\beta_i = 0$ in models (1) or (2), the null distribution of this test statistic is χ^2_1, regardless of the values of $\boldsymbol{\pi}_i$, $(x_{kj})_{k \neq i}$, $\boldsymbol{\beta}_{-i}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\sigma}^2$ in models (1) and (2).

The proposed test utilizes LFA to form an estimate $\widehat{\mathbf{H}}$, replaces $\mathbf{H}$ with $\widehat{\mathbf{H}}$, and carries out the test using model (7) and test statistic (8): $T(\mathbf{x}_i, \mathbf{y}, \widehat{\mathbf{H}})$. This approach directly allows the simultaneous estimation of $\boldsymbol{a}_i$ and b_i for each SNP i under the unconstrained model and the estimation of $\boldsymbol{a}_i$ with $b_i = 0$ under the constraints of the null hypothesis. Because of this, the test allows the uncertainty of the $m \times d$ unknown parameters of $\mathbf{A}$ to be taken into account and it allows b_i to be competitively fit with $\boldsymbol{a}_i$ under the unconstrained, alternative hypothesis model.

Another approach is to first carry out estimation of $\mathbf{F}$ by whatever method the analyst finds appropriate and then base the test on statistic (6) with the π_{ij} replaced with the estimates $\widehat{\pi}_{ij}$: $T(\mathbf{x}_i, \mathbf{y}, \widehat{\boldsymbol{\pi}}_i)$. This has the advantage that it allows for a much broader class of methods to estimate $\mathbf{F}$, but it may be more conservative than the above implementation because b_i is not competitively fit with the π_{ij} under the unconstrained model. In this case, $\mathbf{F}$ may be estimated in a manner that allows for fine-scale levels of inter-individual coancestry and locus-specific models of structure without relying on the lower d-dimensional factorized model $\mathbf{L} = \mathbf{AH}$ that we used here.

Proposed Test Under the Alternative Hypothesis. The proposed association test is based on models (3) and (7). Even though we have proved that the test is immune to population structure, it is also important to demonstrate that the test has favorable statistical power to identify true associations. We have shown that the $\operatorname{logit}\left(\frac{\mathrm{E}[x_{ij}|y_j, \boldsymbol{z}_j]}{2}\right) = a_i + \operatorname{logit}(\pi_{ij}) + b_i y_j$ is a tractable approximation of the model under general configurations of a true alternative hypothesis for SNP i where $\beta_i \neq 0$ (see below). This provides the beginnings of a mathematical framework for characterizing the power of the test.

3 Theorems and Proofs

Because $x_{ij}|\boldsymbol{z}_j \sim$ Binomial$(2, \pi_i(\boldsymbol{z}_j))$ where we write $\pi_{ij} \equiv \pi_i(\boldsymbol{z}_j)$, it follows that $\Pr(x_{ij}|\pi_{ij}, \boldsymbol{z}_j) = \Pr(x_{ij}|\pi_{ij})$. We assume that $\Pr(x_{ij}|\mathbf{h}^j, \boldsymbol{z}_j) = \Pr(x_{ij}|\mathbf{h}^j)$; in other words, all information about the influence of population structure on the genotypes of individual j is captured through column j of $\mathbf{H}$. It therefore follows that $\Pr(x_{ij}|\pi_{ij}, \mathbf{h}^j, \boldsymbol{z}_j) = \Pr(x_{ij}|\pi_{ij}, \mathbf{h}^j) = \Pr(x_{ij}|\pi_{ij})$. We also assume that the SNP genotypes are mutually independent given the structure (which also implies the set of SNPs we consider are in linkage equilibrium, given the structure). These

assumptions yield the following equalities:

$$\Pr(\mathbf{X}|\mathbf{L},\mathbf{H},(\boldsymbol{z}_k)_{k=1}^n) = \Pr(\mathbf{X}|\mathbf{L},\mathbf{H}) = \Pr(\mathbf{X}|\mathbf{L})$$
$$\Pr(\mathbf{X}|(\boldsymbol{z}_k)_{k=1}^n) = \prod_{i=1}^m \prod_{j=1}^n \Pr(x_{ij}|(\boldsymbol{z}_k)_{k=1}^n) = \prod_{i=1}^m \prod_{j=1}^n \Pr(x_{ij}|\boldsymbol{z}_j)$$
$$\Pr(\mathbf{X}|\mathbf{L}) = \prod_{i=1}^m \prod_{j=1}^n \Pr(x_{ij}|\mathbf{L}) = \prod_{i=1}^m \prod_{j=1}^n \Pr(x_{ij}|\pi_{ij})$$
$$\Pr(\mathbf{X}|\mathbf{H}) = \prod_{i=1}^m \prod_{j=1}^n \Pr(x_{ij}|\mathbf{H}) = \prod_{i=1}^m \prod_{j=1}^n \Pr(x_{ij}|\mathbf{h}^j)$$

Theorem 1 *Suppose that* y_j *is distributed according to model* (1) *or* (2)*,* $x_{ij}|\pi_{ij} \sim$ *Binomial*$(2,\pi_{ij})$ *as parameterized above, and the SNP genotypes are mutually independent given the structure as detailed above. Then* $\beta_i = 0$ *in models* (1) *or* (2) *implies that* $b_i = 0$ *in model* (3)*.*

Note: We provide two proofs of this theorem because both provide relevant insights. The first version gives insight into the probabilistic mechanism underlying the proposed approach and has some generality beyond the modeling assumptions made here. The second version directly shows how the terms in models (1) and (2) relate to those in model (3).

Proof (version 1): When $\beta_i = 0$, it follows that $\Pr(y_j|(x_{kj})_{k\neq i}, x_{ij}, \boldsymbol{z}_j) = \Pr(y_j|(x_{kj})_{k\neq i}, \boldsymbol{z}_j)$ by the assumptions of models (1) and (2). Noting that $\Pr((x_{kj})_{k\neq i}|x_{ij},\boldsymbol{z}_j) = \Pr((x_{kj})_{k\neq i}|\boldsymbol{z}_j)$ by the conditional independence assumption, we have:

$$\begin{aligned}\Pr(y_j|x_{ij},\boldsymbol{z}_j) &= \int \Pr(y_j|(x_{kj})_{k\neq i}, x_{ij}, \boldsymbol{z}_j)\Pr((x_{kj})_{k\neq i}|x_{ij},\boldsymbol{z}_j)dP \\ &= \int \Pr(y_j|(x_{kj})_{k\neq i}, \boldsymbol{z}_j)\Pr((x_{kj})_{k\neq i}|\boldsymbol{z}_j)dP \\ &= \Pr(y_j|\boldsymbol{z}_j). \end{aligned} \quad (9)$$

By Bayes theorem we have

$$\Pr(x_{ij}|y_j,\boldsymbol{z}_j) = \frac{\Pr(y_j|x_{ij},\boldsymbol{z}_j)\Pr(x_{ij}|\boldsymbol{z}_j)}{\Pr(y_j|\boldsymbol{z}_j)}.$$

Since $\Pr(y_j|x_{ij},\boldsymbol{z}_j) = \Pr(y_j|\boldsymbol{z}_j)$, this implies that $\Pr(x_{ij}|y_j,\boldsymbol{z}_j) = \Pr(x_{ij}|\boldsymbol{z}_j)$ and it follows that $b_i = 0$ in model (3).

Proof (version 2): For either model (1) or (2), it follows that

$$\begin{aligned}\log \frac{\Pr(x_{ij}=1|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=0|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} &= \log \frac{\Pr(x_{ij}=1|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=0|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} \\ &= \log \frac{\Pr(y_j|x_{ij}=1,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(y_j|x_{ij}=0,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} + \log \frac{\Pr(x_{ij}=1|(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=0|(x_{kj})_{k\neq i},\boldsymbol{z}_j)}\end{aligned} \tag{10}$$

and similarly

$$\log \frac{\Pr(x_{ij}=2|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=1|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} = \log \frac{\Pr(y_j|x_{ij}=2,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(y_j|x_{ij}=1,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} + \log \frac{\Pr(x_{ij}=2|(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=1|(x_{kj})_{k\neq i},\boldsymbol{z}_j)}.$$

By the assumptions detailed above, we have $\Pr(x_{ij}|(x_{kj})_{k\neq i},\boldsymbol{z}_j) = \Pr(x_{ij}|\pi_{ij})$ and therefore:

$$\log \frac{\Pr(x_{ij}=1|(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=0|(x_{kj})_{k\neq i},\boldsymbol{z}_j)} = \log \frac{\pi_{ij}}{1-\pi_{ij}} + \log 2,$$

$$\log \frac{\Pr(x_{ij}=2|(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=1|(x_{kj})_{k\neq i},\boldsymbol{z}_j)} = \log \frac{\pi_{ij}}{1-\pi_{ij}} - \log 2.$$

Under the *quantitative trait* model (1), it follows that

$$\log \frac{\Pr(y_j|x_{ij}=1,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(y_j|x_{ij}=0,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} = \frac{-\beta_i(\beta_i+2\alpha)}{2\sigma_j^2} + \sum_{l\neq i}\frac{-\beta_l\beta_i}{\sigma_j^2}x_{lj} + \frac{-\beta_i}{\sigma_j^2}\lambda_j + \frac{\beta_i}{\sigma_j^2}y_j.$$

Plugging this back into equation (10) shows that

$$\log \frac{\Pr(x_{ij}=1|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=0|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} = a_{ij} + b_{ij}y_j + \mathrm{logit}(\pi_{ij}) + \log(2),$$

where $a_{ij} = \frac{-\beta_i\left(\beta_i/2+\alpha+\sum_{k\neq i}\beta_k x_{kj}+\lambda_j\right)}{\sigma_j^2}$ and $b_{ij} = \frac{\beta_i}{\sigma_j^2}$. Following analogous steps, we find that

$$\log \frac{\Pr(x_{ij}=2|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=1|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} = \tilde{a}_{ij} + b_{ij}y_j + \mathrm{logit}(\pi_{ij}) - \log(2),$$

where $\tilde{a}_{ij} = a_{ij} - \frac{\beta_i^2}{\sigma_j^2}$. When $\beta_i = 0$ in model (1), then $a_{ij} = \tilde{a}_{ij} = b_{ij} = 0$.

Under the *binary trait* model (2), it follows that

$$\log \frac{\Pr(y_j|x_{ij}=1,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(y_j|x_{ij}=0,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} = a_{ij} + b_i y_j,$$

where $a_{ij} = \log \frac{1+\exp(\alpha+\sum_{l\neq i}\beta_l x_{lj}+\lambda_j)}{1+\exp(\alpha+\beta_i+\sum_{l\neq i}\beta_l x_{lj}+\lambda_j)}$ and $b_i = \beta_i$. Plugging this back into equation (10) shows that

$$\log \frac{\Pr(x_{ij}=1|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=0|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} = a_{ij} + b_i y_j + \mathrm{logit}(\pi_{ij}) + \log(2).$$

Following analogous steps, we find that

$$\log \frac{\Pr(x_{ij}=2|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)}{\Pr(x_{ij}=1|y_j,(x_{kj})_{k\neq i},\boldsymbol{z}_j)} = \tilde{a}_{ij} + b_i y_j + \text{logit}(\pi_{ij}) - \log(2),$$

where $\tilde{a}_{ij} = \log \frac{1+\exp(\alpha+\beta_i+\sum_{l\neq i}\beta_l x_{lj}+\lambda_j)}{1+\exp(\alpha+2\beta_i+\sum_{l\neq i}\beta_l x_{lj}+\lambda_j)}$. When $\beta_i = 0$ in model (2), then $a_{ij} = \tilde{a}_{ij} = b_i = 0$.

Putting these together, we have that when $\beta_i = 0$ in models (1) or (2), then model (3) holds with $b_i = 0$.

Corollary 1 *Suppose that the assumptions of Theorem 1 hold and additionally* $\text{logit}(\pi_{ij}) = \sum_{k=1}^{d} a_{ik}h_{kj}$*. Then* $\beta_i = 0$ *in models* (1) *or* (2) *implies that* $b_i = 0$ *in model* (7)*.*

Proof: The proof is the same as that to Theorem 1, except we replace π_{ij} with $\mathbf{h}^j$.

Theorem 2 *Suppose that* y_j *is distributed according to model* (1) *or* (2) *and that* $x_{ij}|\pi_{ij} \sim$ *Binomial*$(2,\pi_{ij})$*. If* $\beta_i = 0$ *in models* (1) *or* (2)*, then the test-statistic* $T(\mathbf{x}_i,\mathbf{y},\boldsymbol{\pi}_i)$ *defined in* (6) *converges in distribution to* χ^2_1 *as* $n \to \infty$*.*

Proof: When $\beta_i = 0$, then $[x_{ij}|y_j,\pi_{ij}] \sim$ Binomial $(2,\pi_{ij})$ by Theorem 1. It then follows that $T(\mathbf{x}_i,\mathbf{y},\boldsymbol{\pi}_i) \to \chi^2_1$ in distribution as $n \to \infty$ by Wilks' theorem [4].

Corollary 2 *Suppose that the assumptions of Theorem 1 hold and additionally* $\text{logit}(\pi_{ij}) = \sum_{k=1}^{d} a_{ik}h_{kj}$*. If* $\beta_i = 0$ *in models* (1) *or* (2)*, then the test-statistic* $T(\mathbf{x}_i,\mathbf{y},\mathbf{H})$ *defined in* (8) *converges in distribution to* χ^2_1 *as* $n \to \infty$*.*

Proof: When $\beta_i = 0$, then $[x_{ij}|y_j,\mathbf{h}^j] \sim$ Binomial $\left(2,\text{logit}^{-1}(\sum_{k=1}^{d} a_{ik}h_{kj})\right)$ by Corollary 1. It then follows that $T(\mathbf{x}_i,\mathbf{y},\mathbf{H}) \to \chi^2_1$ in distribution as $n \to \infty$ by Wilks' theorem [4].

4 Proposed Model Under the Alternative Hypothesis

When the alternative model is true this means that $\beta_i \neq 0$. In this case it is worthwhile to characterize model (3) in terms of the distribution of $x_{ij}|y_j,\boldsymbol{z}_j$. Under trait models (1) or (2), it

follows that:

$$\begin{aligned}
\text{logit}\left(\frac{\text{E}[x_{ij}|y_j, \boldsymbol{z}_j]}{2}\right) &= \log\left(\frac{\frac{1}{2}\text{Pr}(x_{ij}=1|y_j,\boldsymbol{z}_j)+\text{Pr}(x_{ij}=2|y_j,\boldsymbol{z}_j)}{1-\frac{1}{2}\text{Pr}(x_{ij}=1|y_j,\boldsymbol{z}_j)-\text{Pr}(x_{ij}=2|y_j,\boldsymbol{z}_j)}\right) \\
&= \log\left(\frac{\frac{1}{2}\text{Pr}(x_{ij}=1|y_j,\boldsymbol{z}_j)+\text{Pr}(x_{ij}=2|y_j,\boldsymbol{z}_j)}{\frac{1}{2}\text{Pr}(x_{ij}=1|y_j,\boldsymbol{z}_j)+\text{Pr}(x_{ij}=0|y_j,\boldsymbol{z}_j)}\right) \\
&= \log\left(\frac{\frac{1}{2}+\frac{\text{Pr}(x_{ij}=2|y_j,\boldsymbol{z}_j)}{\text{Pr}(x_{ij}=1|y_j,\boldsymbol{z}_j)}}{\frac{1}{2}+\frac{\text{Pr}(x_{ij}=0|y_j,\boldsymbol{z}_j)}{\text{Pr}(x_{ij}=1|y_j,\boldsymbol{z}_j)}}\right)
\end{aligned}$$

This implies that

$$\text{logit}\left(\frac{\text{E}[x_{ij}|y_j, \boldsymbol{z}_j]}{2}\right) = \log\left(\frac{1+\exp\left\{\tilde{a}_{ij}+b_{ij}y_j+\text{logit}(\pi_{ij})\right\}}{1+\exp\left\{-(a_{ij}+b_{ij}y_j+\text{logit}(\pi_{ij}))\right\}}\right),$$

where under model (1) we have $a_{ij} = \frac{-\beta_i\left(\beta_i/2+\alpha+\sum_{k\neq i}\beta_k x_{kj}+\lambda_j\right)}{\sigma_j^2}$, $\tilde{a}_{ij} = a_{ij} - \frac{\beta_i^2}{\sigma_j^2}$, $b_{ij} = \frac{\beta_i}{\sigma_j^2}$ and under model (2) we have $a_{ij} = \log\frac{1+\exp(\alpha+\sum_{l\neq i}\beta_l x_{lj}+\lambda_j)}{1+\exp(\alpha+\beta_i+\sum_{l\neq i}\beta_l x_{lj}+\lambda_j)}$, $\tilde{a}_{ij} = \log\frac{1+\exp(\alpha+\beta_i+\sum_{l\neq i}\beta_l x_{lj}+\lambda_j)}{1+\exp(\alpha+2\beta_i+\sum_{l\neq i}\beta_l x_{lj}+\lambda_j)}$, $b_{ij} = \beta_i$.

