bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Testing for genetic associations in arbitrarily structured populations

Minsun Song[1*+], Wei Hao[1*], and John D. Storey[1,2 †]

1. Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544

2. Department of Molecular Biology, Princeton University, Princeton, NJ 08544

[*] These authors contributed equally to this work

[+] Present address: Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD 20850

[†] To whom correspondence should be addressed: jstorey@princeton.edu

Contents

Introduction 1

Results 3

Discussion 11

Methods 13

References 26

Figures and Tables 31

Supplementary Figures and Tables 34

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We present a new statistical test of association between a trait (either quantitative or binary) and genetic markers, which we theoretically and practically prove to be robust to arbitrarily complex population structure. The statistical test involves a set of parameters that can be directly estimated from large-scale genotyping data, such as that measured in genome-wide associations studies (GWAS). We also derive a new set of methodologies, called a genotype-conditional association test (GCAT), shown to provide accurate association tests in populations with complex structures, manifested in both the genetic and environmental contributions to the trait. We demonstrate the proposed method on a large simulation study and on the Northern Finland Birth Cohort study. In the Finland study, we identify several new significant loci that other methods do not detect. Our proposed framework provides a substantially different approach to the problem from existing methods. We provide some discussion on its similarities and differences with the linear mixed model and principal component approaches.

INTRODUCTION

Performing genome-wide tests of association between a trait and genetic markers is one of the most important research efforts in modern genetics [1–3]. However, a major problem to overcome is how to test for associations in the presence of population structure [4]. Human populations are often structured in the sense that the genotype frequencies at a particular locus are not homogeneous throughout the population. Rather, there are latent variables (such as geography or ancestry) that directly affect the allele frequencies of the genotypes. At the same time, there may be other loci and non-genetic factors that also correlate with these latent variables, which in turn are correlated with the trait of interest. When this occurs, genetic markers become spuriously statistically associated with the trait of interest despite the fact that there is no biological connection.

The importance of addressing association testing in structured populations is evidenced by the existence of a large literature of methods proposed for this problem [5, 6]. The well established methods all take a similar strategy in that the trait is modeled in terms of the genetic markers of interest, while attempting to adjust for genetic structure. Two popular approaches are to correct population structure by including principal components of genotypes as adjust-

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ment variables [7, 8] or by fitting a linear mixed effects model involving an estimated kinship or covariance matrix from the individuals' genotypes [9, 10]. Previous work investigating the limitations of these two methods include Wang, et al. (2013) [11]. These two approaches have been shown to be based on a common model that make differing assumptions about how the kinship or covariance matrices are utilized in the model [5]. This common model does not allow for non-genetic (e.g., environmental) contributions to the trait to be dependent with population structure. The linear mixed effects model requires that the genetic component is composed of small effects that additively are well-approximated by the Normal distribution. The model itself is therefore an approximation, and it is not yet possible to theoretically prove that a test based on this model is robust to structure for the more general class of relevant models that we investigate.

By taking a substantially different approach that essentially reverses the placement of the trait and genotype in the model, we formulate and provide a theoretical solution to the problem of association testing in structured populations for both quantitative and binary traits under general assumptions about the complexity of the population structure and its relationship to the trait through both genetic and non-genetic factors. This theoretical solution directly leads to a method for addressing the problem in practice that differs in key ways from the mixed model and principal component approaches. The method is straightforward: a model of structure is first estimated from the genotypes, and then a logistic regression is performed where the SNP genotypes are logistically regressed on the trait plus an adjustment based on the fitted structure model. The coefficient corresponding to the trait is then tested for statistical significance. The class of models to which this provides a test robust to structure is fairly general.

This association testing framework is robust to population genetic structure, as well as to non-genetic effects that are dependent or correlated with population genetic structure (for example, lifestyle and environment may be correlated with ancestry) and with heteroskedasticity that is dependent on structure. We introduce a test based on this framework, called "genotype conditional association test" (GCAT). We show the proposed method corrects for structure on simulated data with a quantitative trait and compares favorably to existing methods. We also apply the method to the Northern Finland Birth Cohort data [12] and identify several new associated loci that have not been identified by existing methods. For example, the proposed method is the only one to identify a SNP (rs2814982) associated with height, which we note is linked to another SNP (rs2814993) that has been associated with skeletal frame size [13]. We discuss the advantages and disadvantages of the proposed framework with existing ap-

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

proaches, and we conclude that the proposed framework will be useful in future studies as sample sizes and the complexity of structure increase.

RESULTS

Population Structure Model

Suppose that there are n individuals, each with m measured SNP genotypes. The genotype for SNP i in individual j is denoted by $x_{ij} \in \{0, 1, 2\}$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. We collected these SNP genotypes into an $m \times n$ matrix $\mathbf{X}$, where the (i, j) entry is x_{ij}. We denote the genotypes for individual j by $\mathbf{x}^j = (x_{1j}, x_{2j}, \ldots, x_{mj})^T$.

We utilize our recently developed framework that flexibly models complex population structures for diallelic loci [14]. Let $\boldsymbol{Z}$ be an unobserved variable describing how individuals fit into the underlying population structure. For a SNP i, the allele frequency π_i can be viewed as being a function of $\boldsymbol{Z}$, $\pi_i(\boldsymbol{Z})$. For a random sample of n individuals from an overall population, we therefore have sampled population structure positions $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$ with resulting allele frequencies $\pi_i(\boldsymbol{z}_1), \pi_i(\boldsymbol{z}_2), \ldots, \pi_i(\boldsymbol{z}_n)$ for SNP i. In Hao et al. (2013) [14], we formulate and estimate a model for m SNPs simultaneously while providing a flexible parameterization of the form of $\pi_i(\boldsymbol{Z})$.

For shorthand, $\pi_{ij} \equiv \pi_i(\boldsymbol{z}_j)$ is the allele frequency for SNP i conditioned on the ancestry state of individual j. The π_{ij} values are called "individual-specific allele frequencies." These allele frequencies can be collected into an $m \times n$ matrix $\mathbf{F}$, where the (i, j) entry is π_{ij}. Note that $\mathrm{E}[x_{ij}/2|\boldsymbol{z}_j] = \pi_{ij}$, and when Hardy-Weinberg equilibrium holds, $x_{ij}|\boldsymbol{z}_j \sim$ Binomial$(2, \pi_{ij})$. We utilize the framework from Hao et al. (2013) [14] that allows the simultaneous estimation of all π_{ij} from a given genotype data set $\mathbf{X}$. Specifically, it provides estimates of latent variables that form a linear basis of the $\mathrm{logit}(\pi_{ij}) = \log(\frac{\pi_{ij}}{1-\pi_{ij}})$ quantities, which turns out is the most convenient scale on which to estimate a model of structure for the proposed testing framework. The model and estimation procedure is called "logistic factor analysis" (LFA). It should be noted that other well-behaved estimates of π_{ij} may be utilized as well. Further details are provided in METHODS.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Trait Models

We assume a trait (quantitative or binary) has been measured on each individual, which we denote by y_j, $j = 1, 2, \ldots, n$. One way in which spurious associations occur in the presence of population structure is that SNPs become correlated with each other when structure is not taken into account. Therefore, if a SNP is causal for the trait of interest, then any other SNP correlated with this causal SNP may also show an association. For SNPs in linkage disequilibrium due to their physical proximation with the causal SNP, one expects these to be associated with the trait regardless of structure. However, in the presence of structure, there may be many unlinked SNPs that also show associations with the trait due the fact that structure induces correlations of these SNPs with the causal SNP. Indeed, one of the early methods for detecting structure in association studies was to show that many randomly chosen, unlinked SNPs show associations to the trait [4]. This source of confounding is typically the main focus of association tests designed for structured populations.

Another key issue that is less often considered is the fact that lifestyle and environment are also often related to ancestry (Figure 1a). This implies that non-genetic effects may also be directly related to structure. We therefore extend the concept of the latent variable $\boldsymbol{Z}$ to include not only population genetic structure, but also lifestyle and environment: $\boldsymbol{Z} =$ (structure, lifestyle, environment). For each observed individual j, there is an underlying latent variable $\boldsymbol{z}_j$ that contains the information about structure, lifestyle, and environment for individual j. We allow for the case that structure or ancestry may be directly influential on or related to lifestyle and environment, and that all three of these variables may influence the trait of interest. An association test that is immune to structure should also be immune to the non-genetic effects that are confounded with structure.

We consider the following models of quantitative and binary traits. We write the trait models in terms of additive genetic effects, but the framework can be extended to account for dominance models and interactions, and the models can also incorporate adjustment variables that capture known sources of trait variation.

The quantitative trait model is

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

where β_i is the genetic effect of SNP i on the trait, λ_j is the random non-genetic effect, and ϵ_j is the random noise variation. To allow the interdependence of structure, lifestyle, and

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

environment, we assume that $\mathbf{x}^j = (x_{1j}, \ldots, x_{m,j})^T$, λ_j, and σ_j^2 may all be functions of $\boldsymbol{z}_j$. We assume that $\mathrm{E}[\epsilon_j|\boldsymbol{z}_j] \sim$ Normal$(0, \sigma_j^2(\boldsymbol{z}_j))$, which allows for heteroskedasticity of the random noise variation. The distribution of λ_j can remain unspecified, although we assume that λ_j and $\boldsymbol{z}_j$ may be dependent random variables. The population genetic model summarized shows how the distribution of $(x_{ij})_{i=1}^m$ depends on $\boldsymbol{z}_j$. Without having observed $\boldsymbol{z}_j$, it follows that $(x_{ij})_{i=1}^m$, λ_j, and ϵ_j are dependent random variables; however, we assume that conditional on $\boldsymbol{z}_j$, these random variables are independent.