In the case that $a_{ij} = \tilde{a}_{ij}$, it is the case that

$$\text{logit}\left(\frac{\text{E}[x_{ij}|y_j, \boldsymbol{z}_j]}{2}\right) = a_{ij} + b_{ij}y_j + \text{logit}(\pi_{ij}).$$

However, this exact equality is only the case when $\beta_i = 0$. For the typical effect sizes seen in GWAS, it will nevertheless be true that $a_{ij} \approx \tilde{a}_{ij}$, in which case the above functional form will be approximately true. This allows for an approximation that can be utilized in practice for power calcuations.

5 Simulated Allele Frequencies

In order to simulate the $m \times n$ matrix of genotypes $\mathbf{X}$, we first needed to simulate the $m \times n$ matrix of allele frequencies $\mathbf{F}$. Recall that we model the allele frequencies by forming $\mathbf{L} = \text{logit}(\mathbf{F})$ and then utilizing the model $\mathbf{L} = \mathbf{AH}$ from equation (4).

Instead of simulating allele frequencies from the $\mathbf{L} = \mathbf{AH}$ model we use to perform the proposed association test, we instead simulated them from a different model to demonstrate the flexibility of the $\mathbf{L} = \mathbf{AH}$ model. Specifically, we let $\mathbf{F} = \mathbf{\Gamma S}$ where $\mathbf{\Gamma}$ is $m \times d$ and $\mathbf{S}$ is $d \times n$ with $d \leq n$. The $d \times n$ matrix $\mathbf{S}$ encapsulates the genetic population structure for these individuals since $\mathbf{S}$ is not SNP-specific but is shared across SNPs. The $m \times d$ matrix $\mathbf{\Gamma}$ maps

how the structure is manifested in the allele frequencies of each SNP. We have shown that the model $\mathbf{F} = \mathbf{\Gamma S}$ includes as special cases discrete subpopulations, the Balding-Nichols model, and the Pritchard-Stephens-Donnelly model.

We formed $\mathbf{\Gamma}$ and $\mathbf{S}$ for the 11 different population structure configurations exactly as carried out in Hao et al. (2013) [1]. These constructions are summarized as follows from Hao et al. (2013).

Balding-Nichols Model (Balding-Nichols). The HapMap data set was deliberately sampled to be from three discrete populations, which allowed us to populate each row i of Γ with three independent and identically distributed draws from the Balding-Nichols model: $\boldsymbol{\gamma}_{ik} \overset{i.i.d.}{\sim}$ BN(p_i, F_i), where $k \in \{1, 2, 3\}$. Each $\boldsymbol{\gamma}_{ik}$ is interpreted to be the allele frequency for subpopulation k at SNP i. The pairs (p_i, F_i) were computed by randomly selecting a SNP in the HapMap data set, calculating its observed allele frequency, and estimating its $\mathrm{F_{ST}}$ value using the Weir & Cockerham estimator [5]. The columns of $\mathbf{S}$ were populated with indicator vectors such that each individual was assigned to one of the three subpopulations. The subpopulation assignments were drawn independently with probabilities $60/210$, $60/210$, and $90/210$, which reflect the subpopulation proportions in the HapMap data set. The dimensions of the simulated data were $m = 100,000$ SNPs and $n = 5000$ individuals.

1000 Genomes Project (TGP). We started with the TGP data set from Hao et al. (2013) [1]. The matrix $\mathbf{\Gamma}$ was generated by sampling $\gamma_{ik} \overset{i.i.d.}{\sim} 0.9 \times$ Uniform$(0, 1/2)$ for $k = 1, 2$ and setting $\gamma_{i3} = 0.05$. In order to generate $\mathbf{S}$, we computed the first two principal components of the TGP genotype matrix after mean centering each SNP. We then transformed each principal component to be between $(0, 1)$ and set the first two rows of $\mathbf{S}$ to be the transformed principal components. The third row of $\mathbf{S}$ was set to 1, i.e. an intercept. The dimensions of the simulated data were $m = 100,000$ and $n = 1500$, where n was determined by the number of individuals in the TGP data set.

Human Genome Diversity Project (HGDP). We started with the HGDP data set from Hao et al. (2013) [1] and applied the same simulation scheme as for the TGP scenario. The dimensions of the simulated data were $m = 100,000$ and $n = 940$, where n was determined by the number of individuals in the HGDP data set.

Pritchard-Stephens-Donnelly (PSD). The PSD model assumes individuals to be an admixture of ancestral subpopulations. The rows of $\mathbf{\Gamma}$ were again created by three independent and identically distributed draws from the Balding-Nichols model: $\boldsymbol{\gamma}_{ik} \overset{i.i.d.}{\sim}$ BN(p_i, F_i), where

$k \in \{1, 2, 3\}$. For this scenario, the pairs (p_i, F_i) were computed from analyzing the HGDP data set for observed allele frequency and estimated $\mathrm{F_{ST}}$ via the Weir & Cockerham estimate [5]. The estimator requires each individual to be assigned to a subpopulation, which were made according to the $K = 5$ subpopulations from the analysis in Rosenberg et al. (2002) [6]. The columns of $\mathbf{S}$ were sampled $(s_{1j}, s_{2j}, s_{3j}) \overset{i.i.d.}{\sim}$ Dirichlet$(\boldsymbol{\alpha})$ for $j = 1, \ldots, n$. There were four PSD scenarios with parameter values $\boldsymbol{\alpha} = (0.01, 0.01, 0.01)$, $\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$, $\boldsymbol{\alpha} = (0.5, 0.5, 0.5)$, and $\boldsymbol{\alpha} = (1, 1, 1)$. $\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$ was chosen as the representative structure for Figure 2. The dimensions of the simulated data were $m = 100,000$ SNPs and $n = 5000$ individuals.

Spatial. We seek to simulate genotypes such that the population structure relates to the spatial position of the individuals. The matrix $\mathbf{\Gamma}$ was populated by sampling $\gamma_{ik} \overset{i.i.d.}{\sim} 0.9 \times$ Uniform$(0, 1/2)$ for $k = 1, 2$ and setting $\gamma_{i3} = 0.05$. The first two rows of $\mathbf{S}$ correspond to coordinates for each individual on the unit square and were set to be independent and identically distributed samples from Beta(a, a), while the third row of $\mathbf{S}$ was set to be 1, i.e. an intercept. There were four spatial scenarios with parameter values of $a = 0.1, 0.25, 0.5,$ and 1. As $a \to 0$, the individuals are placed closer to the corners of the unit square, while when $a = 1$, the individuals are distributed uniformly. $a = 0.1$ was chosen as the representative structure for Figure 2. The dimensions of the simulated data were $m = 100,000$ SNPs and $n = 5000$ individuals.

6 Simulated Traits

For each of the 11 simulations scenarios, we generated 100 independent studies. For each study, $\mathbf{X}$ was formed by simulating $x_{ij} \sim$ Binomial$(2, \pi_{ij})$ where $\mathbf{F}$ was constructed as described above. In order to simulate a quantitative trait, we needed to simulate α, $\sum_{i=1}^{m} \beta_i x_{ij}$, λ_j, and ϵ_j from model (1).

First, we set $\alpha = 0$. Without loss of generality SNPs $i = 1, 2, \ldots, 10$ were set to be true alternative SNPs (where $\beta_i \neq 0$); we simulated $\beta_i \overset{i.i.d.}{\sim}$ Normal$(0, 0.5)$ for $i = 1, 2, \ldots, 10$. We set $\beta_i = 0$ for $i > 10$. Note that $\mathbf{X}$ is influenced by the latent variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ through $\mathbf{S}$ in the model $\mathbf{F} = \mathbf{\Gamma S}$ described above. In order to simulate λ_j and ϵ_j so that they are also influenced by the latent variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, we performed the following:

1. Perform K-means clustering on the columns of $\mathbf{S}$ with $K = 3$ using Euclidean distance. This assigns each individual j to one of three mutually exclusive cluster sets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$

where $\mathcal{S}_k \subset \{1, 2, \ldots, n\}$.

2. Set $\lambda_j = k$ for all $j \in \mathcal{S}_k$ for each $k = 1, 2, 3$.

3. Let $\tau_1^2, \tau_2^2, \tau_3^2 \overset{i.i.d.}{\sim}$ InvGamma$(3, 1)$ and set $\sigma_j^2 = \tau_k^2$ for all $j \in \mathcal{S}_i$ for each $k = 1, 2, 3$.

4. Draw $\epsilon_j \sim$ Normal$(0, \sigma_j^2)$ independently for $j = 1, 2, \ldots, n$.

This strategy simulates non-genetic effects and random variation that manifest among K discrete groups over a more continuous population genetic structure defined by $\mathbf{S}$. This is meant to emulate the fact that environment (specifically lifestyle) may partition among individuals in a manner distinct from, but highly related to population structure.

This yields three values $\sum_{i=1}^m \beta_i x_{ij}$, λ_j, and ϵ_j for each individual $j = 1, 2, \ldots, n$. In order to set the variances of these three values to pre specified levels ν_{gen}, ν_{env} and ν_{noise}, we rescaled each quantity as follows:

$$\sum_{i=1}^m \beta_i x_{ij} \leftarrow \left[\frac{\sqrt{\nu_{\text{gen}}}}{\text{s.d.}\left\{\sum_{i=1}^m \beta_i x_{ik}\right\}_{k=1}^n} \right] \sum_{i=1}^m \beta_i x_{ij}$$

$$\lambda_j \leftarrow \left[\frac{\sqrt{\nu_{\text{env}}}}{\sqrt{\frac{\sum_{k=1}^n (\lambda_k - \bar{\lambda})^2}{n-1}}} \right] \lambda_j$$

$$\epsilon_j \leftarrow \left[\frac{\sqrt{\nu_{\text{noise}}}}{\sqrt{\frac{\sum_{k=1}^n (\epsilon_k - \bar{\epsilon})^2}{n-1}}} \right] \epsilon_j$$

The trait for a given study was then formed according to

$$y_j = \sum_{i=1}^m \beta_i x_{ij} + \lambda_j + \epsilon_j$$

for $j = 1, 2, \ldots, n$. For each of the 11 simulation scenarios, we considered the following three configurations of $(\nu_{\text{gen}}, \nu_{\text{env}}, \nu_{\text{noise}})$: (5%, 5%, 90%), (10%, 0%, 90%) and (10%, 20%, 70%).

In total, there were 11 different types of structures considered over three different configurations of genetic, environmental, and noise variances for a total of 33 settings. For each setting, we simulated 100 independent studies where each involved $m = 100,000$ SNPs and up to $n = 5000$ individuals.

7 Northern Finland Birth Cohort Data

Genotype data was downloaded from dbGaP (Study Accession: phs000276.v1.p1). Individuals were filtered for completeness (maximum 1% missing genotypes) and pregnancy. (Pregnant women were excluded because we did not receive IRB approval for these individuals.) SNPs were first filtered for completeness (maximum 5% missing genotypes) and minor allele frequency (minimum 1% minor allele frequency), then tested for Hardy-Weinberg equilibrium (p-value $< \frac{1}{328348}$.). The final dimensions of the genotype matrix are $m = 324,160$ SNPs and $n = 5027$ individuals.

A Box-Cox transform was applied to each trait, where the parameter was chosen such that the values in the median 95% value of the trait was as close to the normal distribution as possible. Indicators for sex, oral contraception, and fasting status were added as adjustment variables. For glucose, the individual with the minimum value was removed from the analysis as an extreme outlier. All analyses were performed with $d = 6$ logistic factors, which was determined based on the Hardy-Weinberg equilibrium method described in ref. [1]. The association tests were performed exactly as described in the main text.

8 Linear Mixed Effects Model and Principal Component Analysis Approaches

In order to explain the assumptions made by the linear mixed effects model approach (LMM) and principal components approach (PCA), we first re-write model (1) as follows:

$$y_j = \alpha + \beta_i x_{ij} + \sum_{k \neq i} \beta_k x_{kj} + \lambda_j + \epsilon_j,$$

where the object of inference is β_i for each SNP $i = 1, \ldots, m$. As explained in Astle and Balding (2009) [7], these approaches assume that $\lambda_j + \epsilon_j \overset{i.i.d.}{\sim} \text{Normal}(0, \sigma_e^2)$, meaning that the non-genetic effects are independent from population structure and there is no heteroskedasticity among individuals.

The LMM approach also makes the assumption that we can approximate the genetic contribution by a multivariate Normal distribution:

$$\left\{ \sum_{k \neq i} \beta_k x_{kj} \right\}_{j=1}^{n} \dot{\sim} \text{MVN}(\mathbf{0}, \sigma_g^2 \mathbf{\Phi}),$$

where $\mathbf{\Phi}$ is the $n \times n$ kinship matrix. If we define $\eta_j^{(i)} = \sum_{k \neq i} \beta_k x_{kj} + \lambda_j + \epsilon_j$, we can write the above model as

$$y_j = \alpha + \beta_i x_{ij} + \eta_j^{(i)},$$

where it is assumed that $\left\{\eta_j^{(i)}\right\}_{j=1}^{n} \dot{\sim}$ MVN$(\mathbf{0}, \sigma_g^2 \mathbf{\Phi} + \sigma_e^2 \mathbf{I})$. Since it is not the case in general that the $\eta_j^{(i)}$ are identically distributed for all SNPs $i = 1, \ldots, m$, one can either estimate a different pair of parameters (σ_g^2, σ_e^2) for each SNP or assume that these parameters change very little between SNPs. Since the former tends to be computationally demanding, algorithms such as EMMAX [8] propose to estimate a single pair of parameters (σ_g^2, σ_e^2) from a null model and then utilize this single estimate for every SNP. More recently, algorithms such as GEMMA have been proposed to relax this assumption [9].

The $n \times n$ kinship matrix $\mathbf{\Phi}$ is estimated from the genotype data $\mathbf{X}$. This involves the simultaneous estimation of $(n^2 - n)/2$ parameters, which is particularly large for sample sizes considered in current GWAS (on the order of 10^8 for $n = 10,000$). The uncertainty in the estimated $\mathbf{\Phi}$ is typically not taken into account, and there is so far no regularization of the high-dimensional estimator of $\mathbf{\Phi}$. Unregularized estimates of large covariance matrices have been shown to be problematic [10, 11], a concern that is also applicable to estimates of $\mathbf{\Phi}$. Estimating (σ_g^2, σ_e^2) involves manipulations of the estimated $\mathbf{\Phi}$ matrix, which can pose numerical challenges due to the fact that the estimated $\mathbf{\Phi}$ is both high-dimensional and nonsingular. The LMM approach therefore makes assumptions that are important to verify for each given study and it involves some challenging calculations and estimations.