The binary trait model is

$$\log\left(\frac{\Pr(y_j = 1)}{\Pr(y_j = 0)}\right) = \alpha + \sum_{i=1}^{m} \beta_i x_{ij} + \lambda_j, \tag{2}$$

where again β_i is the genetic effect of SNP i on the trait, λ_j is the non-genetic effect, and we allow for the case that $\mathbf{x}^j$ and λ_j may be dependent due to the common confounding latent variable $\boldsymbol{z}_j$ as described for the quantitative trait model.

We have shown that the linear mixed effects model and principal components approaches involve more restrictive assumptions about the trait models utilized in testing for associations (METHODS).

Motivation and Rationale of the Proposed Test

The rationale for the proposed test is schematized in Figure 1. The SNP X_i and the trait Y become spuriously associated because they are under the influence of a common latent variable $\boldsymbol{Z}$. This latent variable contains information on population structure, lifestyle, and environment, all of which may be interdependent and play a determining role in the trait. The problem is that we cannot directly observe $\boldsymbol{Z}$ and we would like to avoid making assumptions about its mathematical form. If we can successfully construct either $X_i|\boldsymbol{Z}$ (the distribution of X_i conditional on $\boldsymbol{Z}$) or $Y|\boldsymbol{Z}$, then it is possible to perform a test of association between X_i and Y that is immune to the effects of $\boldsymbol{Z}$. Possible association tests should occur between $X_i|\boldsymbol{Z}$ and Y, between X_i and $Y|\boldsymbol{Z}$, or between $X_i|\boldsymbol{Z}$ and $Y|\boldsymbol{Z}$.

The linear mixed model and principal components approaches can be interpreted as attempts to estimate a model of $Y|\boldsymbol{Z}$. This requires additional assumptions about non-genetic and genetic effects, and their relationship to $\boldsymbol{Z}$, specifically there is no relationship between structure and non-genetic effects in the trait model (METHODS and ref. [5]). Due to the massive number of SNPs that have been measured in GWAS, trying to construct $X_i|\boldsymbol{Z}$ is appealing

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

since we have an abundance of information about the effect of latent variables on the genotypes. (For example, this can easily be visualized in principal components constructed from the genotypes.) Our approach is therefore to carry out an association test between $X_i|\boldsymbol{Z}$ and Y by specifically testing whether there is equality or not between $\Pr(X_i|Y, \boldsymbol{Z})$ and $\Pr(X_i|\boldsymbol{Z})$ (Figure 1b). If $\Pr(X_i|Y, \boldsymbol{Z}) = \Pr(X_i|\boldsymbol{Z})$ then there is no association between the SNP X_i and the trait Y; if $\Pr(X_i|Y, \boldsymbol{Z}) \neq \Pr(X_i|\boldsymbol{Z})$, then there is an association. This test of association is in theory immune to population structure because we have taken into account $\boldsymbol{Z}$.

One remaining problem is that we cannot observe $\boldsymbol{Z}$. However, it is straightforward to show that when there is no association $\Pr(X_i|Y, \boldsymbol{Z}, \pi_i(\boldsymbol{Z})) = \Pr(X_i|Y, \pi_i(\boldsymbol{Z}))$ and $\Pr(X_i|\boldsymbol{Z}, \pi_i(\boldsymbol{Z})) = \Pr(X_i|\pi_i(\boldsymbol{Z}))$. In other words, in order to capture $X_i|\boldsymbol{Z}$, it suffices to capture $X_i|\pi_i(\boldsymbol{Z})$, the effect of $\boldsymbol{Z}$ on the allele frequency of SNP i. We have recently developed a framework that flexibly parameterizes and estimates $X_i|\pi_i(\boldsymbol{Z})$ [14]. In order to test whether $\Pr(X_i|Y, \pi_i(\boldsymbol{Z})) = \Pr(X_i|\pi_i(\boldsymbol{Z}))$, we perform a logistic regression of the SNP genotypes X_i on the trait Y plus the transformed individual-specific allele frequencies, $\mathrm{logit}(\pi_i(\boldsymbol{Z}))$, where $\mathrm{logit}(p) = \log(\frac{p}{1-p})$ for $0 < p < 1$. This inverse regression approach is a substantial departure from the most commonly employed methods that attempt to adjust for population structure.

Association Test Immune to Population Structure

We have derived a statistical hypothesis test of association that is equivalent to testing whether $\beta_i = 0$ for each SNP i in the above trait models (1) and (2), 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 (METHODS). 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.

We have proved a theorem (METHODS) that shows that $\beta_i = 0$ in models (1) and (2) implies that $b_i = 0$ in the following model:

$$\begin{aligned} x_{ij}|y_j, \boldsymbol{z}_j &\sim \text{Binomial}\left(2, \mathrm{logit}^{-1}(a_i + b_i y_j + \mathrm{logit}(\pi_{ij}))\right), \\ \mathrm{logit}\left(\frac{\mathrm{E}[x_{ij}|y_j, \boldsymbol{z}_j]}{2}\right) &= a_i + b_i y_j + \mathrm{logit}(\pi_{ij}) \end{aligned} \tag{3}$$

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). Note that the non-genetic effects, heteroskedasticity, and polygenic background do not appear in the above model used to test for associations. This is

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

important because under our general assumptions, these terms can be difficult or even impossible to estimate in practice. Furthermore, testing for association under this model means that the test will have a valid null distribution regardless of the form of the non-genetic effects, heteroskedasticity, and polygenic background.

As fully detailed in METHODS, an association statistic whose null distribution is known can be constructed by testing whether $b_i = 0$ in the above model, which we have shown is a valid test of $\beta_i = 0$ in traits models (1) and (2). Briefly, the testing procedure works as follows:

1. Formulate and estimate a model of population structure that provides well-behaved estimates of the $\mathrm{logit}(\pi_{ij})$ values. We specifically use the logistic factor analysis (LFA) approach of ref. [14], which has been shown to provide a accurate linear basis of the $\mathrm{logit}(\pi_{ij})$ values.

2. For each SNP i, perform a logistic regression of the SNP genotypes on the trait values plus the model terms that estimate the $\{\mathrm{logit}(\pi_{ij})\}_{j=1}^n$ values[1]. Also, perform a logistic regression of the SNP genotypes on only the model terms that estimate $\{\mathrm{logit}(\pi_{ij})\}_{j=1}^n$, where the trait is now excluded from the fit. These two model fits are compared via a likelihood ratio statistic, where the larger the statistic, the more evidence there is that $b_i \neq 0$.

3. Calculate a p-value for each SNP based on our result that when the null hypothesis of no association is true, $\beta_i = 0$ in models (1) and (2), then the above statistic follows a χ_1^2 distribution for large sample sizes.

We call our proposed test the "genotype conditional association test" (GCAT). As a general concept, such an approach is sometimes called an inverse regression model because we consider $\mathrm{E}[x|y]$ rather than $\mathrm{E}[y|x]$.

Simulation Studies

We performed an extensive set of simulations to demonstrate that the proposed test is robust to population structure and to assess its power to detect true associations (full technical details in METHODS). We compared the proposed test to its oracle version where model (3) and test-statistic (6) are used with the true π_{ij} values. We also included in the simulations studies three

[1] In our implementation, the logistic factors are included as covariates, which serve as the model terms that estimate the $\{\mathrm{logit}(\pi_{ij})\}_{j=1}^n$ values.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

important and popular methods: (i) the method of adjusting the trait and genotypes by principal components computed from the full set of genotypes [8] and (ii) two implementations of the linear mixed effects model approach [9, 10], specifically EMMAX by Kang et al. (2010) [10] and GEMMA by Zhou and Stephens (2012) [15]. The methods are abbreviated as "PCA," "LMM-EMMAX," and "LMM-GEMMA."

The complete simulation study on quantitative traits involved population structure constructed in 11 different ways for each of three different apportionments of variance among genetic effects, non-genetic effects, and random variation that all contribute to variation in the trait. Therefore, each configuration involved a constructed allele frequency matrix $\mathbf{F}$ and values assigned to variances $\mathrm{Var}(\sum_{i=1}^{n} \beta_i x_{ij})$, $\mathrm{Var}(\sum_{j=1}^{n} \lambda_j)$, and $\mathrm{Var}(\epsilon_j)$ from model (1). For each of these $33 = 11 \times 3$ configurations, we simulated 100 GWAS data sets, for a grand total of 3300 studies.

We simulated allele frequencies: (i) subject to structure estimated from three real data sets: HapMap, Human Genome Diversity Project (HGDP), and the 1000 Genomes Project (TGP), where the HapMap structure was simulated according to the Balding-Nichols model; (ii) at four different levels of admixture in the Pritchard-Stephens-Donnelly (PSD) model, which is an extension of the Balding-Nichols model; and (iii) for four different types of spatially defined structure. We intentionally simulated challenging population structures, having in mind that future GWAS such as the forthcoming "Genotype Tissue Expression" program (GTEx) data may involve particularly challenging forms of structure.

In order to provide an extra challenge to the proposed test, we simulated the allele frequencies from a model that differs from LFA model (4). We generated allele frequencies parameterized by $\mathbf{F} = \mathbf{\Gamma S}$, where $\mathbf{F}$ is the matrix of π_{ij} values, $\mathbf{\Gamma}$ is an $m \times d$ matrix and $\mathbf{S}$ is the $d \times n$ matrix that encapsulates the structure. This model captures as special cases the Balding-Nichols model and the PSD model [14]. It was also intended to provide an advantage to the PCA and LMM methods because the structure is manifested on the observed genotype scale [14], which is the same scale on which both methods estimate structure.

We simulated 10 truly associated SNPs whose effect sizes are distributed according to a Normal distribution. All genotypes were simulated to be in linkage equilibrium so that true and false positives are unambiguous. We set the variances $\mathrm{Var}(\sum_{i=1}^{n} \beta_i x_{ij})$, $\mathrm{Var}(\sum_{j=1}^{n} \lambda_j)$, and $\mathrm{Var}(\epsilon_j)$ to be: (5%, 5%, 90%), (10%, 0%, 90%), and (10%, 20%, 70%). Setting these variances enforced a certain overall level of genetic contribution to the trait; therefore our simulation study results were minimally affected by the choice of 10 truly associated SNPs

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and the Normal distribution on their effect sizes. In each simulation scenario, we simulated data for $m = 100,000$ SNPs and $n = 5000$ individuals, except HGDP necessarily restricted us to $n = 940$ individuals and TGP to $n = 1500$ individuals. The dimension of the structure was set to $d = 3$, although we carried out the same simulations for $d = 6$ and the results were quantitatively very similar and qualitatively equivalent.

Each simulation configuration involved analyzing 100 GWAS data sets $(\mathbf{X}, \mathbf{y})$, where the Oracle method, the proposed GCAT method, and the PCA method were applied to each study. For a given simulated study, we obtained a set of $m =$ 100,000 p-values. So-called "spurious associations" occur when the p-values corresponding to null (non-associated) SNPs are artificially small. For a given p-value threshold t, we expect there to be $m_0 \times t$ false positives among the m_0 p-values corresponding to null SNPs, where $m_0 = 100,000 - 10$ in our case. At the same time, we can calculate the observed number of false positive simply by counting how many of the null SNP p-values are less than or equal to t. The excess observed false positives are spurious associations. A method properly accounts for structure when the average difference is zero. The best one can do on a study-by-study basis is captured by the Oracle method, which according to our theory is immune to structure and provides the correct null distribution.

We found from using the distributed binary executable EMMAX software and our own implementation that EMMAX required a 10-fold increase in computational time over the proposed method and PCA when analyzing $n = 5000$ individuals. Therefore, it was not reasonable to apply EMMAX to all 3300 simulated GWAS data sets. We limited comparisons with EMMAX to five representative structure configurations of the full 11 for a single apportionment of the variances assigned to genetic effects, non-genetic effects, and random variation. GEMMA was computationally more efficient, though still significantly slower than GCAT or our implementation of PCA. Figure 2 shows the excess in observed false positives vs. the expected number of false positives for the Oracle, GCAT (proposed), PCA, and both implementations of LMM methods under five configurations of structure for the variance configuration corresponding to genetic=5%, environmental=5%, and noise=90%. It can be seen that the LFA implementation of the proposed GCAT method performs similarly to the Oracle test, whereas PCA tends to suffer from an excess of spurious associations. Figures S1-S8 is a more complete set of simulations with results from the all three sets of variances for the full 11 configurations of structure. Due to the computational constraints mentioned above, the additional simulations feature only results from GEMMA for LMM methods.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In comparing the statistical power among the methods (Figures S9-S17), we found that the Oracle, GCAT, and PCA performed similarly well, while the two LMM methods often suffered from a loss of power. We also carried out analogous simulations on binary traits simulated from model (2) and we found that all methods performed similarly well in terms of producing correct p-values that were robust to structure. This result agrees with the comparisons made between PCA and a linear mixed effects model in Astle and Balding (2009) [5].

Analysis of the Northern Finland Birth Cohort Data

We applied the proposed method to the Northern Finland Birth Cohort (NFBC) genome-wide association study data [12], which includes several metabolic traits and height. This study has also been analyzed by the LMM and PCA methods, as well as a standard analysis uncorrected for structure [10]. We carried out association analyses with the proposed method on the 10 traits that were also analyzed using the other methods (Table 1). After processing the data, including filtering for missing data, minor allele frequencies, and departures from Hardy-Weinberg equilibrium, the data were composed of $m = 324,160$ SNPs and $n = 5027$ individuals (METHODS). The logistic factors were computed on a subset of the data where markers were at least 200 kbp apart.

Most traits showed only approximate Normal distributions, so we applied a Box-Cox Normal transformation to all traits so that they satisfy the model assumptions. We noted that C-reactive Protein (CRP) and Triglycerides (TG) traits followed an exponential distribution more closely, so it was unnecessary to transform these two traits. The developed theory can be extended to exponential distributed quantitative traits as well.

The 20 most significant SNPs for each of the 10 traits are shown in Table S1. Kang et al. (2010) utilized a genome-wide significance threshold of p-value $< 7.2 \times 10^{-8}$ as proposed in ref. [16], so we also utilized this threshold for comparative purposes. The number of loci found to be significant for each method are shown in Table 1. Whereas our proposed method identifies 16 significant loci, the other methods identify 11 to 14 loci.

We identified three new loci that were not identified by the other methods. None of the other methods identified any significant associations for the height trait. However, we identified rs2814982 on chromosome 6 as being statistically associated with height (Table S1). This SNP is located ~ 70kbp from another SNP, rs2814993, which has been associated with skeletal frame size in a previous study [13]. Additionally, rs2814993 was the fifth most signifi-

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cant SNP for height. For the LDL cholesterol trait, we identified a significant association with rs11668477, which was significantly associated with LDL cholesterol in a different study [17]. Finally, there were significant associations between the glucose (GLU) trait and a cluster of SNPs (rs3847554, rs1387153, rs1447352, rs7121092) proximal to the *MTNR1B* locus; variation at this locus has been associated with glucose in a previous study [18].

As described in Sabatti et al. (2009) [12], the NFBC data show modest levels of inflation due to population structure as measured by the genomic control inflation factor (GCIF) [19] of test statistics from an uncorrected analysis. The population structure present among these individuals may be subtler and manifested on a finer scale than other settings. Noting that the GCAT approach does not attempt to adjust for a polygenic background, the GCIF values calculated for the proposed method (Table S2) were found to be in line with what is expected for polygenic traits where no structure is present [20], providing evidence that the proposed method adequately accounts for structure.

Discussion

We considered models of quantitative and binary traits involving genetic effects and non-genetic effects in the presence of arbitrarily complex population structure. We allowed for the non-genetic effects to be confounded with population genetic structure since structure, ancestry, lifestyle, and environment – all factors potentially involved in complex traits – may be highly dependent with one another. A causal model provided the intuition that under these models, it is most reasonable to account for this confounding in the genotypes, but it is not tractable to do so in the non-genetic effects. This follows because we have many instances of genotypes that can be jointly modeled to provide reliable estimates of structure, but the non-genetic effects are never directly observed and we do not have repeated instances of them. In general it is not possible to estimate a latent variable that accounts for the confounding between structure and non-genetic effects.

These observations led us to propose an inverse regression approach to testing for associations, where the association is tested by modeling genotype variation in terms of the trait plus model terms accounting for structure. In this model, the terms accounting for structure were based on the logistic factor analysis approach that we have proposed [14], although the general form of the association test can incorporate other methods that estimate population struc-

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ture. We mathematically proved under general assumptions that the trait term in the model is non-zero only when the genetic marker is truly associated with the trait, regardless of the population structure. We demonstrated that the implemented test properly accounts for structure in a large body of simulated studies that included a wide range of population structures. We also applied the method to 10 traits from the Northern Finland Birth Cohort genome-wide association study. The proposed method identified three new loci associated with the traits, including being the only method among those we considered that identifies a locus associated with the height trait. Overall, we showed that the proposed method compares favorably to existing methods and we also noted that it has favorable computational requirements compared to existing methods.

As GWAS increase in sample size and levels of complexity of population structure, it is important to develop methods that properly account for structure and scale well with sample size. Whereas we found that the popular principal components adjustment does not properly account for structure, we also found that the mixed model approach performs reasonably well. However, the mixed model approach involves estimating a $n \times n$ kinship matrix and its current implementation does not scale well with sample size. The kinship matrix quickly becomes computationally unwieldy when n grows large, and the possibility of the estimated kinship matrix becoming overwhelmed by noise is a concern [21]. In the Northern Finland Birth Cohort data, the mixed model approach required us to estimate 12 million parameters, whereas the proposed method involved estimating 25-thousand parameters, a $\sim$500-fold decrease. A study involving $n = 10,000$ individuals with the same complexity of structure requires estimating about 50-million parameters in the mixed model kinship matrix, whereas the proposed method requires estimating 50-thousand parameters, a $\sim$1000-fold decrease. In addition, estimating the structure in the proposed method primarily uses singular value decomposition, for which a rich literature of computational techniques exist. We utilized a Lanczos bidiagonalization algorithm [22] which scales approximately linearly with respect to n for $d \ll n$. The proposed method is well equipped to scale to massive GWAS and can take advantage of future advances for computing singular value decomposition.

The key assumption to verify in utilizing the proposed GCAT approach is that population structure observed in the SNP genotypes is adequately modeled and estimated. One can test for associations among SNPs that show convincing empirical evidence that the model of structure is reasonably well-behaved; this can be directly tested on the genotype data as previously demonstrated in our logistic factor analysis (LFA) model of structure [14]. For example, on the

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Northern Finland Birth Cohort Study, we empirically verified that utilizing the LFA model with dimension $d = 6$ accounted for structure reasonably well for the great majority of SNPs. The linear mixed effects model (LMM) approach and principal components (PCA) approach make trait model assumptions that may be difficult to verify in practice (METHODS).

We anticipate that the proposed genotype conditional association test (GCAT) will be useful for future studies. The mathematical framework we have developed should facilitate its extension to traits modeled according to distributions not considered here while maintaining our theoretical proof that the test accounts for population structure in the presence of non-genetic effects also confounded with structure.

METHODS

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 [14]. 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 detailed below, 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. [14] that this model performs well in estimating structure

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

resulting from discrete subpopulations, admixed populations, the Balding-Nichols model [23], the Pritchard-Stephens-Donnelly model [24], 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 [14]. 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), $\mathbf{L} = \mathbf{AH}$.

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 above trait models (1) and (2), 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 $(\beta_k)_{k \neq i}$.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 $\mathrm{E}[x|y]$ rather than $\mathrm{E}[y|x]$.