The PCA approach first calculates the top d principal components on a normalized version of the genotype matrix $\mathbf{X}$. In the method proposed by Price et al. (2006) [12], these principal components are then regressed out of each SNP i and the trait (regardless of whether it is binary or quantitative). A correlation statistic is calculated between each adjusted SNP genotype and the adjusted trait, and the p-value that tests for equality to 0 is reported. As shown in Hao et al. (2013) [1], the top d principal components form a high-quality estimate of a linear basis of the allele frequencies π_{ij}. Extracting the residuals after linearly regressing the genotype data for SNP i onto these principal components is equivalent to estimating the quantity $x_{ij} - \pi_{ij}$. Using the trait as the response variable in this regression adjustment is equivalent to estimating $\sum_{k=1}^{n} \beta_k (x_{kj} - \pi_{kj})$ under the assumptions on the trait model given above (where this quantitative trait model is assumed regardless of whether the trait is quantitative or binary). Therefore, the association test carried out in the PCA approach implicitly involves an estimated

form of the model:

$$y_j = \alpha + \beta_i(x_{ij} - \pi_{ij}) + \sum_{k \neq i} \beta_k(x_{kj} - \pi_{ik}) + \lambda_j + \epsilon_j,$$

where it is assumed that $\lambda_j + \epsilon_j$ are approximately i.i.d. Normal$(0, \sigma_e^2)$. When a correlation between the adjusted trait and the adjusted genotype for SNP i is carried out, then the residual variation is based on the joint distribution of $\sum_{k \neq i} \beta_k(x_{kj} - \pi_{ik}) + \lambda_j + \epsilon_j$ for $j = 1, \ldots, n$.

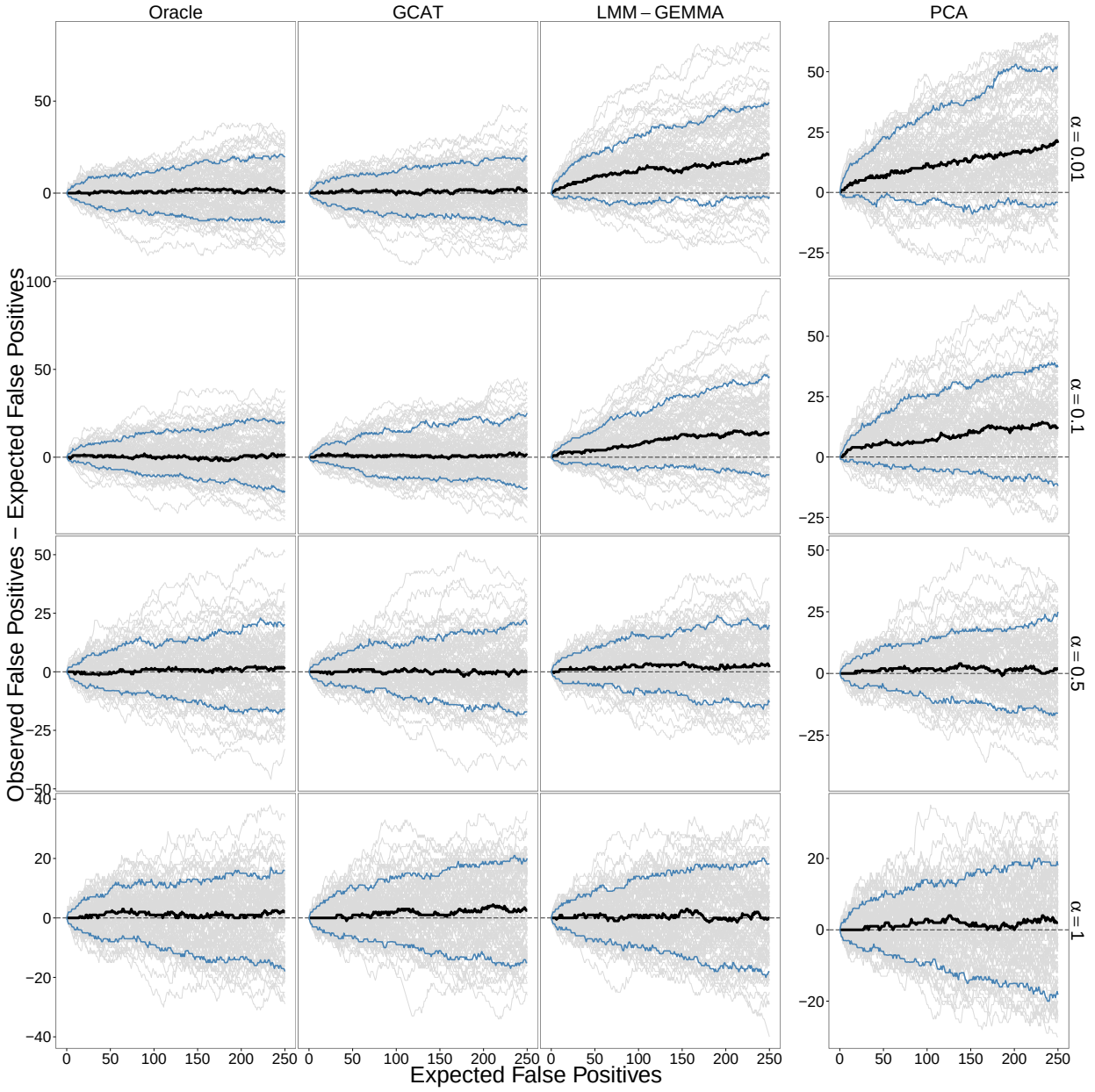
Let us denote $\xi_j^{(i)} = \sum_{k \neq i} \beta_k(x_{kj} - \pi_{ik}) + \lambda_j + \epsilon_j$. Since $\mathrm{Var}(x_{ij} - \pi_{ij}) = 2\pi_{ij}(1 - \pi_{ij})$ and $\mathrm{Var}(x_{kj} - \pi_{kj}) = 2\pi_{kj}(1 - \pi_{kj})$, it follows that $(x_{ij} - \pi_{ij})$ and $(x_{kj} - \pi_{kj})$ for $i, k = 1, \ldots, m$ and $j = 1, \ldots, n$ still suffer from confounding due to structure through their variances. Therefore, the implicit assumption made by the PCA approach that the $\xi_1^{(i)}, \xi_2^{(i)}, \ldots, \xi_n^{(i)}$ are independent and identically distributed in the above model is violated. This is our interpretation of why the PCA approach shows poor performance in adjusting for structure under our quantitative trait simulations. Astle and Balding (2009) [7] make further mathematical characterizations of the relationship between the implicit models in the PCA and LMM approaches, which we also found to be helpful.

Interestingly, when considering the binary trait model (2), the Bernoulli distributed trait does not involve a mean and variance term as in the Normal distributed quantitative trait. It may be the case that this difference contributes to explaining why the PCA approach shows similar behavior to the GCAT and LMM approaches for binary traits (see ref. [7]). Specifically, the PCA approach appears to perform reasonably well in adjusting for structure for the binary trait simulations that we considered.

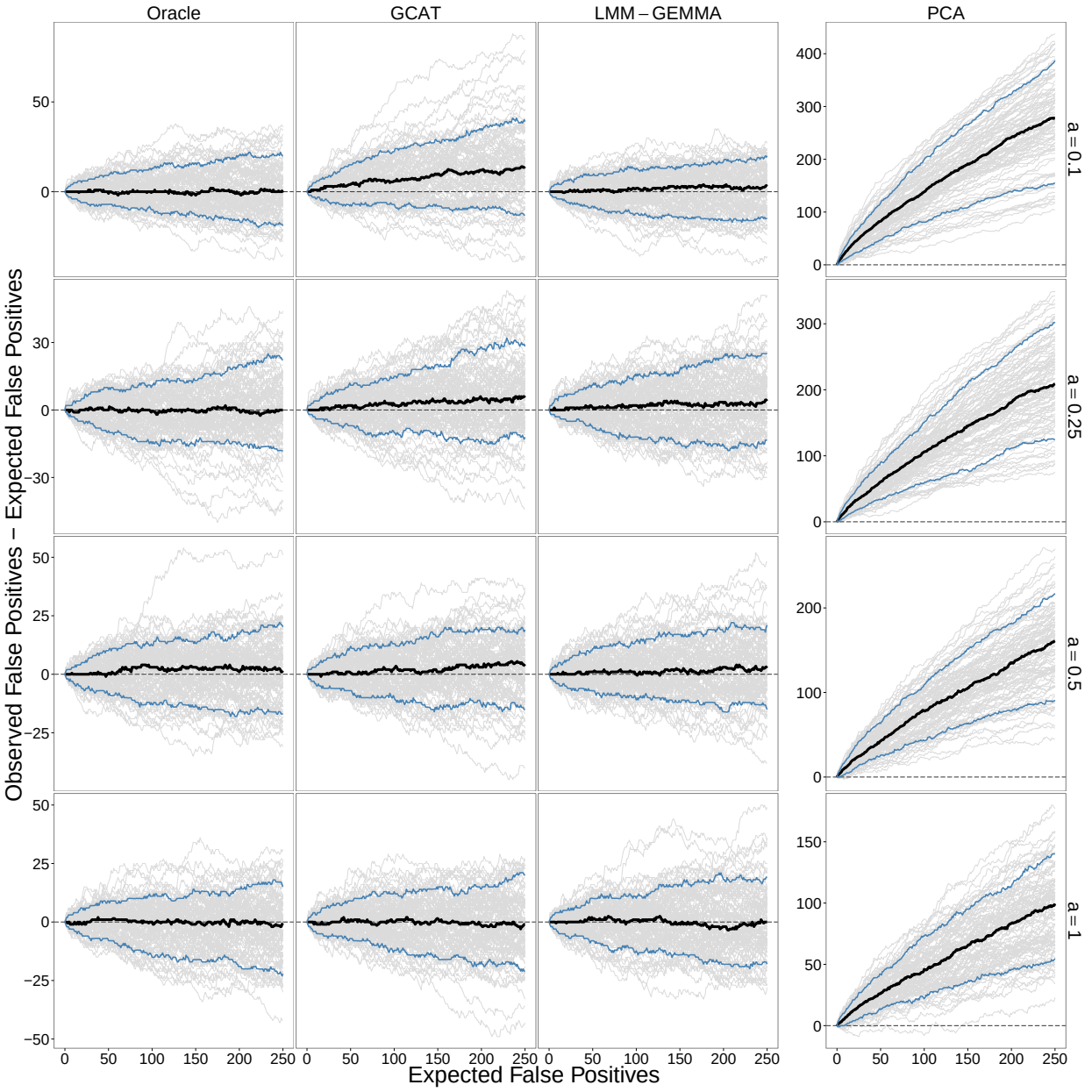
9 Software Implementation

The proposed method has been implemented in open source software, which is available at `https://github.com/StoreyLab/gcat`.