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) 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 [25]:

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

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 (METHODS) 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_j^2 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 $\mathrm{logit}(\pi_{ij}) = \sum_{k=1}^{d} a_{ik}h_{kj}$. 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:

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

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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:

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

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 $a_{ij} = \tilde{a}_{ij} = b_{ij} = 0$.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

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* $\mathrm{logit}(\pi_{ij}) = \sum_{k=1}^{d} a_{ik}h_{kj}$*. 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 [25].

Corollary 2 *Suppose that the assumptions of Theorem 1 hold and additionally* $\mathrm{logit}(\pi_{ij}) = \sum_{k=1}^{d} a_{ik}h_{kj}$*. 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 $\left(2, \mathrm{logit}^{-1}(\sum_{k=1}^{d} a_{ik}h_{kj})\right)$ 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 [25].

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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) [14]. These constructions are summarized as follows from Hao et al. (2013).

Balding-Nichols Model. 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 [26]. 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) [14]. 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) [14] 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.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 [26]. 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) [27]. 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.

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

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. [14]. The association tests were performed exactly as described in the main text.

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) [5], these approaches assume that $\lambda_j + \epsilon_j \overset{i.i.d.}{\sim}$ 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

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 [10] 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 [15].

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 [28,29], 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) [8], 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) [14], 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

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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) [5] 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 RESULTS and ref. [5]). Specifically, the PCA approach appears to perform reasonably well in adjusting for structure for the binary trait simulations that we considered.

Software Implementation

The proposed method has been implemented in open source software, which will be made publicly available upon publication.

References

[1] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* **9**(5), 356–369 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[2] Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**(4), 241–251, Apr (2009).

[3] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145), 661–678, Jun (2007).

[4] Pritchard, J. K. and Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* **65**(1), 220–228, Jul (1999).

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

[6] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**(7), 459–463, Jun (2010).

[7] Zhang, S., Zhu, X., and Zhao, H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic epidemiology* **24**(1), 44–56 (2003).

[8] 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).

[9] Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**(2), 203–208 (2006).

[10] 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).

[11] Wang, K., Hu, X., and Peng, Y. An analytical comparison of the principal component method and the mixed effects model for association studies in the presence of cryptic relatedness and population stratification. *Hum. Hered.* **76**(1), 1–9 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[12] Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., Sovio, U., Ruokonen, A., Laitinen, J., Jakkula, E., Coin, L., Hoggart, C., Collins, A., Turunen, H., Gabriel, S., Elliot, P., McCarthy, M. I., Daly, M. J., Järvelin, M.-R., Freimer, N. B., and Peltonen, L. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics* **41**(1), 35–46 (2009).

[13] Soranzo, N., Rivadeneira, F., Chinappen-Horsley, U., Malkina, I., Richards, J. B., Hammond, N., Stolk, L., Nica, A., Inouye, M., Hofman, A., Stephens, J., Wheeler, E., Arp, P., Gwilliam, R., Jhamai, P. M., Potter, S., Chaney, A., Ghori, M. J. R., Ravindrarajah, R., Ermakov, S., Estrada, K., Pols, H. A. P., Williams, F. M., McArdle, W. L., van Meurs, J. B., Loos, R. J. F., Dermitzakis, E. T., Ahmadi, K. R., Hart, D. J., Ouwehand, W. H., Wareham, N. J., Barroso, I., Sandhu, M. S., Strachan, D. P., Livshits, G., Spector, T. D., Uitterlinden, A. G., and Deloukas, P. Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size. *PLoS genetics* **5**(4), e1000445 (2009).

[14] 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).

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

[16] Dudbridge, F. and Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology* **32**(3), 227–234 (2008).

[17] Sandhu, M. S., Waterworth, D. M., Debenham, S. L., Wheeler, E., Papadakis, K., Zhao, J. H., Song, K., Yuan, X., Johnson, T., Ashford, S., Inouye, M., Luben, R., Sims, M., Hadley, D., McArdle, W., Barter, P., Kesäniemi, Y. A., Mahley, R. W., McPherson, R., Grundy, S. M., Wellcome Trust Case Control Consortium, Bingham, S. A., Khaw, K.-T., Loos, R. J. F., Waeber, G., Barroso, I., Strachan, D. P., Deloukas, P., Vollenweider, P., Wareham, N. J., and Mooser, V. Ldl-cholesterol concentrations: a genome-wide association study. *Lancet* **371**(9611), 483–491 (2008).

[18] Prokopenko, I., Langenberg, C., Florez, J. C., Saxena, R., Soranzo, N., Thorleifsson, G., Loos, R. J. F., Manning, A. K., Jackson, A. U., Aulchenko, Y., Potter, S. C., Erdos, M. R.,

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Sanna, S., Hottenga, J.-J., Wheeler, E., Kaakinen, M., Lyssenko, V., Chen, W.-M., Ahmadi, K., Beckmann, J. S., Bergman, R. N., Bochud, M., Bonnycastle, L. L., Buchanan, T. A., Cao, A., Cervino, A., Coin, L., Collins, F. S., Crisponi, L., de Geus, E. J. C., Dehghan, A., Deloukas, P., Doney, A. S. F., Elliott, P., Freimer, N., Gateva, V., Herder, C., Hofman, A., Hughes, T. E., Hunt, S., Illig, T., Inouye, M., Isomaa, B., Johnson, T., Kong, A., Krestyaninova, M., Kuusisto, J., Laakso, M., Lim, N., Lindblad, U., Lindgren, C. M., McCann, O. T., Mohlke, K. L., Morris, A. D., Naitza, S., Orrù, M., Palmer, C. N. A., Pouta, A., Randall, J., Rathmann, W., Saramies, J., Scheet, P., Scott, L. J., Scuteri, A., Sharp, S., Sijbrands, E., Smit, J. H., Song, K., Steinthorsdottir, V., Stringham, H. M., Tuomi, T., Tuomilehto, J., Uitterlinden, A. G., Voight, B. F., Waterworth, D., Wichmann, H.-E., Willemsen, G., Witteman, J. C. M., Yuan, X., Zhao, J. H., Zeggini, E., Schlessinger, D., Sandhu, M., Boomsma, D. I., Uda, M., Spector, T. D., Penninx, B. W., Altshuler, D., Vollenweider, P., Jarvelin, M. R., Lakatta, E., Waeber, G., Fox, C. S., Peltonen, L., Groop, L. C., Mooser, V., Cupples, L. A., Thorsteinsdottir, U., Boehnke, M., Barroso, I., Van Duijn, C., Dupuis, J., Watanabe, R. M., Stefansson, K., McCarthy, M. I., Wareham, N. J., Meigs, J. B., and Abecasis, G. R. Variants in mtnr1b influence fasting glucose levels. *Nature genetics* **41**(1), 77–81 (2009).

[19] Devlin, B. and Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

[20] Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'Connell, J. R., Mangino, M., Mägi, R., Madden, P. A., Heath, A. C., Nyholt, D. R., Martin, N. G., Montgomery, G. W., Frayling, T. M., Hirschhorn, J. N., McCarthy, M. I., Goddard, M. E., Visscher, P. M., and GIANT Consortium. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics* **19**(7), 807–812 (2011).

[21] Witten, D. M., Tibshirani, R., and Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**(3), 515–534 (2009).

[22] Baglama, J. and Reichel, L. Restarted block lanczos bidiagonalization methods. *Numerical Algorithms* **43**, 251–272 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[23] 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).

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

[25] 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).

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

[27] 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).

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

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

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures and Tables

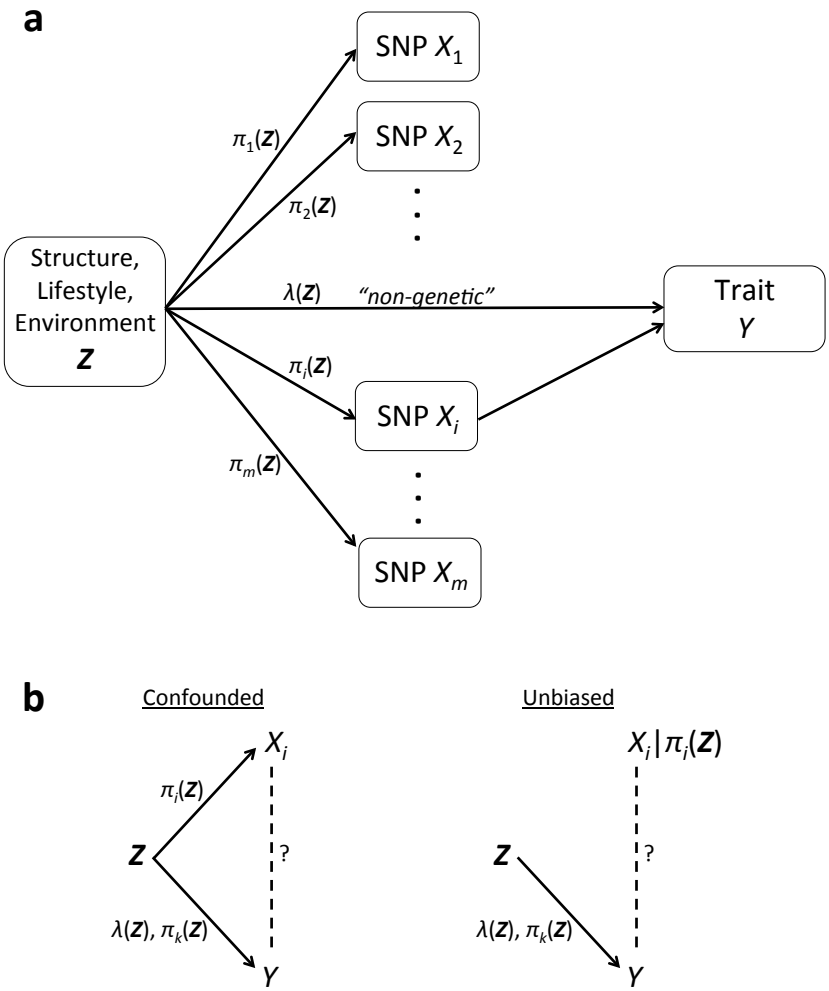

Figure 1: Rationale for the proposed test of association. (a) A graphical model describing population structure and its effects on traits. Population structure serves as a latent causal variable common among a set of loci, via the allele frequencies. When one locus has a causal effect on the trait, this induces spurious associations with other loci affected by population structure. At the same time, population structure may be highly related to lifestyle and the environment as these are all possibly related to ancestry and geography. (b) Accounting for confounding due to latent population structure. Left panel: A test for association between the ith SNP X_i and trait Y without taking into account $\boldsymbol{Z}$ will produce a spurious association due to the fact that both X_i and Y are confounded with $\boldsymbol{Z}$. Right panel: A test for association between $X_i|\pi_i(\boldsymbol{Z})$ and Y will be an unbiased because condition on $\pi_i(\boldsymbol{Z})$ breaks the relationship between $\boldsymbol{Z}$ and X_i.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

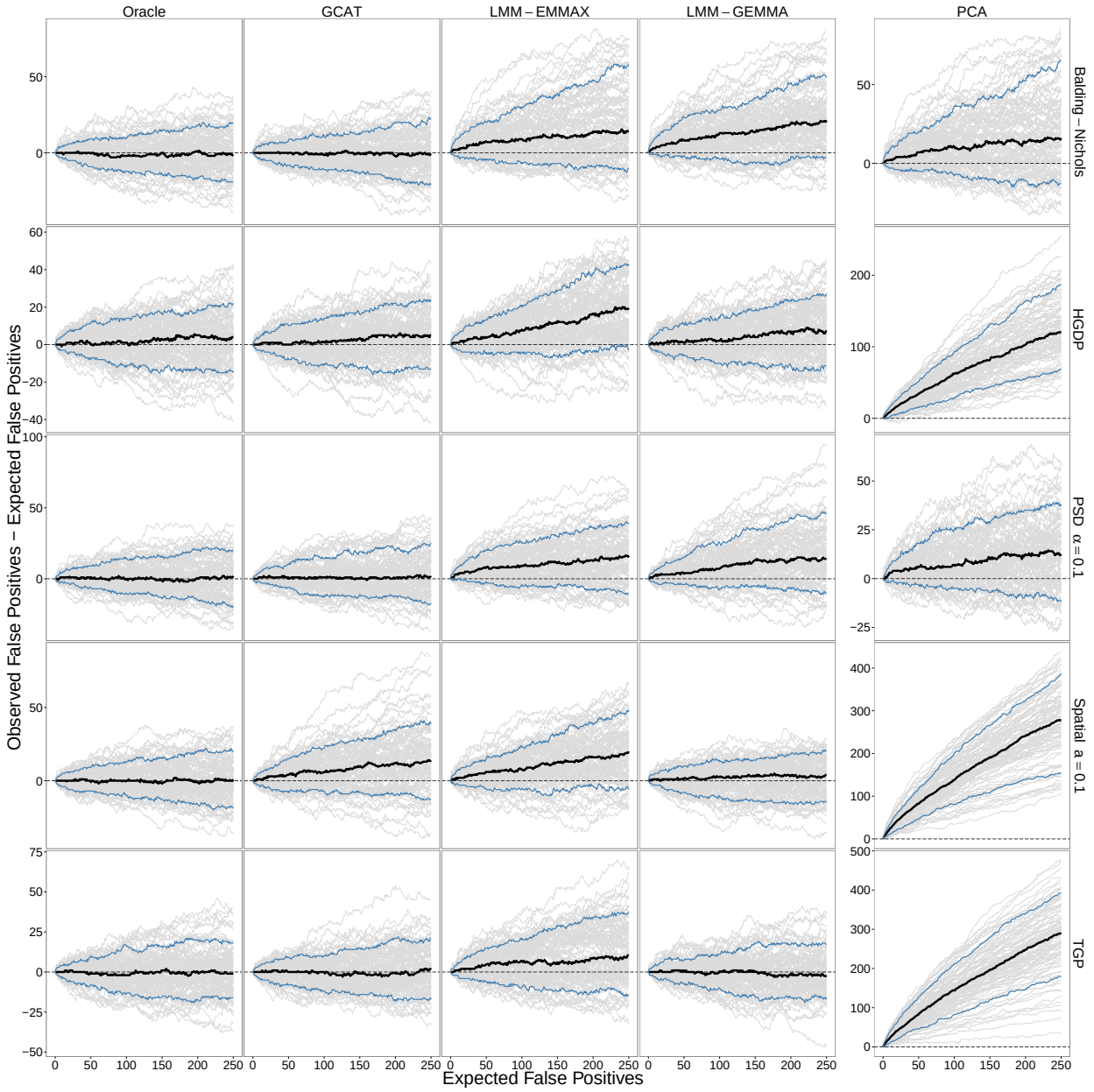

Figure 2: Performance of association tests on 100 simulated studies based off the Balding-Nichols, HGDP, TGP, PSD $\alpha = 0.1$, and spatial $a = 0.1$ simulation scenarios comparing the Oracle, GCAT (proposed), LMM (EMMAX), LMM (GEMMA), and PCA tests. The quantitative traits are based on model (1). The variance contributions to the trait are genetic=5%, environmental=5%, and noise=90%. The differences between the observed number of false positives and expected number of false positives versus the expected number of false positives under the null are plotted for each simulated study (grey lines), the average of those differences (black line), and the middle 90% (blue lines). All simulations had $m = 100,000$ SNPs, so the range of the x-axis corresponds to choosing a significance threshold of up to 0.0025. The difference on the y-axis is the number of "spurious associations." PCA is shown on a separate y-axis since it usually has a much larger maximum than the other methods. The Oracle method is where the true allele frequencies are inputted into the proposed test, which we have theoretically proven always corrects for structure.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: Number of significant loci at genome-wide significance (p-value $< 7.2 \times 10^{-8}$) for each of the 10 traits from the Northern Finland Birth Cohort. The counts for LMM+GC, PCA+GC, and Uncorr+GC are derived from Table 2 in Kang et al. (2010).

Trait	Abbreviation	GCAT+GC	LMM+GC	PCA+GC	Uncorr+GC
Body Mass Index	BMI	0	0	0	0
C-reactive Protein	CRP	2†	2	2	2
Diastolic blood pressure	DBP	0	0	0	0
Glucose	GLU	3	2	2	2
HDL Cholesterol	HDL	4	4	2	4
Height	Height	1	0	0	0
Insulin	INS	0	0	0	0
LDL Cholesterol	LDL	4	3	3	3
Systolic blood pressure	SBP	0	0	0	0
Triglycerides	TG	2	3	2	2
	Total	16	14	11	13

†Result when the Box-Cox transformation was not applied to the CRP trait. The result is 1 when the transformation is applied.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figures and Tables

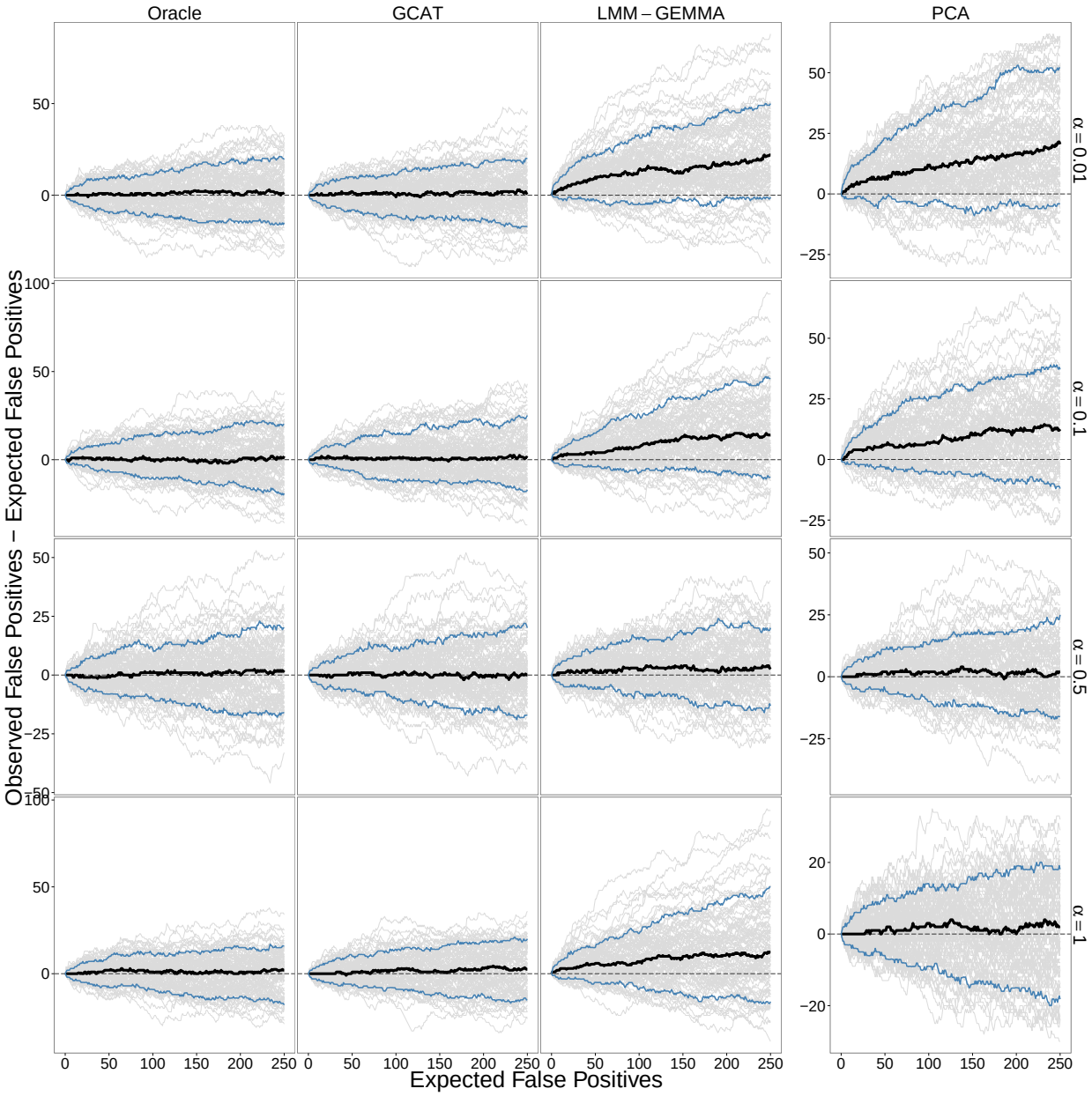

Figure S1: Performance of association tests on 100 simulated studies based off the PSD model of structure for various α comparing the Oracle, GCAT (proposed), LMM (GEMMA), and PCA tests. The variance contributions to the trait are genetic=5%, environmental=5%, and noise=90%. The remaining details are equivalent to Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