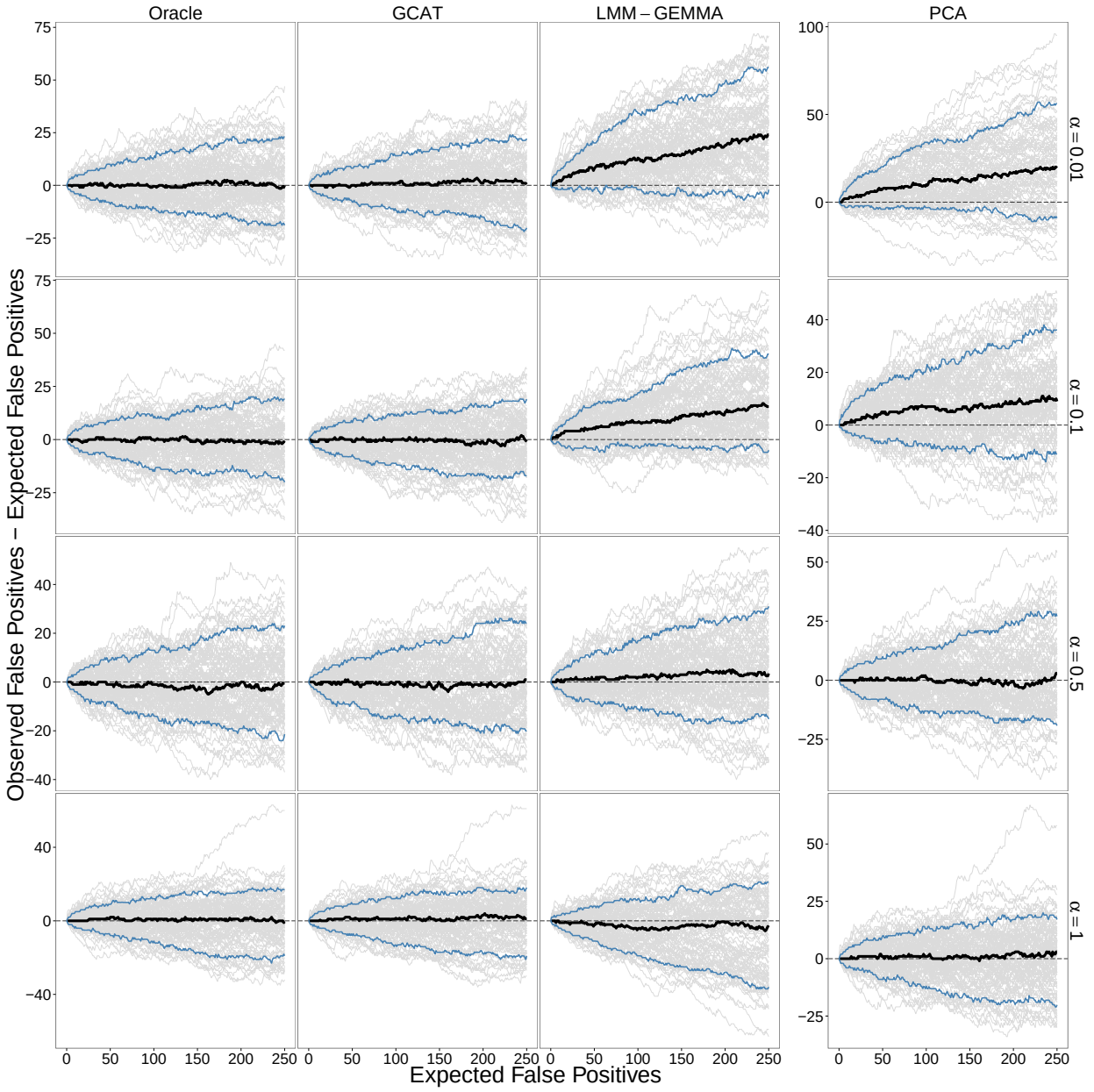
SUPPLEMENTARY FIGURES



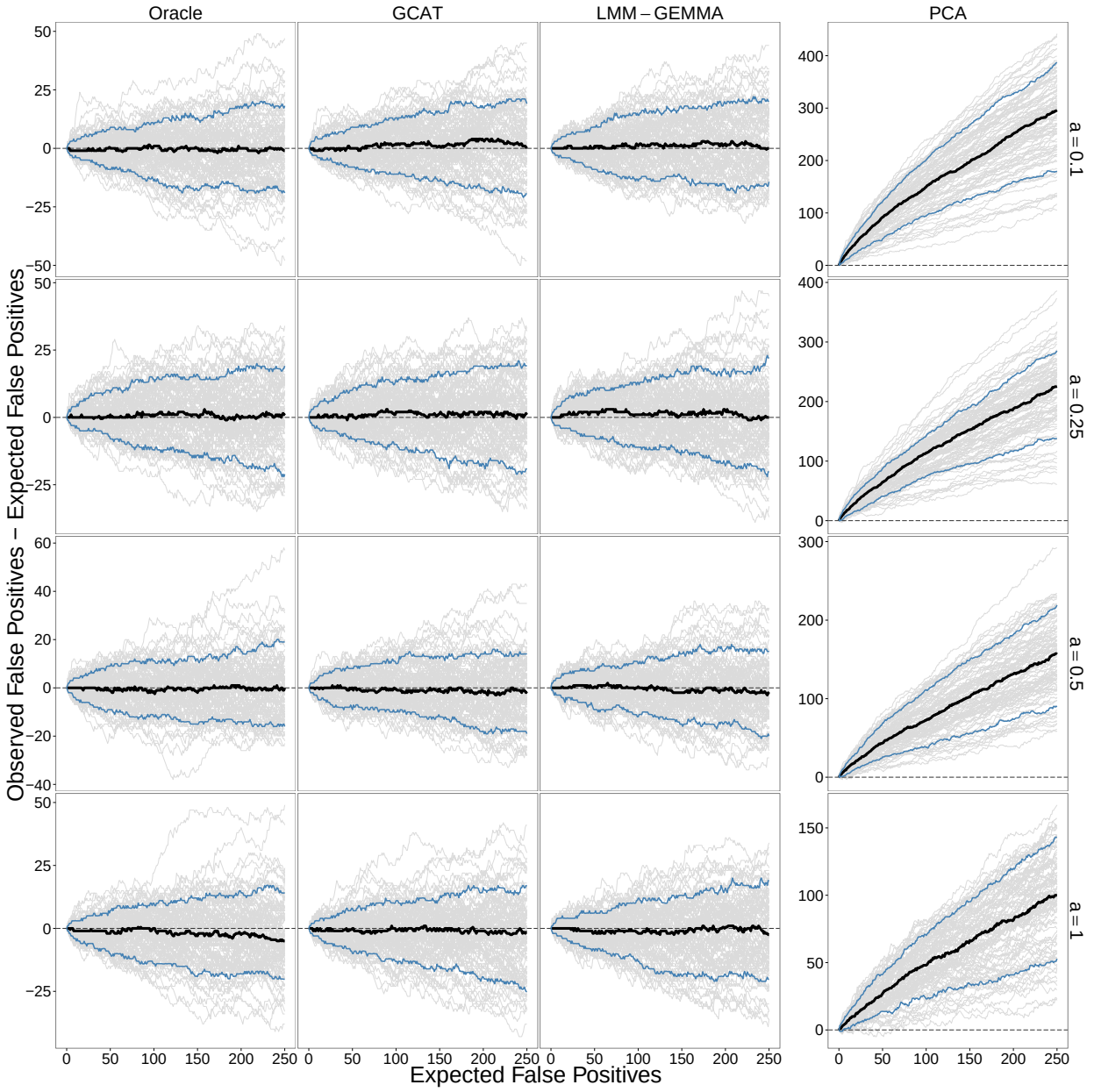
Supplementary Figure 1: Performance of association tests on 100 simulated studies from the PSD model of structure for various α comparing the Oracle, GCAT (proposed), LMM-GEMMA, and PCA tests. The variance contributions to the trait are genetic=5%, environmental=5%, and noise=90%. The remaining details are equivalent to Figure 2.



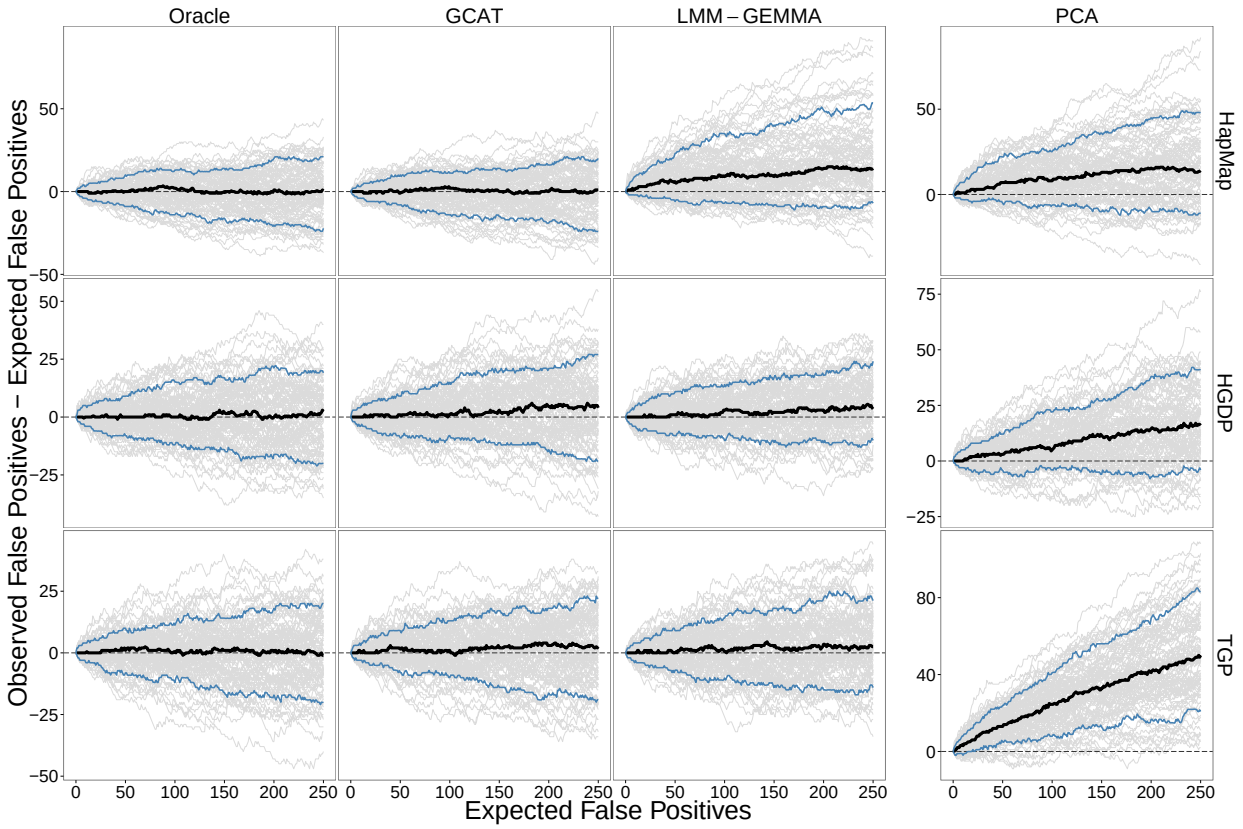
Supplementary Figure 2: Performance of association tests on 100 simulated studies from the spatial model of structure for various a comparing the Oracle, GCAT (proposed), LMM-GEMMA, and PCA tests. The variance contributions to the trait are genetic=5%, environmental=5%, and noise=90%. The remaining details are equivalent to Figure 2.



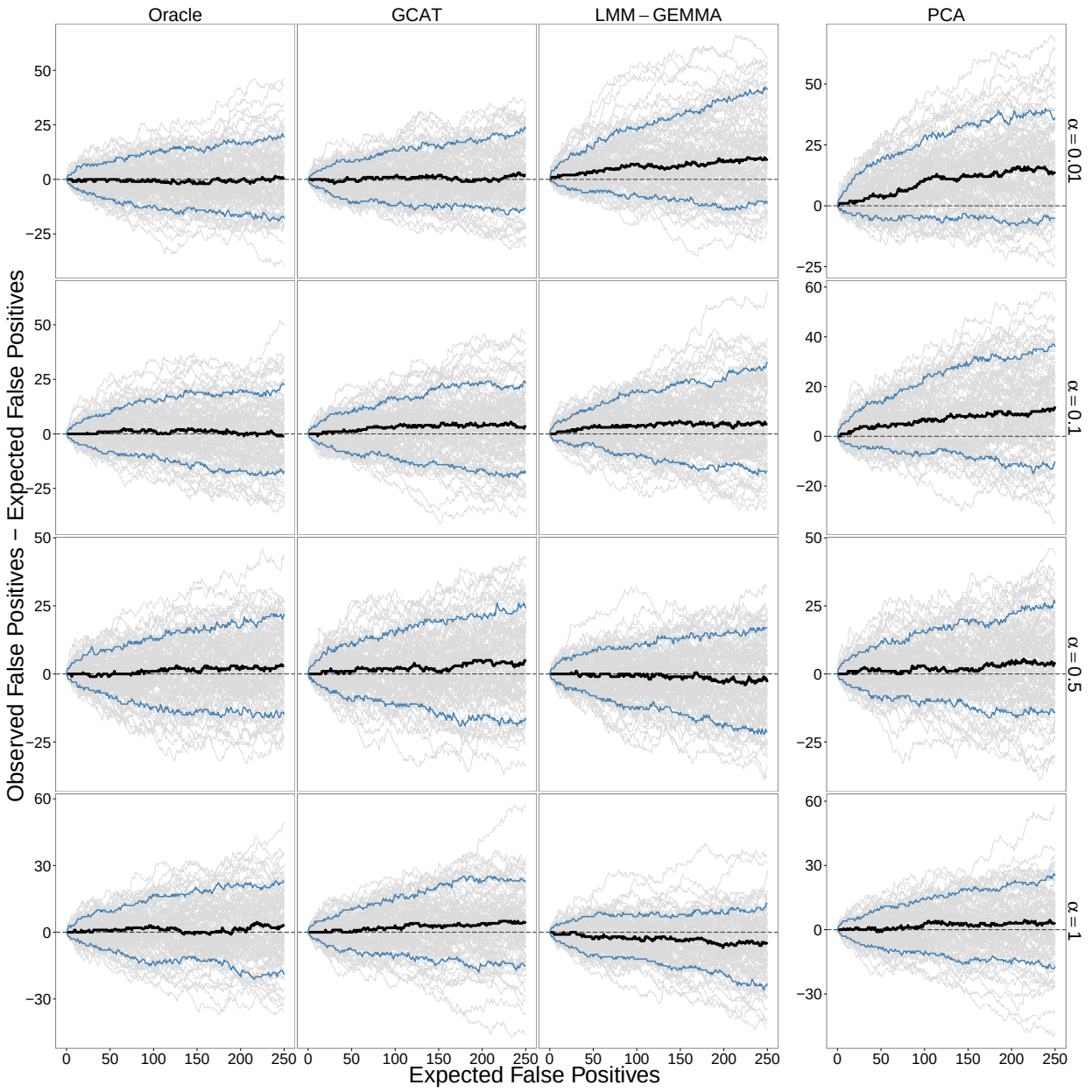
Supplementary Figure 3: Performance of association tests on 100 simulated studies from the PSD model of structure for various α comparing the Oracle, GCAT (proposed), LMM-GEMMA, and PCA tests. The variance contributions to the trait are genetic=10%, environmental=0%, and noise=90%. The remaining details are equivalent to Figure 2.



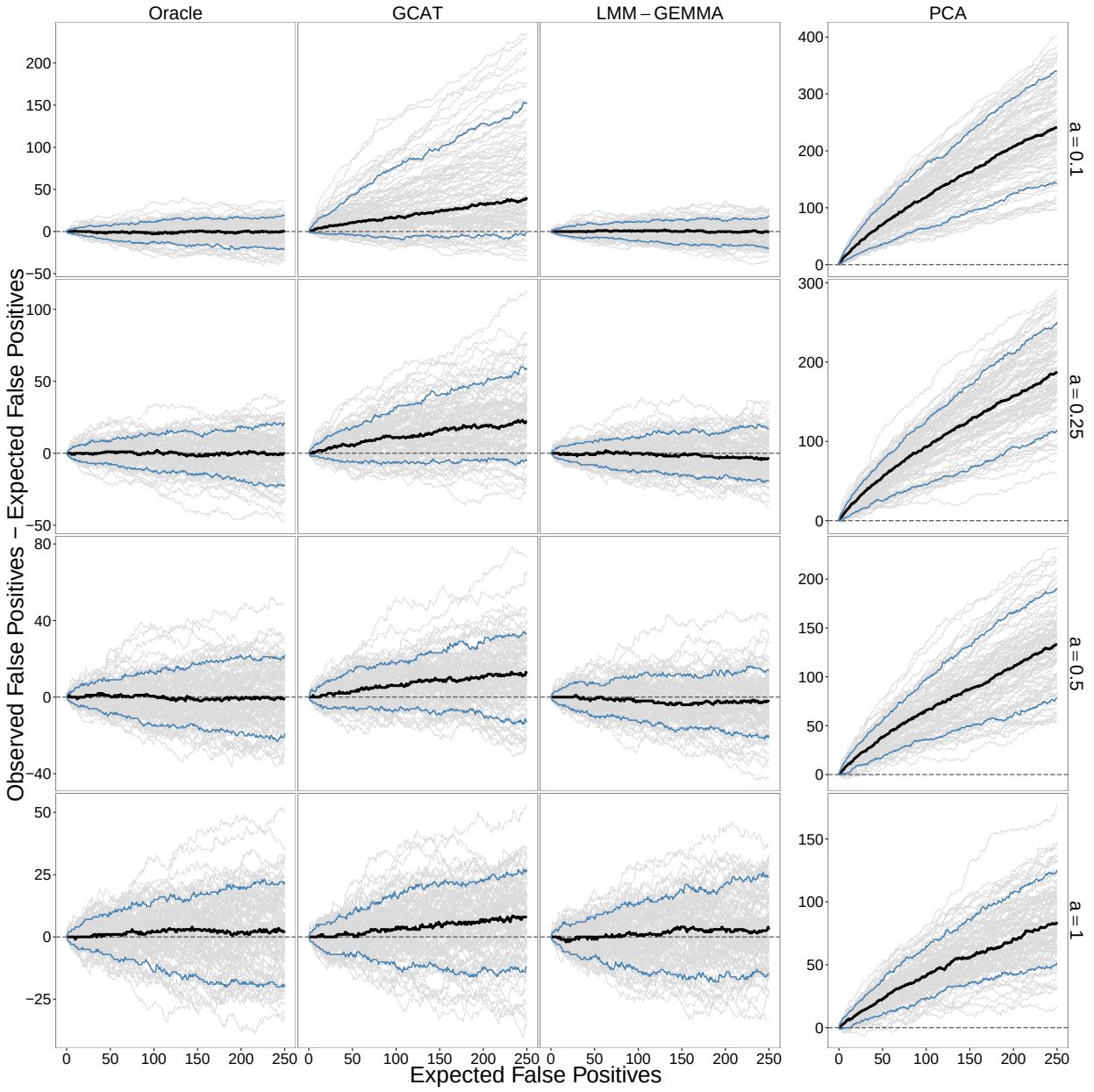
Supplementary Figure 4: Performance of association tests on 100 simulated studies from the spatial model of structure for various a comparing the Oracle, GCAT (proposed), LMM-GEMMA, and PCA tests. The variance contributions to the trait are genetic=10%, environmental=0%, and noise=90%. The remaining details are equivalent to Figure 2.



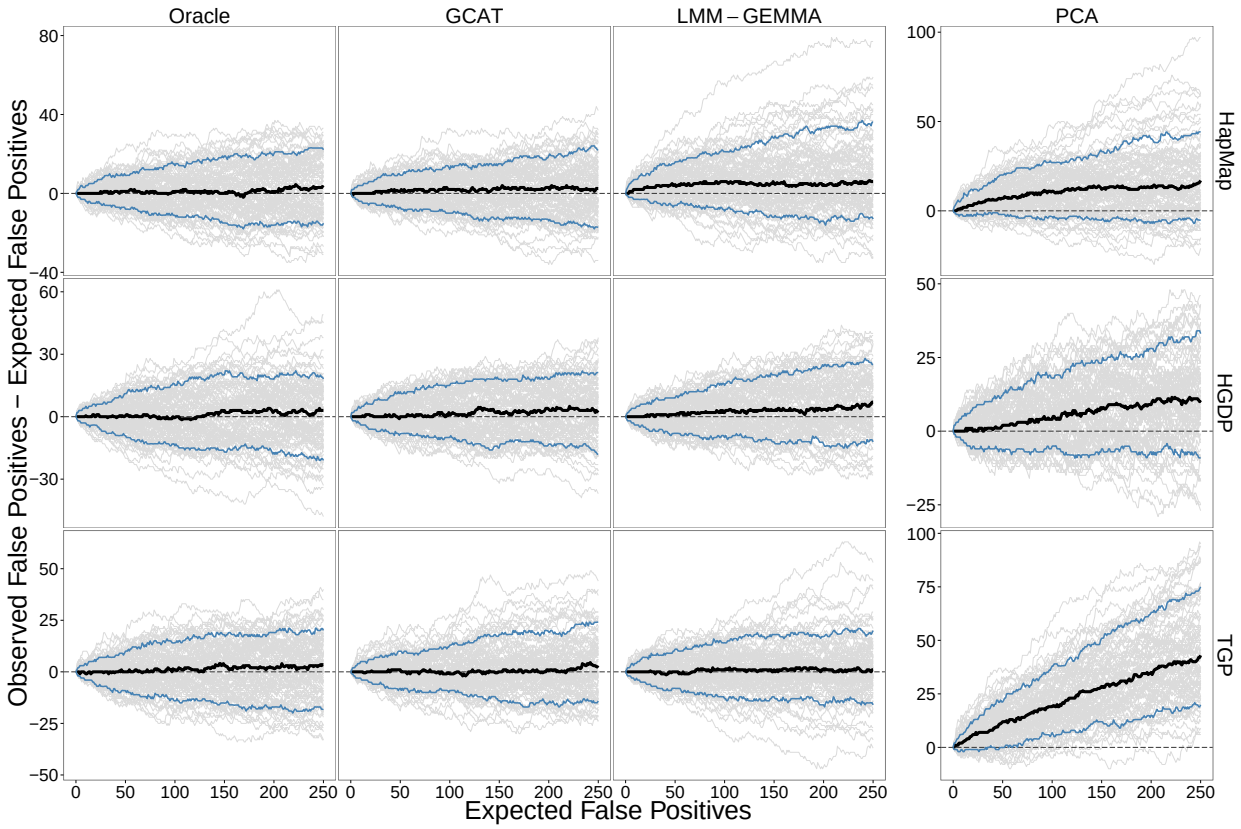
Supplementary Figure 5: Performance of association tests on 100 simulated studies from the Balding-Nichols, HGDP, and TGP simulation scenarios comparing the Oracle, GCAT (proposed), LMM-GEMMA, and PCA tests. The variance contributions to the trait are genetic=10%, environmental=0%, and noise=90%. The remaining details are equivalent to Figure 2.



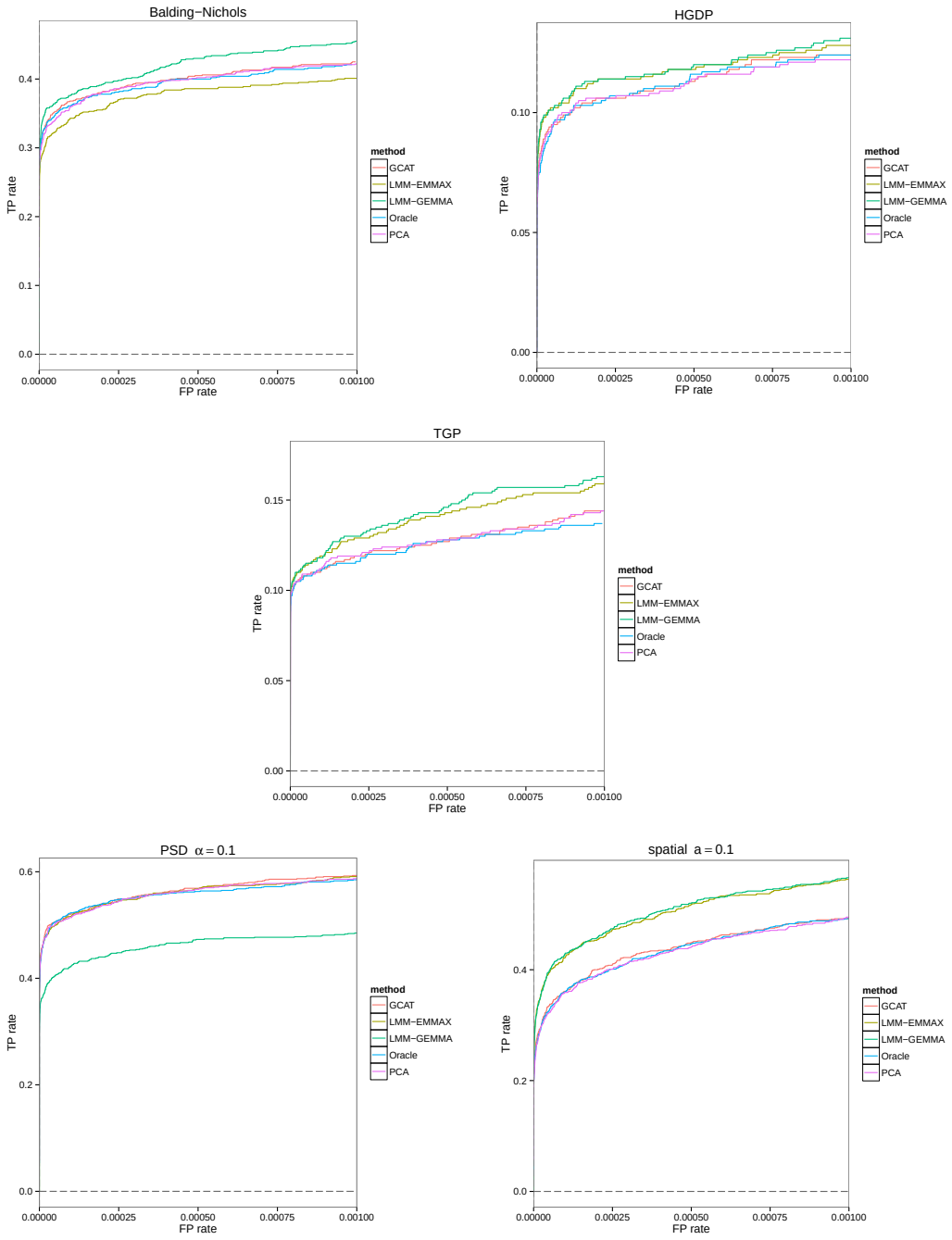
Supplementary Figure 6: Performance of association tests on 100 simulated studies from the PSD model of structure for various α comparing the Oracle, GCAT (proposed), LMM-GEMMA, and PCA tests. The variance contributions to the trait are genetic=20%, environmental=10%, and noise=70%. The remaining details are equivalent to Figure 2.



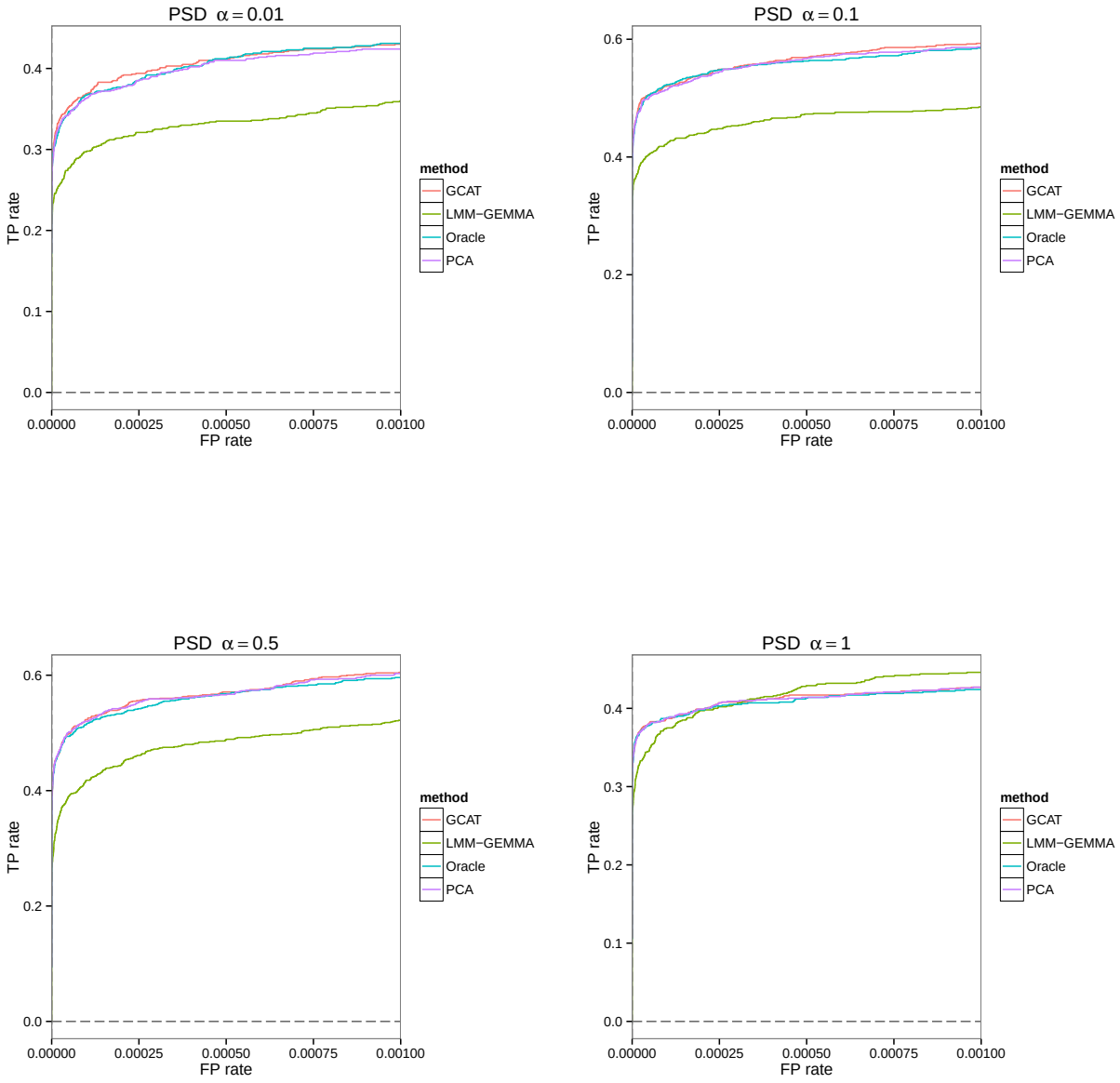
Supplementary Figure 7: Performance of association tests on 100 simulated studies from the spatial model of structure for various a comparing the Oracle, GCAT (proposed), LMM-GEMMA, and PCA tests. The variance contributions to the trait are genetic=20%, environmental=10%, and noise=70%. The remaining details are equivalent to Figure 2. The difference in results between Oracle and GCAT is due to the fact that the π_{ij} values are estimated in GCAT whereas the true π_{ij} values are utilized in Oracle. In this particular simulation scenario, the error in π_{ij} estimation results in a difference.



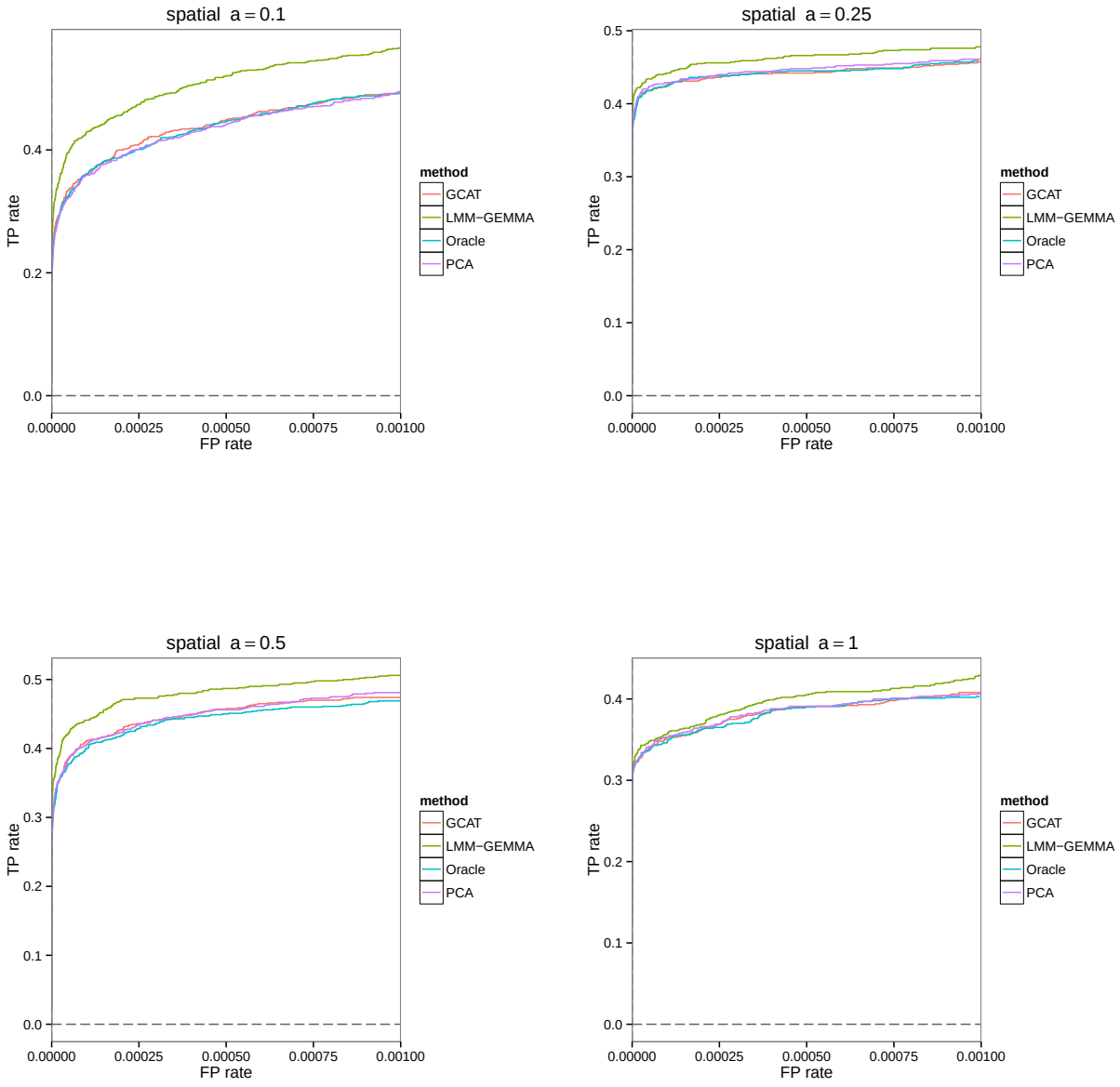
Supplementary Figure 8: Performance of association tests on 100 simulated studies from the Balding-Nichols, HGDP, and TGP simulation scenarios comparing the Oracle, GCAT (proposed), LMM-GEMMA, and PCA tests. The variance contributions to the trait are genetic=20%, environmental=10%, and noise=70%. The remaining details are equivalent to Figure 2.