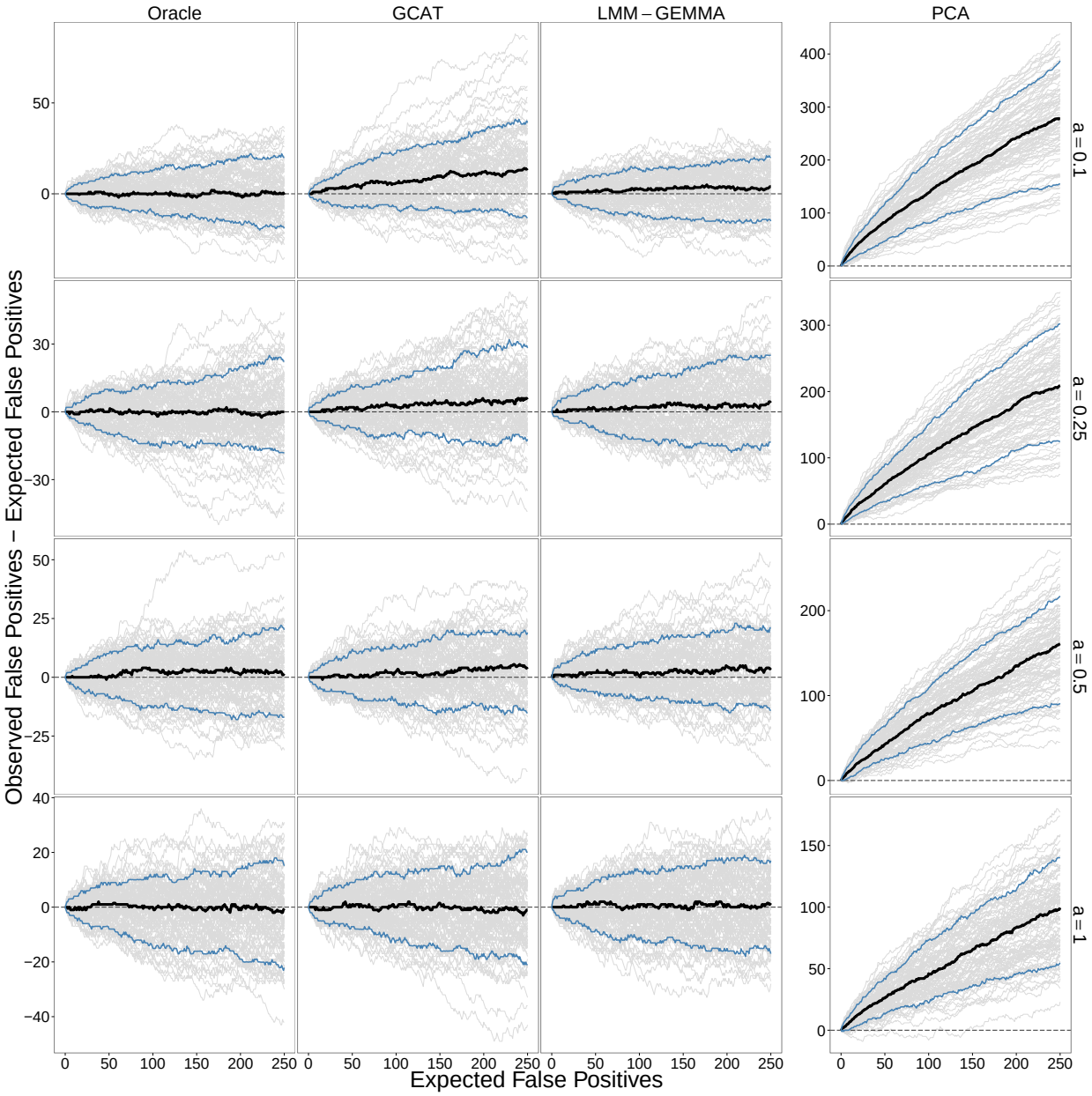

Figure S2: Performance of association tests on 100 simulated studies based off the spatial model of structure for various a comparing the Oracle, GCAT (proposed), LMM (GEMMA), and PCA tests. The variance contributions to the trait are genetic=5%, environmental=5%, and noise=90%. The remaining details are equivalent to Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

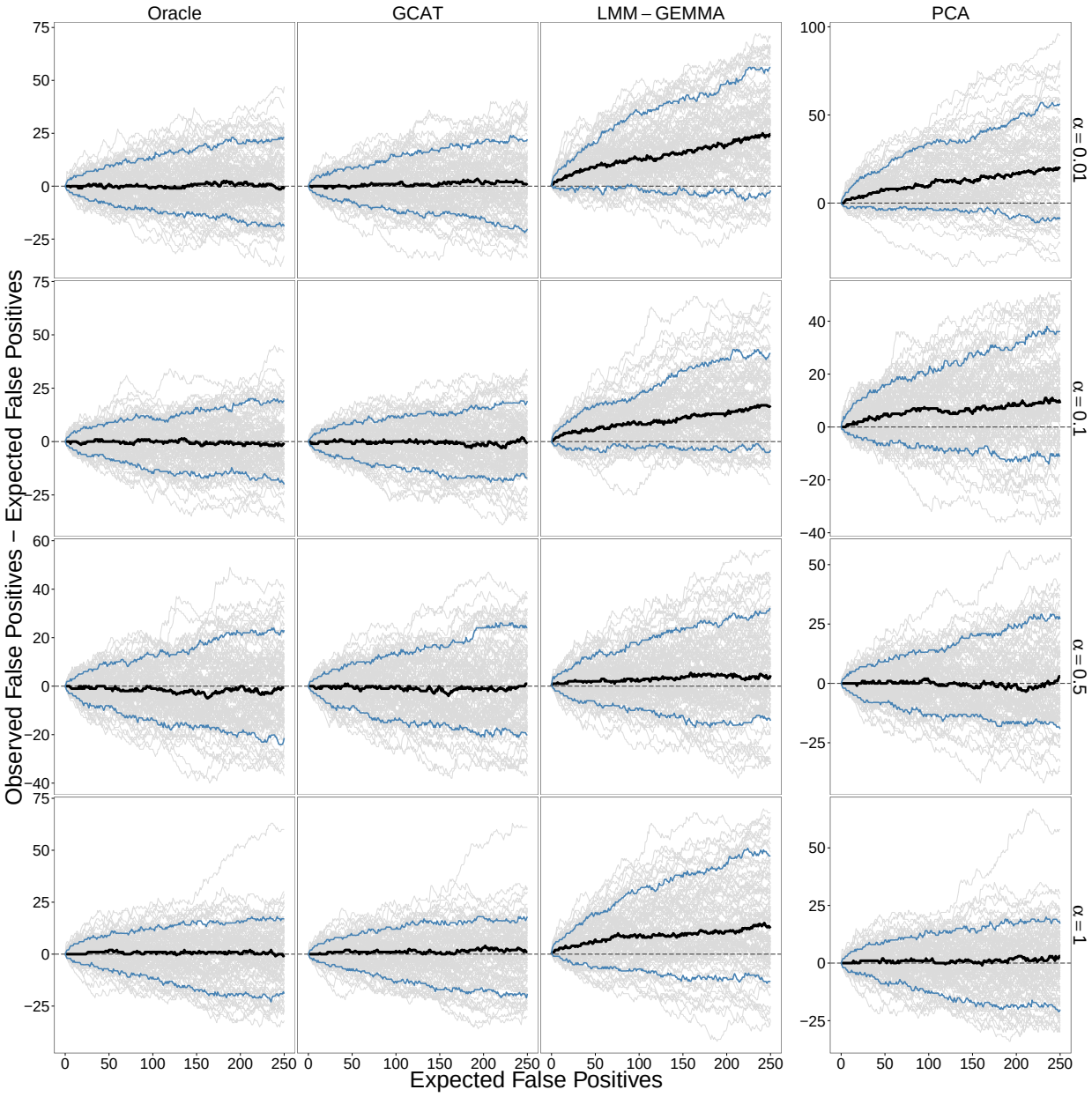

Figure S3: Performance of association tests on 100 simulated studies based off the PSD model of structure for various α comparing the Oracle, GCAT (proposed), LMM (GEMMA), and PCA tests. The variance contributions to the trait are genetic=10%, environmental=0%, and noise=90%. The remaining details are equivalent to Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

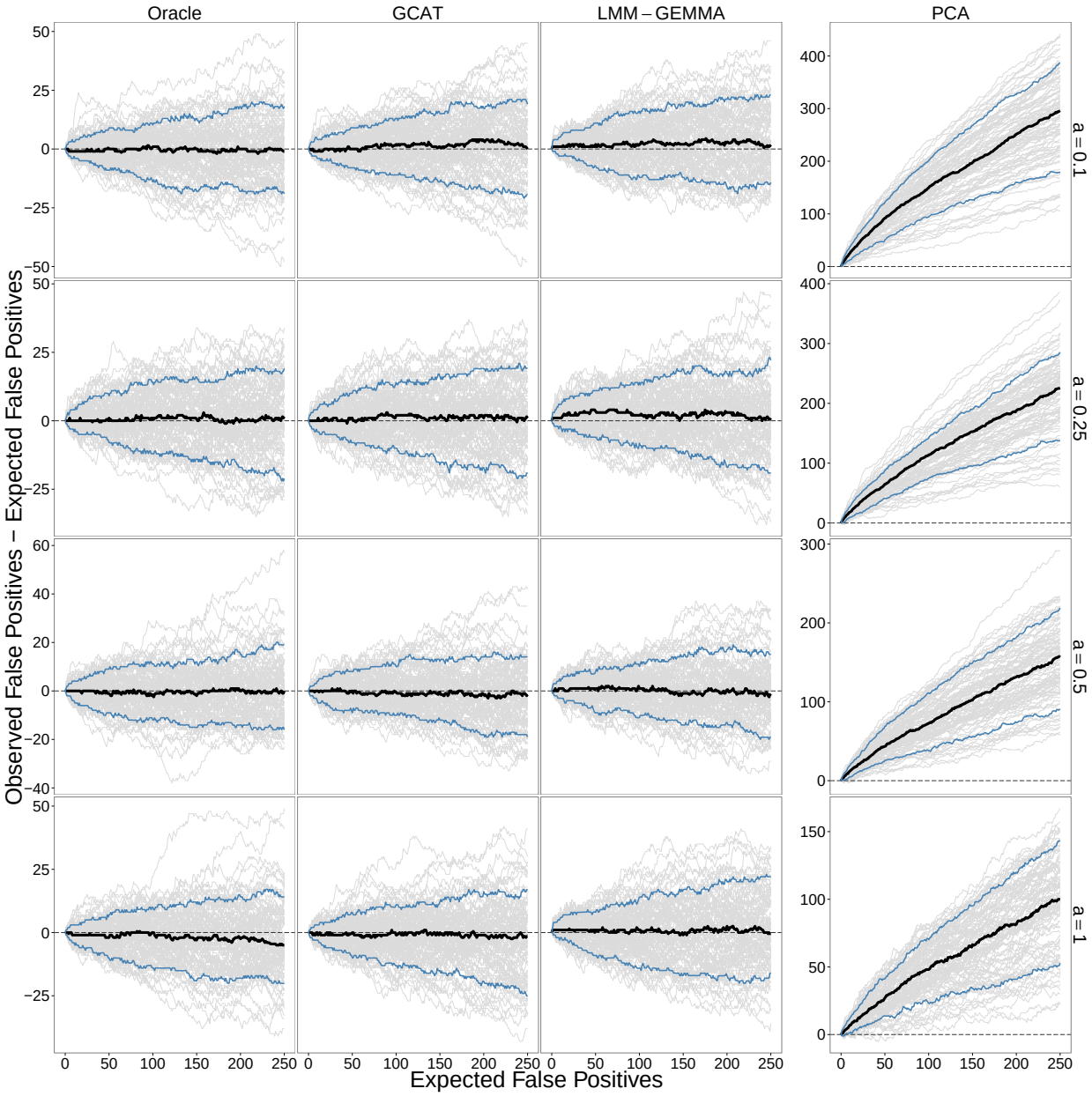

Figure S4: Performance of association tests on 100 simulated studies based off the spatial model of structure for various a comparing the Oracle, GCAT (proposed), LMM (GEMMA), and PCA tests. The variance contributions to the trait are genetic=10%, environmental=0%, and noise=90%. The remaining details are equivalent to Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

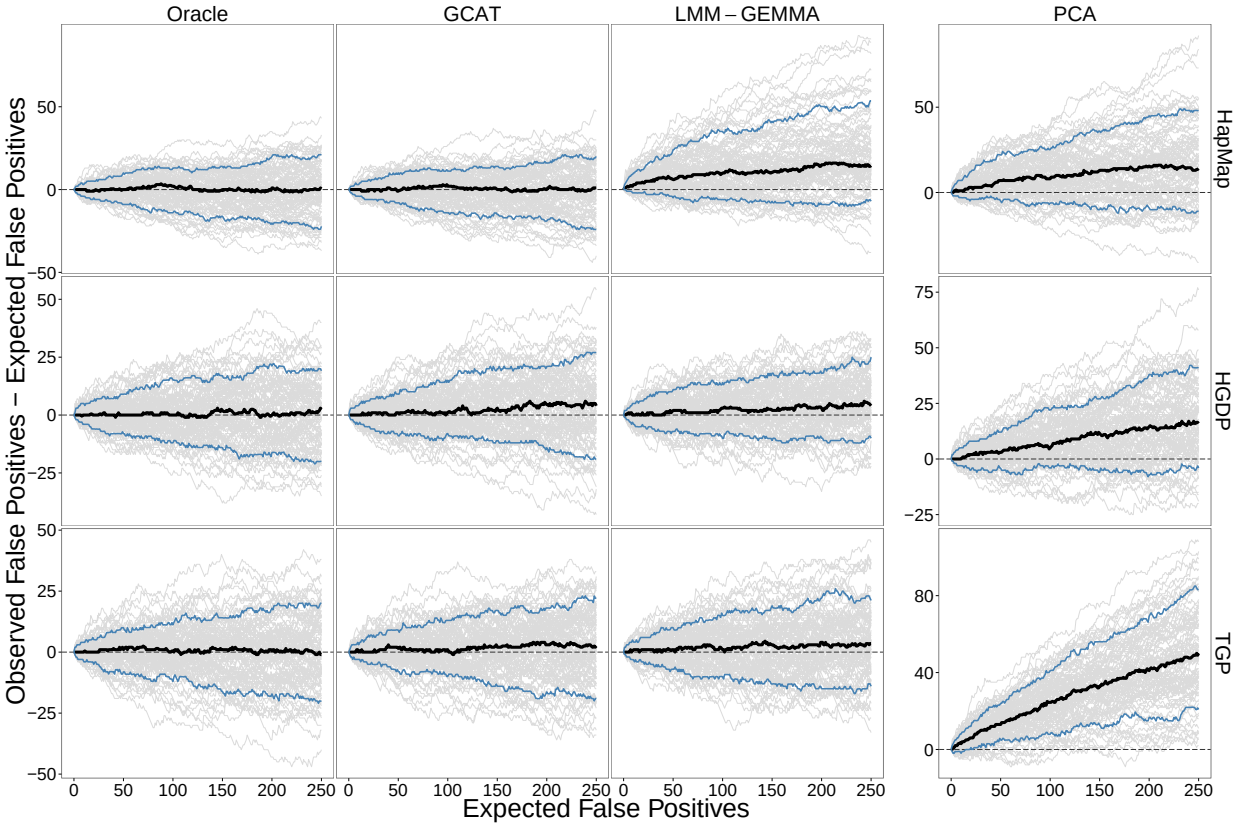

Figure S5: Performance of association tests on 100 simulated studies based off the Balding-Nichols, HGDP, and TGP simulation scenarios comparing the Oracle, GCAT (proposed), LMM (GEMMA), and PCA tests. The variance contributions to the trait are genetic=10%, environmental=0%, and noise=90%. The remaining details are equivalent to Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