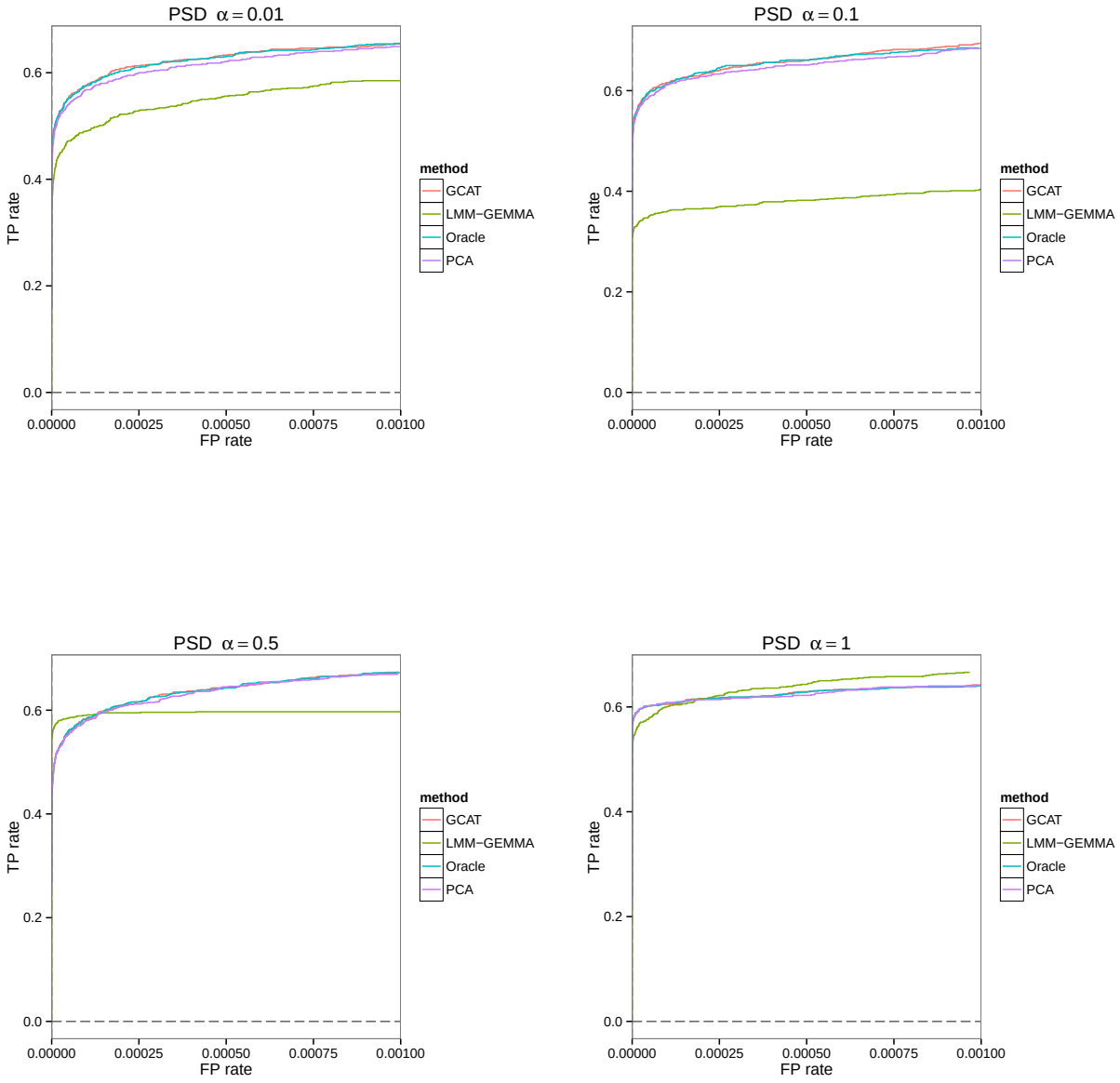
Supplementary Figure 9: Statistical power of the Oracle, GCAT (proposed), PCA, and both LMM association tests. The results are for the simulated data sets shown in Figure 2. The quantitative traits are simulated from model (1) from Online Methods. The variance contributions to the trait are genetic=5%, environmental=5%, and noise=90%.



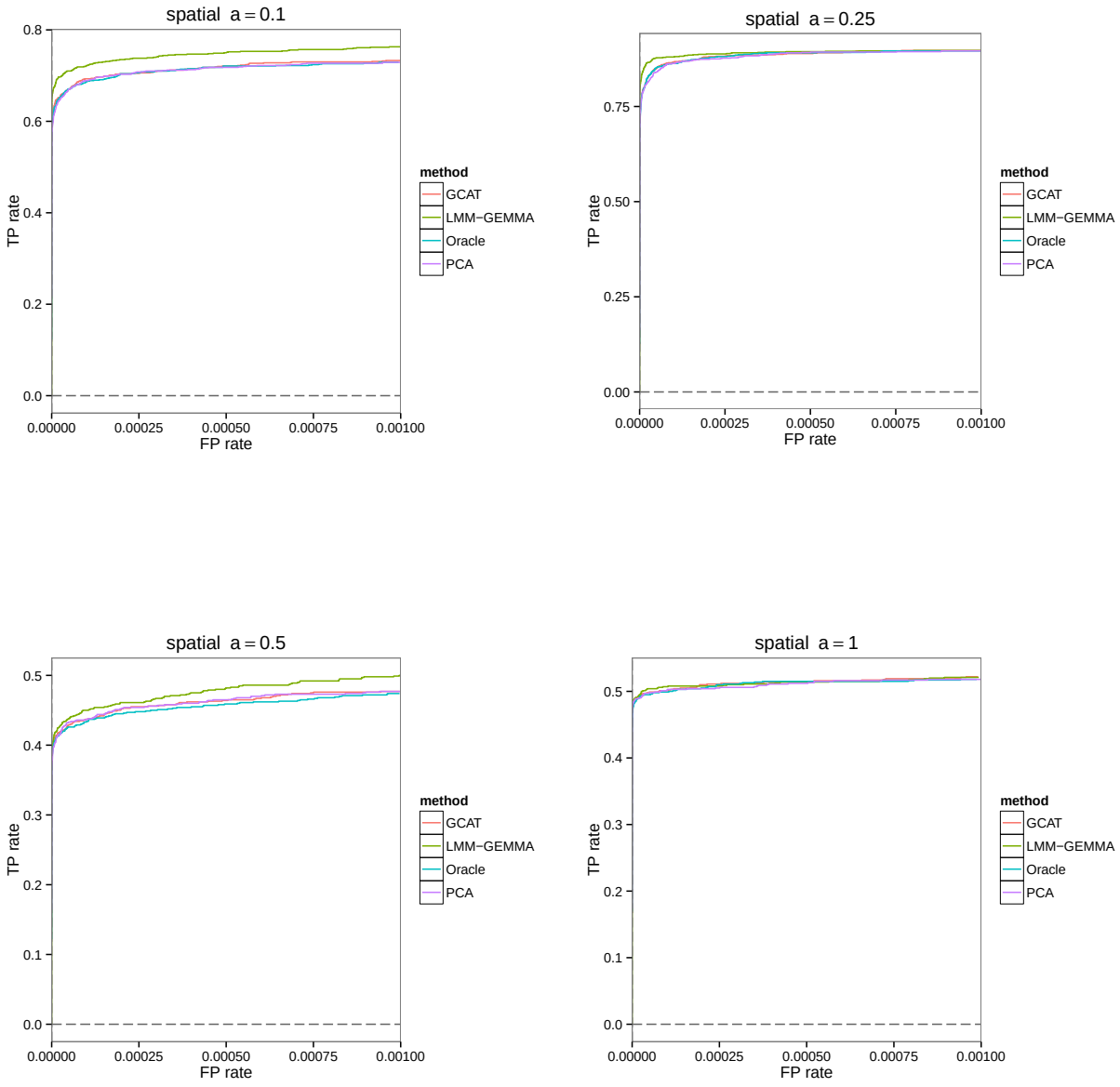
Supplementary Figure 10: Power analysis for the simulation studies presented in Supplementary Figure 1.



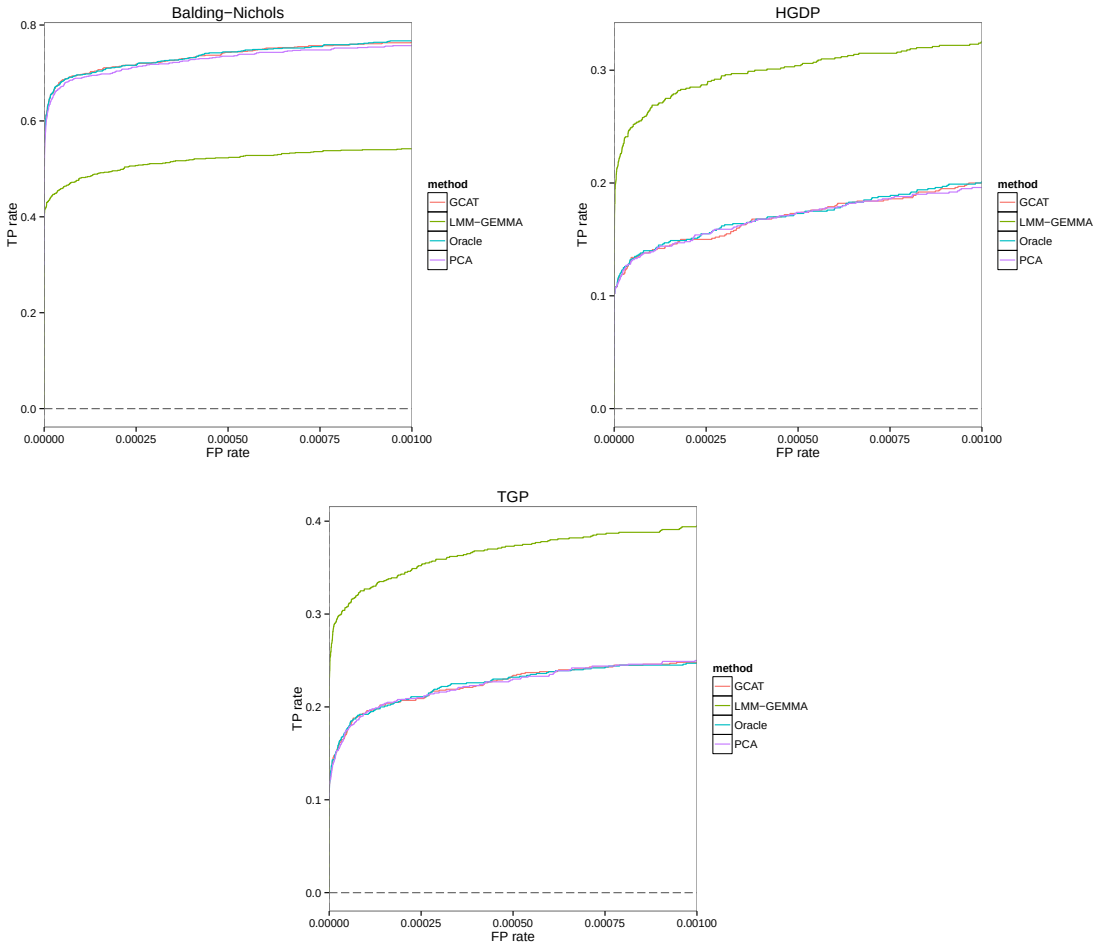
Supplementary Figure 11: Power analysis for the simulation studies presented in Supplementary Figure 2.



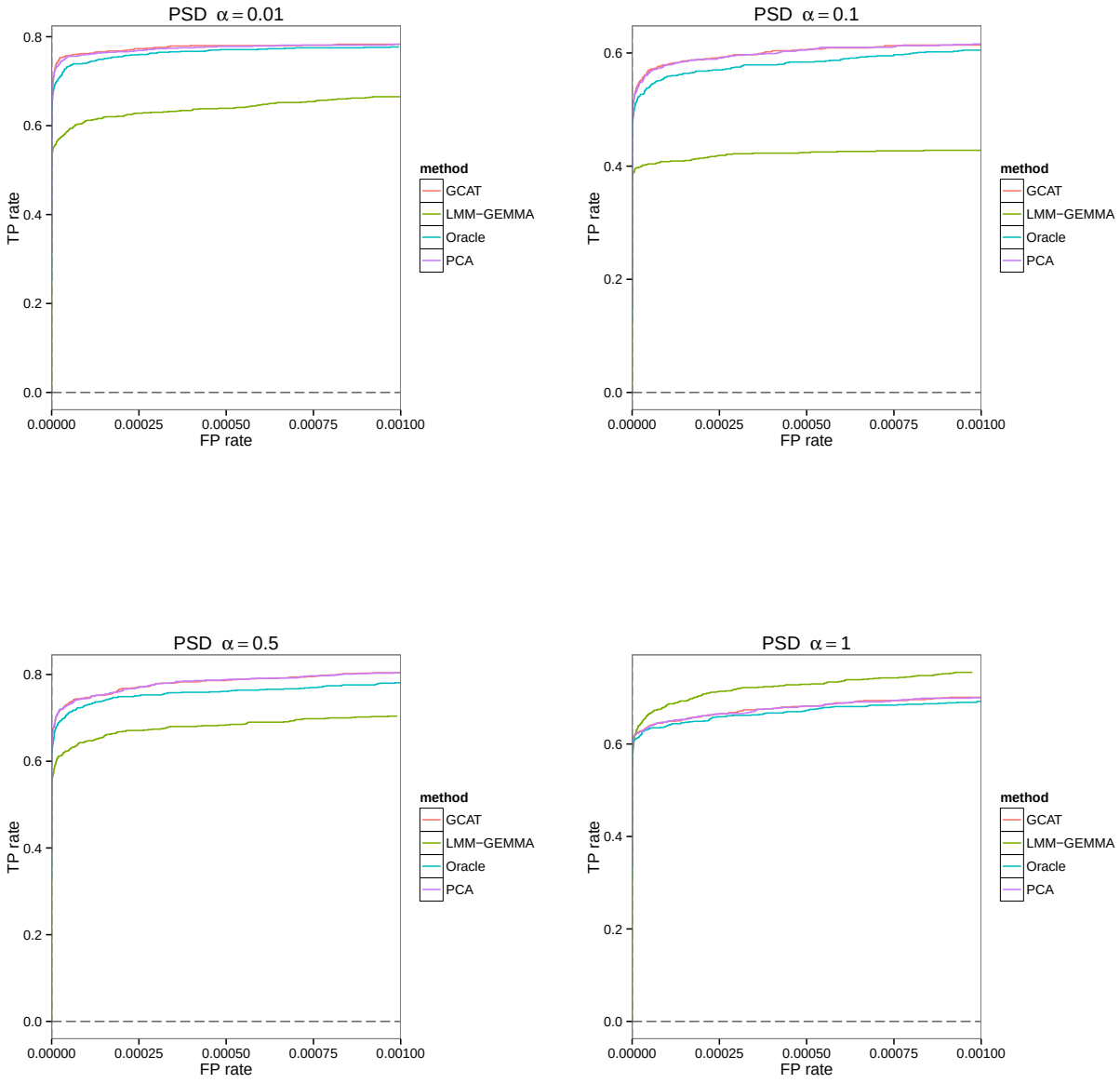
Supplementary Figure 12: Power analysis for the simulation studies presented in Supplementary Figure 3.



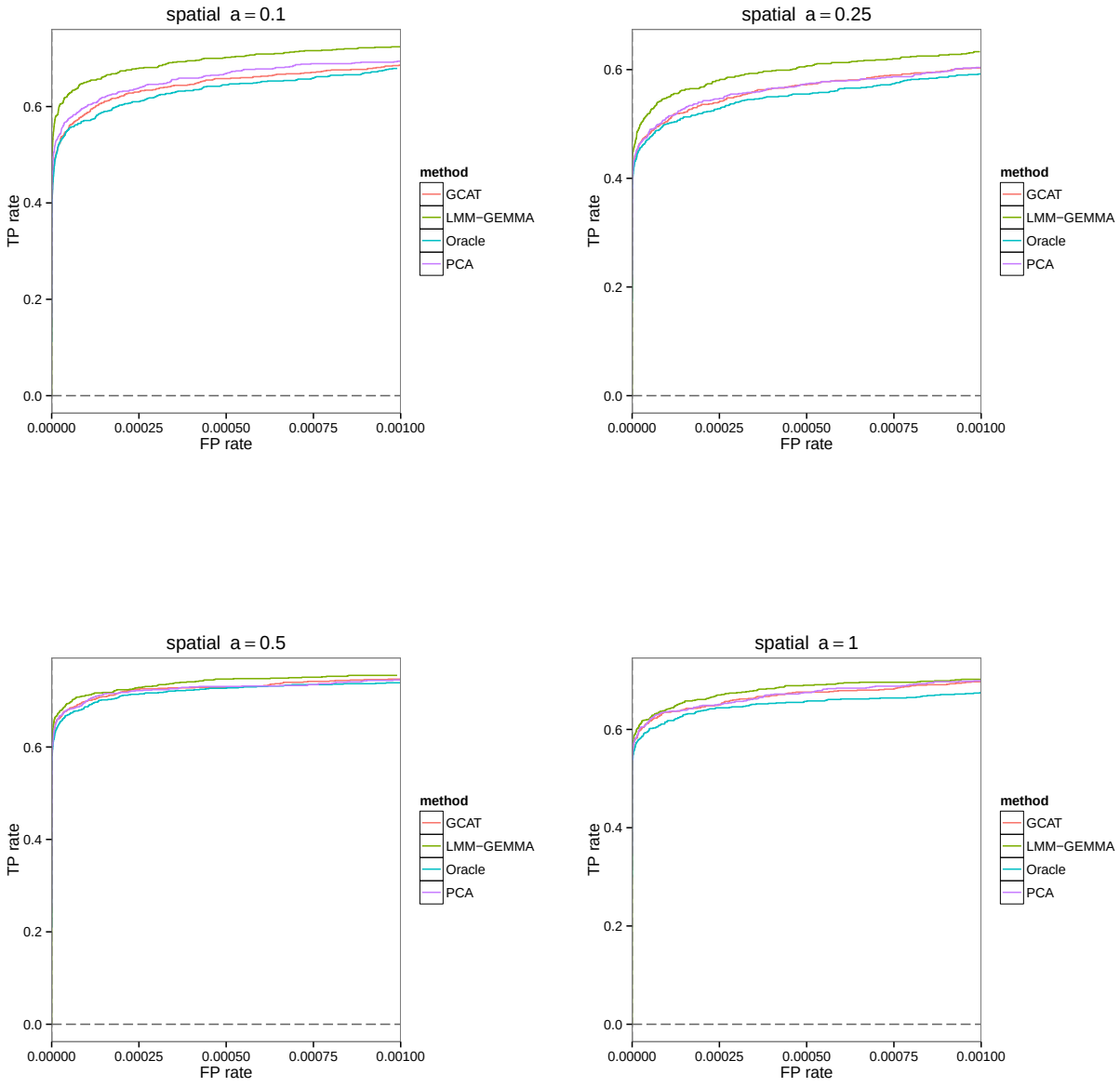
Supplementary Figure 13: Power analysis for the simulation studies presented in Supplementary Figure 4.



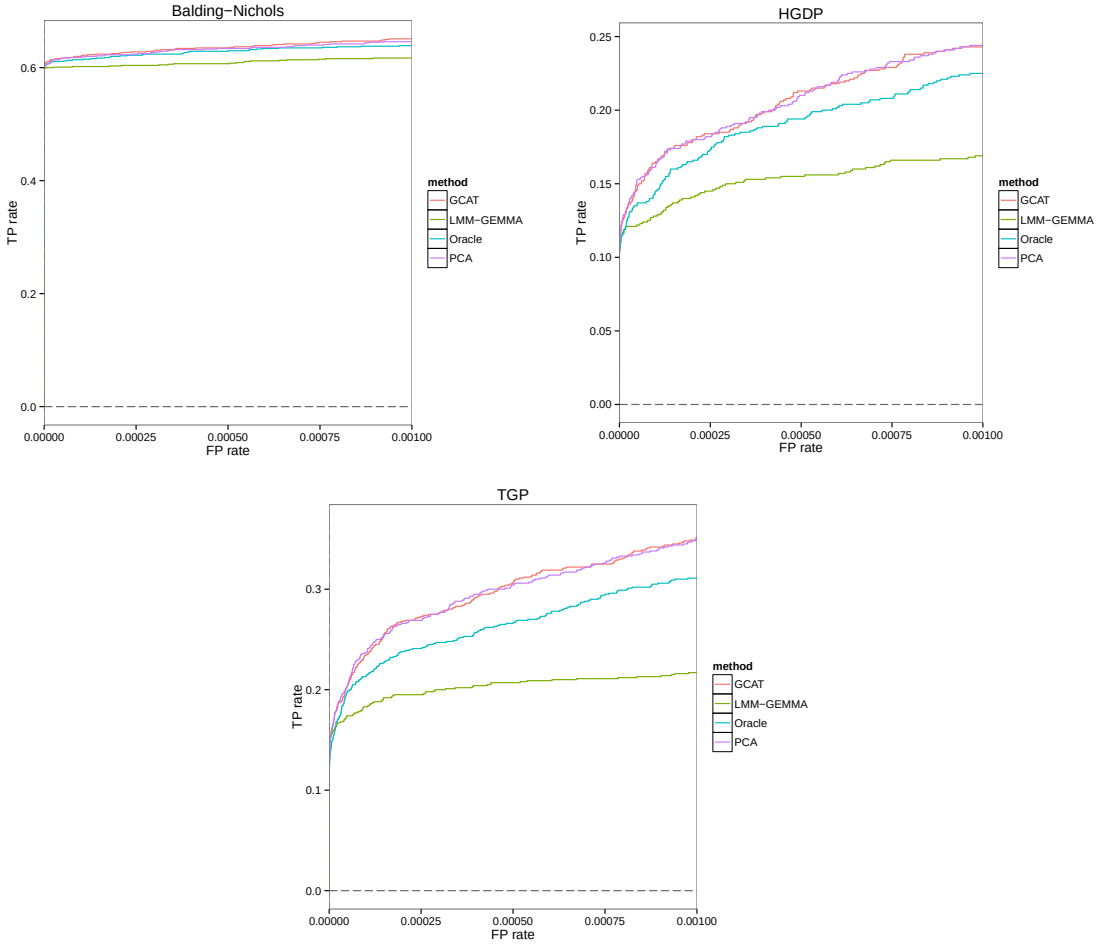
Supplementary Figure 14: Power analysis for the simulation studies presented in Supplementary Figure 5.



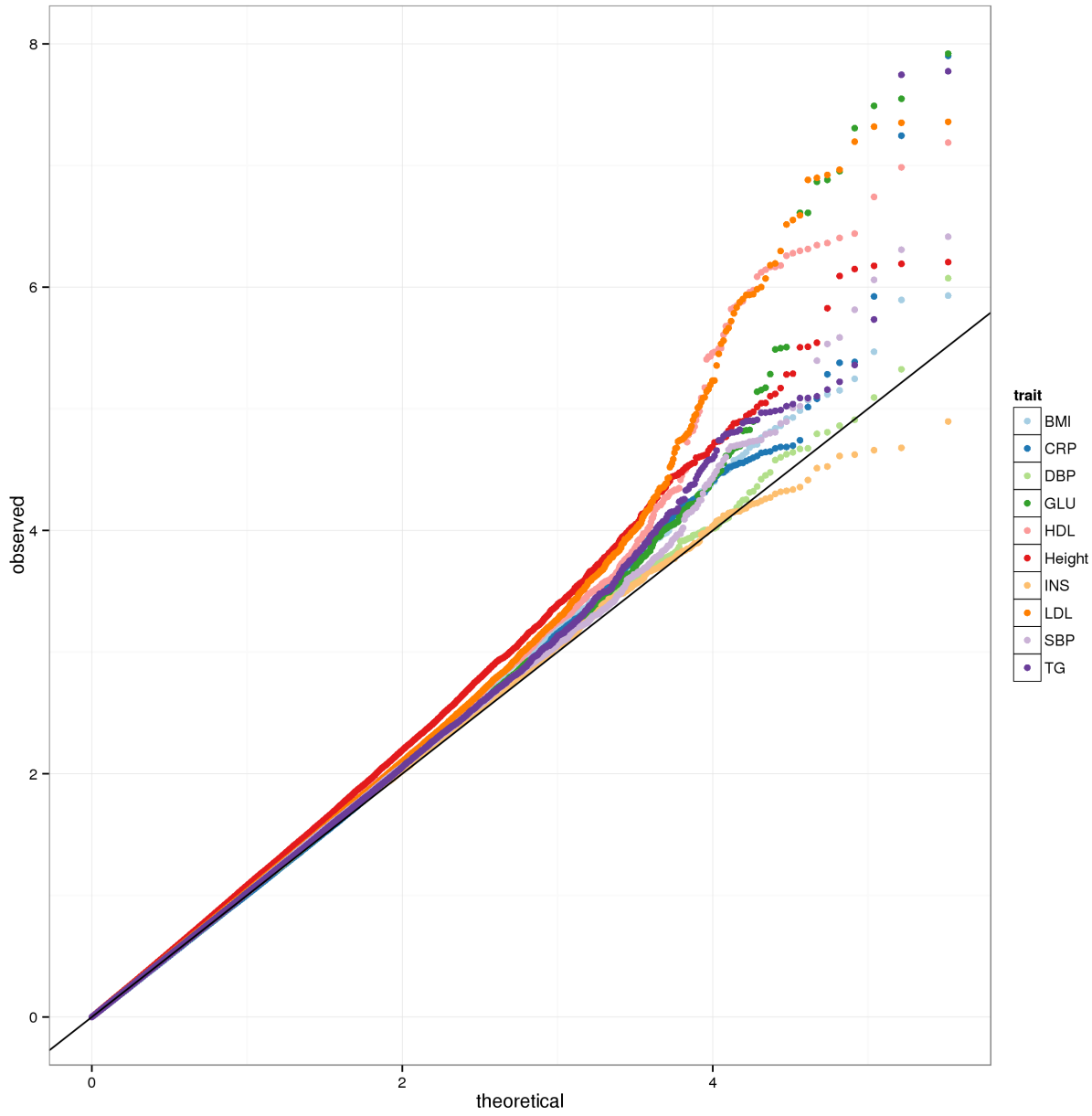
Supplementary Figure 15: Power analysis for the simulation studies presented in Supplementary Figure 6.



Supplementary Figure 16: Power analysis for the simulation studies presented in Supplementary Figure 7.



Supplementary Figure 17: Power analysis for the simulation studies presented in Supplementary Figure 8.



Supplementary Figure 18: Theoretical versus observed quantiles of $-\log_{10}$(p-value) from the GCAT association tests on the Northern Finland Birth Cohort traits. The y-axis was truncated at p-value $< 10^{-8}$; see Supplementary Table 1 for the smallest p-values for each trait.

SUPPLEMENTARY TABLES

(Supplementary Tables begin on next page.)

Supplementary Table 1: The top 20 most associated SNPs for each of the 10 traits considered in the Northern Finland Birth Cohort study. The GCAT p-value and GCAT+GC p-value (genomic control adjusted GCAT p-value) are shown for each SNP. SNPs that achieved GCAT+GC p-value $< 7.2 \times 10^{-8}$ are colored, and each locus for a given trait is given a different color.

BMI

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs987237	6	50911009	1.1740e-06	1.8102e-06
2	rs11759809	6	51063040	1.2745e-06	1.9597e-06
3	rs710139	1	10767145	3.3937e-06	5.0475e-06
4	rs1001729	6	2540477	5.6701e-06	8.2880e-06
5	rs943005	6	50973779	7.0516e-06	1.0231e-05
6	rs6871982	5	56807391	7.6186e-06	1.1025e-05
7	rs12636212	3	86287913	7.9311e-06	1.1462e-05
8	rs8085349	18	55884408	8.5149e-06	1.2276e-05
9	rs4953198	2	45248172	1.0358e-05	1.4834e-05
10	rs7925000	11	8665565	1.1783e-05	1.6803e-05
11	rs6567030	18	54679876	1.2041e-05	1.7157e-05
12	rs8050136	16	52373776	1.3787e-05	1.9556e-05
13	rs1350341	18	55993513	1.4471e-05	2.0492e-05
14	rs12658762	5	18615363	1.5436e-05	2.1811e-05
15	rs633265	18	55982448	1.6156e-05	2.2793e-05
16	rs3751812	16	52375961	1.7325e-05	2.4386e-05
17	rs17207196	7	74939001	1.9619e-05	2.7499e-05
18	rs6447118	4	41550330	1.9832e-05	2.7787e-05
19	rs13386897	2	236764149	2.0884e-05	2.9210e-05
20	rs10484665	6	51050509	2.2783e-05	3.1773e-05

CRP

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs2794520	1	157945440	4.8203e-13	6.1981e-13
2	rs12093699	1	157914612	1.6766e-10	2.0421e-10
3	rs2592887	1	157919563	1.2559e-08	1.4700e-08
4	rs1811472	1	157908973	5.6824e-08	6.5599e-08
5	rs402681	4	104634397	1.1920e-06	1.3383e-06
6	rs7694802	4	104621696	4.1179e-06	4.5715e-06
7	rs2708104	12	119968332	4.1802e-06	4.6401e-06
8	rs7178765	15	23672266	5.2013e-06	5.7622e-06
9	rs340468	4	104637688	8.2712e-06	9.1245e-06
10	rs10774580	12	119960806	9.6851e-06	1.0669e-05
11	rs4259763	10	133291511	1.8144e-05	1.9875e-05
12	rs10107791	8	101040128	2.0076e-05	2.1970e-05
13	rs4534508	10	98272976	2.0584e-05	2.2521e-05
14	rs35779764	10	98309845	2.0584e-05	2.2521e-05
15	rs1510889	12	77295462	2.1194e-05	2.3182e-05
16	rs4656241	1	157880610	2.2403e-05	2.4492e-05
17	rs7538364	1	85711938	2.2729e-05	2.4846e-05
18	rs33964467	10	98310922	2.3057e-05	2.5202e-05
19	rs1403955	1	85712693	2.4406e-05	2.6662e-05
20	rs488797	18	33224625	2.5203e-05	2.7525e-05

DBP

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs472594	1	226668261	8.4419e-07	1.1356e-06
2	rs1491313	4	44480146	4.7333e-06	6.1230e-06
3	rs7783562	7	106704674	8.0721e-06	1.0317e-05
4	rs17305647	21	13962089	1.2297e-05	1.5568e-05
5	rs4548444	1	204956761	1.3747e-05	1.7360e-05
6	rs952061	12	100502356	1.5578e-05	1.9617e-05
7	rs11669309	19	34584137	1.6056e-05	2.0205e-05
8	rs2304586	17	4045747	2.1122e-05	2.6417e-05
9	rs2212853	18	57474627	2.1370e-05	2.6720e-05
10	rs6942973	7	3134277	2.2787e-05	2.8451e-05
11	rs1079199	11	6384682	2.3648e-05	2.9500e-05
12	rs10171678	2	204863117	2.5030e-05	3.1186e-05
13	rs7256832	19	34586645	2.6413e-05	3.2869e-05
14	rs11119265	1	204907336	3.3342e-05	4.1275e-05
15	rs6454393	6	85438647	3.5259e-05	4.3592e-05
16	rs4782509	16	87354279	3.7928e-05	4.6815e-05
17	rs3736338	16	75519348	4.3515e-05	5.3547e-05
18	rs6703170	1	225041893	4.7666e-05	5.8534e-05
19	rs6437523	3	105772154	4.8726e-05	5.9806e-05
20	rs6819019	4	23630880	5.5619e-05	6.8065e-05

GLU

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs560887	2	169471394	3.7754e-12	7.6825e-12
2	rs3847554	11	92308474	9.4364e-10	1.6428e-09
3	rs2971671	7	44177862	4.6022e-09	7.6621e-09
4	rs1387153	11	92313476	6.6178e-09	1.0906e-08
5	rs563694	2	169482317	1.2029e-08	1.9494e-08
6	rs1447352	11	92362409	2.8260e-08	4.4711e-08
7	rs7121092	11	92363999	3.2323e-08	5.0946e-08
8	rs2166706	11	92331180	4.9250e-08	7.6715e-08
9	rs2908290	7	44182662	1.1147e-07	1.6970e-07
10	rs853778	2	169519470	1.3122e-07	1.9886e-07
11	rs10244051	7	15030358	1.3606e-07	2.0598e-07
12	rs2083567	13	110223844	2.4468e-07	3.6441e-07
13	rs2191348	7	15030780	2.4483e-07	3.6462e-07
14	rs2685814	2	169506865	5.2640e-07	7.6737e-07
15	rs12196601	6	65351159	3.1123e-06	4.3181e-06
16	rs763913	14	41907455	3.1755e-06	4.4032e-06
17	rs1893292	18	523191	3.2493e-06	4.5027e-06
18	rs478333	2	169487402	5.1875e-06	7.0959e-06
19	rs497692	2	169497262	6.7055e-06	9.1072e-06
20	rs2073741	22	18369890	7.0091e-06	9.5080e-06

HDL

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs3764261	16	55550825	2.3773e-32	4.9288e-31
2	rs1532624	16	55562980	7.5555e-22	5.5951e-21
3	rs7499892	16	55564091	9.6861e-16	3.9504e-15
4	rs1532085	15	56470658	1.7492e-13	5.7275e-13
5	rs7120118	11	47242866	3.7380e-09	8.0480e-09
6	rs1800961	20	42475778	4.2849e-09	9.1729e-09
7	rs2167079	11	47226831	4.7891e-09	1.0205e-08
8	rs9989419	16	55542640	5.4462e-09	1.1543e-08
9	rs415799	15	56478046	9.2636e-09	1.9202e-08
10	rs255052	16	66582496	6.4830e-08	1.2390e-07
11	rs255049	16	66570972	1.0342e-07	1.9385e-07
12	rs2575875	9	106702315	1.8123e-07	3.3185e-07
13	rs2271293	16	66459571	3.6319e-07	6.4610e-07
14	rs6499137	16	66229305	3.9424e-07	6.9896e-07
15	rs4743764	9	106668925	4.3381e-07	7.6606e-07
16	rs673548	2	21091049	4.5146e-07	7.9591e-07
17	rs1975802	16	66843348	4.8554e-07	8.5340e-07
18	rs8058517	16	66937361	5.0286e-07	8.8257e-07
19	rs6728178	2	21047434	5.2562e-07	9.2082e-07
20	rs676210	2	21085029	5.5106e-07	9.6350e-07

Height

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs2814982	6	34654538	5.7467e-09	2.6103e-08
2	rs2744972	6	34767032	6.2207e-07	1.9050e-06
3	rs2814983	6	34699185	6.4332e-07	1.9645e-06
4	rs2815005	6	34746825	6.6764e-07	2.0325e-06
5	rs2814993	6	34726871	7.0911e-07	2.1478e-06
6	rs2814985	6	34656274	8.1011e-07	2.4266e-06
7	rs2814944	6	34660775	1.4897e-06	4.2405e-06
8	rs6719545	2	218160079	2.8640e-06	7.7191e-06
9	rs4911494	20	33435328	3.0941e-06	8.2857e-06
10	rs6088813	20	33438595	3.1259e-06	8.3639e-06
11	rs9462014	6	34836231	5.1444e-06	1.3205e-05
12	rs1042630	15	87203055	5.2251e-06	1.3394e-05
13	rs2272023	15	87192164	6.7490e-06	1.6936e-05
14	rs8050499	16	66985827	7.5368e-06	1.8740e-05
15	rs2679184	2	232487467	7.8748e-06	1.9509e-05
16	rs6058154	20	33049495	8.9542e-06	2.1947e-05
17	rs6476514	9	36036596	8.9881e-06	2.2024e-05
18	rs4932439	15	87202113	9.6408e-06	2.3486e-05
19	rs13250548	8	35627942	1.0594e-05	2.5606e-05
20	rs9395041	6	44707121	1.0970e-05	2.6439e-05

Supplementary Table 1 continued.

INS

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs7068299	10	72992635	1.2712e-05	1.6927e-05
2	rs7241379	18	64306982	2.0943e-05	2.7508e-05
3	rs6502762	17	3819013	2.1891e-05	2.8719e-05
4	rs11041941	11	1918445	2.3782e-05	3.1129e-05
5	rs885014	10	72997827	2.4419e-05	3.1939e-05
6	rs521184	8	41720842	2.9795e-05	3.8759e-05
7	rs11726701	4	133207690	3.0767e-05	3.9988e-05
8	rs11175040	12	62233961	3.8519e-05	4.9758e-05
9	rs1444858	15	93597363	4.4007e-05	5.6641e-05
10	rs4953198	2	45248172	4.6139e-05	5.9308e-05
11	rs12373385	18	52170174	4.7328e-05	6.0794e-05
12	rs7644598	3	129631215	4.8166e-05	6.1841e-05
13	rs2969344	2	177090835	5.0070e-05	6.4217e-05
14	rs2303164	19	8028737	5.3696e-05	6.8736e-05
15	rs7148454	14	94841177	5.5013e-05	7.0376e-05
16	rs4801020	18	52179034	5.7137e-05	7.3017e-05
17	rs877783	10	72985946	5.9507e-05	7.5962e-05
18	rs932052	12	62081496	6.0274e-05	7.6913e-05
19	rs998223	2	64824633	6.1922e-05	7.8959e-05
20	rs2400541	8	83042101	6.5601e-05	8.3518e-05

LDL

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs646776	1	109620053	3.0987e-11	8.0825e-11
2	rs693	2	21085700	7.3555e-11	1.8507e-10
3	rs754524	2	21165046	3.5409e-09	7.5849e-09
4	rs4844614	1	205941798	4.5687e-09	9.6838e-09
5	rs11668477	19	11056030	9.2904e-09	1.9121e-08
6	rs207150	1	55579053	4.3743e-08	8.4446e-08
7	rs1541596	19	10848013	4.4530e-08	8.5900e-08
8	rs157580	19	50087106	4.7932e-08	9.2182e-08
9	rs3923037	2	21011755	6.3663e-08	1.2101e-07
10	rs6754295	2	21059688	1.0839e-07	2.0156e-07
11	rs754523	2	21165196	1.1943e-07	2.2120e-07
12	rs6728178	2	21047434	1.2624e-07	2.3327e-07
13	rs1429974	2	21154275	1.3113e-07	2.4193e-07
14	rs611917	1	109616775	2.5699e-07	4.6117e-07
15	rs10198175	2	20997364	2.8066e-07	5.0184e-07
16	rs174556	11	61337211	3.0497e-07	5.4344e-07
17	rs3737002	1	205827396	5.0495e-07	8.8133e-07
18	rs207127	1	55588172	6.3836e-07	1.1035e-06
19	rs10495712	2	21049609	6.5972e-07	1.1389e-06
20	rs174546	11	61326406	8.4890e-07	1.4504e-06

SBP

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs782588	2	55695144	3.8489e-07	5.1179e-07
2	rs782586	2	55689669	4.9242e-07	6.5145e-07
3	rs782602	2	55702813	8.7091e-07	1.1387e-06
4	rs2627759	2	55706845	2.5932e-06	3.3154e-06
5	rs2291336	2	55698855	2.9326e-06	3.7399e-06
6	rs1754154	1	43243353	4.0200e-06	5.0935e-06
7	rs10496050	2	55659817	8.3039e-06	1.0366e-05
8	rs1565198	5	8208254	9.5277e-06	1.1860e-05
9	rs782606	2	55740106	9.8327e-06	1.2232e-05
10	rs782652	2	55716279	1.2745e-05	1.5772e-05
11	rs2216322	2	56228414	1.3187e-05	1.6307e-05
12	rs12740489	1	97069523	1.5592e-05	1.9214e-05
13	rs782637	2	55747751	1.6244e-05	2.0002e-05
14	rs12992408	2	55602589	1.7976e-05	2.2089e-05
15	rs7710144	5	92015872	1.8254e-05	2.2423e-05
16	rs2586954	2	55745765	1.8477e-05	2.2691e-05
17	rs480801	11	117018041	1.8653e-05	2.2903e-05
18	rs3741353	11	3085350	1.9268e-05	2.3643e-05
19	rs9791555	7	33211653	1.9431e-05	2.3838e-05
20	rs10486523	7	33208521	1.9832e-05	2.4320e-05

TG

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs1260326	2	27584444	1.7072e-09	3.0005e-09
2	rs10096633	8	19875201	1.6803e-08	2.7606e-08
3	rs780094	2	27594741	1.7955e-08	2.9441e-08
4	rs6447066	4	41102425	1.8445e-06	2.6403e-06
5	rs1260333	2	27602128	4.3671e-06	6.0963e-06
6	rs10499276	6	154351501	6.9589e-06	9.5836e-06
7	rs2083637	8	19909455	7.8991e-06	1.0838e-05
8	rs2304130	19	19650528	9.1440e-06	1.2493e-05
9	rs6447065	4	41101723	1.0246e-05	1.3952e-05
10	rs2190174	7	78817283	1.0389e-05	1.4142e-05
11	rs2907632	17	50223911	1.0624e-05	1.4453e-05
12	rs261336	15	56529710	1.0734e-05	1.4598e-05
13	rs673548	2	21091049	1.0768e-05	1.4642e-05
14	rs676210	2	21085029	1.2314e-05	1.6679e-05
15	rs12179536	6	31101569	1.2519e-05	1.6950e-05
16	rs10060710	5	156213134	1.2655e-05	1.7128e-05
17	rs2364913	7	78861440	1.3088e-05	1.7697e-05
18	rs28397289	6	31305386	1.4835e-05	1.9986e-05
19	rs2075650	19	50087459	1.5417e-05	2.0747e-05
20	rs6728178	2	21047434	1.5585e-05	2.0967e-05

CRP (untransformed)

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs2464196	12	119919810	1.6254e-09	2.3469e-09
2	rs1169300	12	119915608	1.9049e-09	2.7420e-09
3	rs2794520	1	157945440	2.9924e-08	4.0861e-08
4	rs2650000	12	119873345	2.7614e-07	3.6141e-07
5	rs735396	12	119923227	3.3146e-07	4.3231e-07
6	rs2592887	1	157919563	3.4052e-07	4.4390e-07
7	rs10160939	12	128430312	8.3779e-07	1.0736e-06
8	rs2009800	17	72026460	3.0208e-06	3.7779e-06
9	rs10035541	5	7592712	6.2592e-06	7.7210e-06
10	rs2098930	3	153371624	6.8292e-06	8.4103e-06
11	rs7953249	12	119888107	6.8385e-06	8.4215e-06
12	rs390623	9	118028734	7.4386e-06	9.1459e-06
13	rs924796	11	11067701	1.1358e-05	1.3854e-05
14	rs12093699	1	157914612	1.4562e-05	1.7679e-05
15	rs2072081	17	39683019	1.9668e-05	2.3743e-05
16	rs8015588	14	55230657	1.9845e-05	2.3953e-05
17	rs10483644	14	55171632	2.6387e-05	3.1679e-05
18	rs1811472	1	157908973	2.6710e-05	3.2059e-05
19	rs7637998	3	54061623	2.7532e-05	3.3027e-05
20	rs1169302	12	119916685	3.5850e-05	4.2793e-05

TG (untransformed)

	RSID	Chr	Pos	GCAT	GCAT+GC
1	rs1260326	2	27584444	4.8574e-09	5.6817e-09
2	rs10096633	8	19875201	9.7234e-09	1.1305e-08
3	rs780094	2	27594741	3.0158e-08	3.4722e-08
4	rs673548	2	21091049	7.1013e-06	7.8005e-06
5	rs3923037	2	21011755	7.8905e-06	8.6596e-06
6	rs6581439	12	38608113	9.4246e-06	1.0328e-05
7	rs676210	2	21085029	9.8160e-06	1.0753e-05
8	rs784622	1	39877401	1.1086e-05	1.2132e-05
9	rs6122161	20	61857331	1.2809e-05	1.4000e-05
10	rs261336	15	56529710	1.3401e-05	1.4641e-05
11	rs1836882	11	88871809	1.7570e-05	1.9151e-05
12	rs2286276	7	72625290	1.7695e-05	1.9287e-05
13	rs12179536	6	31101569	1.9373e-05	2.1100e-05
14	rs6728178	2	21047434	1.9395e-05	2.1123e-05
15	rs3811644	2	27656309	2.5212e-05	2.7397e-05
16	rs12805061	11	116058235	2.5845e-05	2.8079e-05
17	rs6472088	8	64381899	2.6590e-05	2.8881e-05
18	rs10234070	7	44504221	2.7333e-05	2.9681e-05
19	rs7700248	4	89073818	2.8885e-05	3.1352e-05
20	rs6843164	4	95838010	2.9945e-05	3.2492e-05

Supplementary Table 2: The genomic control inflation factor (GCIF) was calculated for each trait in the association analysis of the Northern Finland Birth Cohort traits. The calculation was based on SNPs spaced at $\sim$250kbp. The 95% Bonferroni adjusted simultaneous confidence interval under the assumption that the median statistic follows the theoretical null distribution is $(0.9389, 1.0666)$. We calculated GCIF for the proposed statistics $T(\mathbf{x}_i, \mathbf{y}, \widehat{\mathbf{H}})$ and $T(\mathbf{x}_i, \mathbf{y}, \widehat{\boldsymbol{\pi}}_i)$ defined in the text.

Trait	Abbreviation	$T(\mathbf{x}_i, \mathbf{y}, \widehat{\mathbf{H}})$	$T(\mathbf{x}_i, \mathbf{y}, \widehat{\boldsymbol{\pi}}_i)$
Body Mass Index	BMI	1.0633	1.0445
C-reactive Protein	CRP	1.0073	1.0050
Diastolic blood pressure	DBP	1.0487	1.0306
Glucose	GLU	1.0225	0.9886
HDL Cholesterol	HDL	1.0418	1.0206
Height	Height	1.0798	1.1017
Insulin	INS	1.0471	1.0636
LDL Cholesterol	LDL	1.0651	1.0264
Systolic blood pressure	SBP	1.0319	1.0336
Triglycerides	TG	1.0708	1.0327

References

[1] Hao, W., Song, M., and Storey, J. D. Probabilistic models of genetic variation in structured populations applied to global human studies. arXiv:1312.2041 (2013).

[2] Balding, D. J. and Nichols, R. A. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**(1-2), 3–12 (1995).

[3] Pritchard, J. K., Stephens, M., and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**(2), 945–959, Jun (2000).

[4] Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* **9**(1), 60–62 (1938).

[5] Weir, B. and Cockerham, C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).

[6] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).

[7] Astle, W. and Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Statistical Science* **24**, 451–471 (2009).

[8] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**(4), 348–354 (2010).

[9] Zhou, X. and Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**(7), 821–824, Jul (2012).

[10] Bickel, P. J. and Levina, E. Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**(1), 199–227, 02 (2008).

[11] Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3), 432–441 (2008).

[12] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**(8), 904–909, Aug (2006).