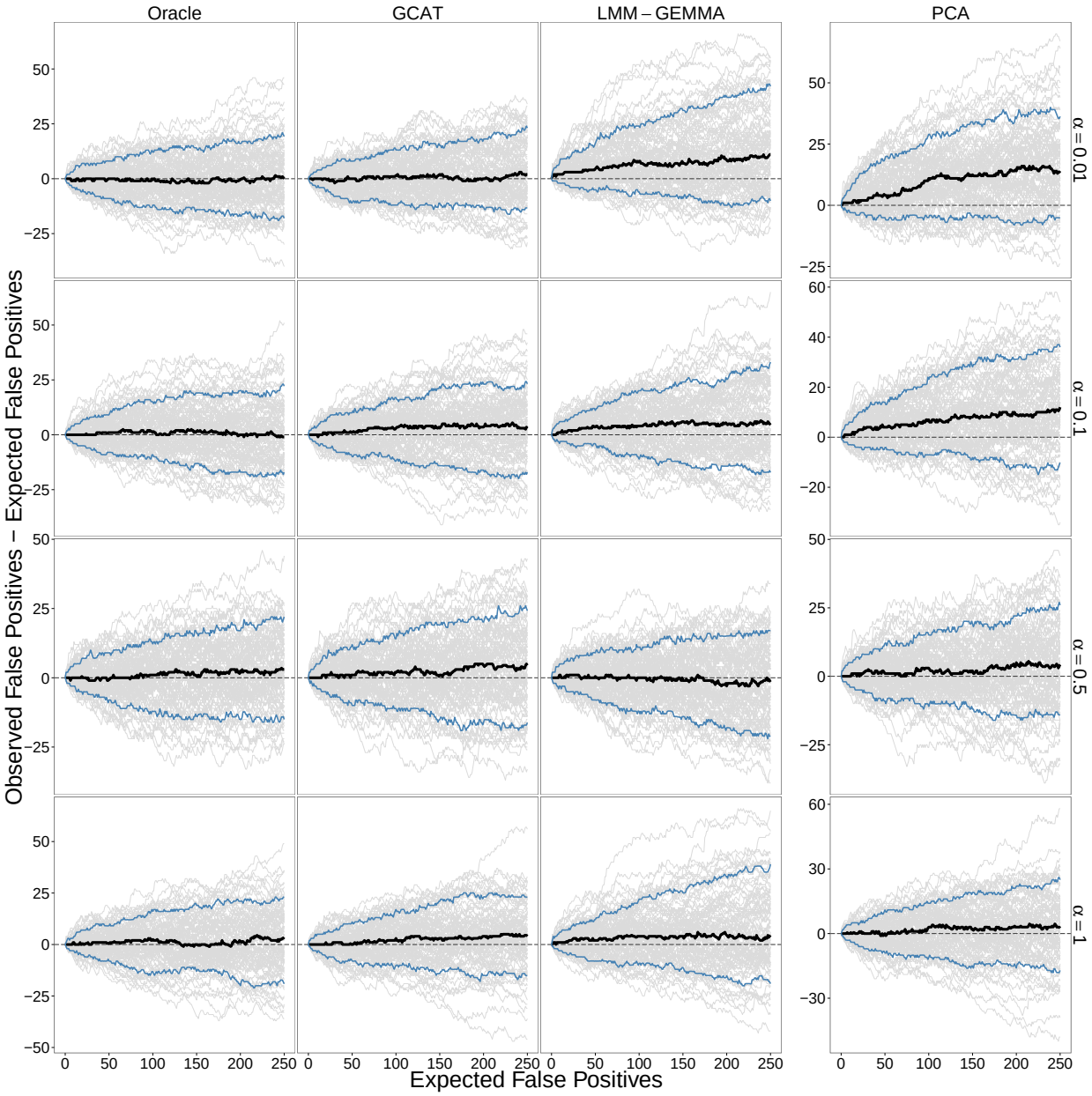

Figure S6: Performance of association tests on 100 simulated studies based off the PSD model of structure for various α comparing the Oracle, GCAT (proposed), LMM (GEMMA), and PCA tests. The variance contributions to the trait are genetic=20%, environmental=10%, and noise=70%. The remaining details are equivalent to Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

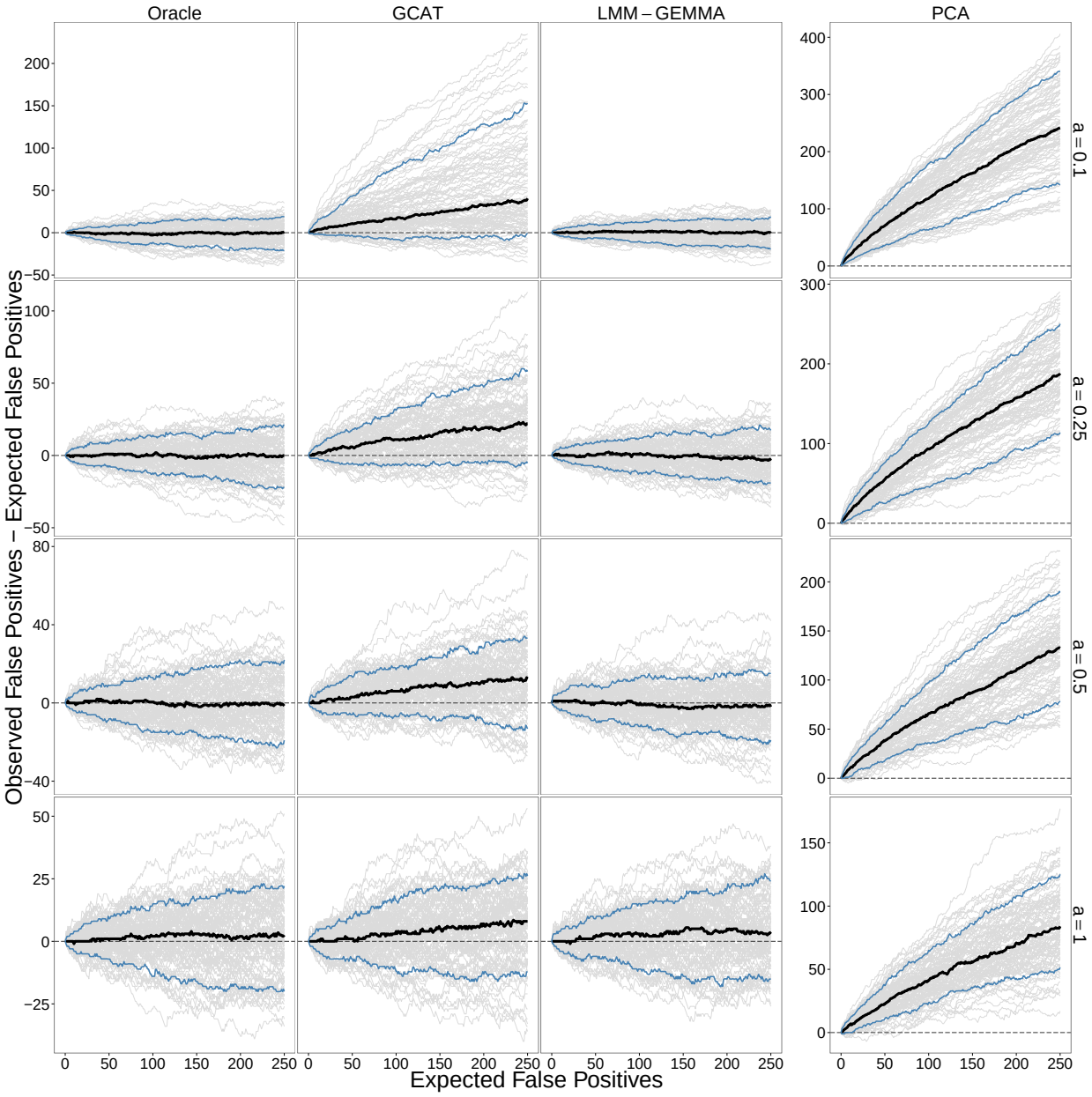

Figure S7: Performance of association tests on 100 simulated studies based off the spatial model of structure for various a comparing the Oracle, GCAT (proposed), LMM (GEMMA), and PCA tests. The variance contributions to the trait are genetic=20%, environmental=10%, and noise=70%. The remaining details are equivalent to Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

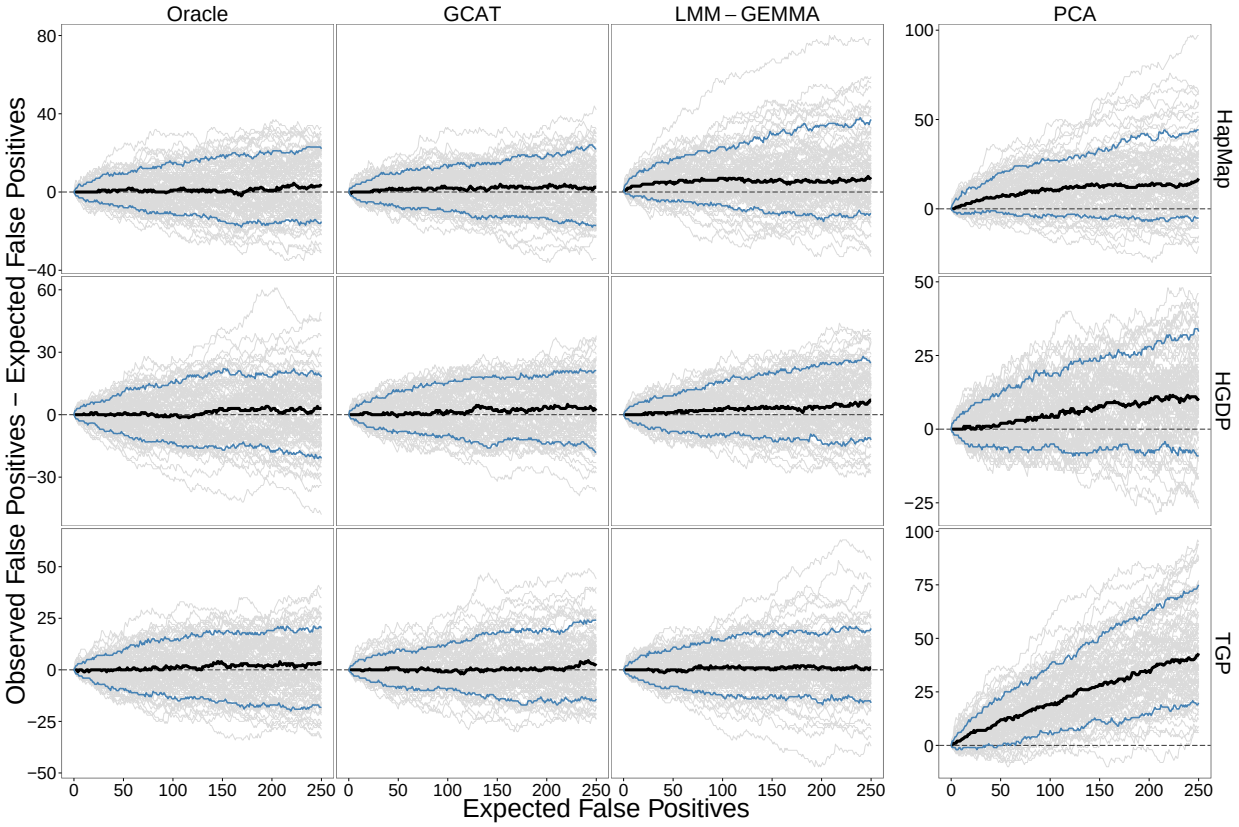

Figure S8: Performance of association tests on 100 simulated studies based off the Balding-Nichols, HGDP, and TGP simulation scenarios comparing the Oracle, GCAT (proposed), LMM (GEMMA), and PCA tests. The variance contributions to the trait are genetic=20%, environmental=10%, and noise=70%. The remaining details are equivalent to Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

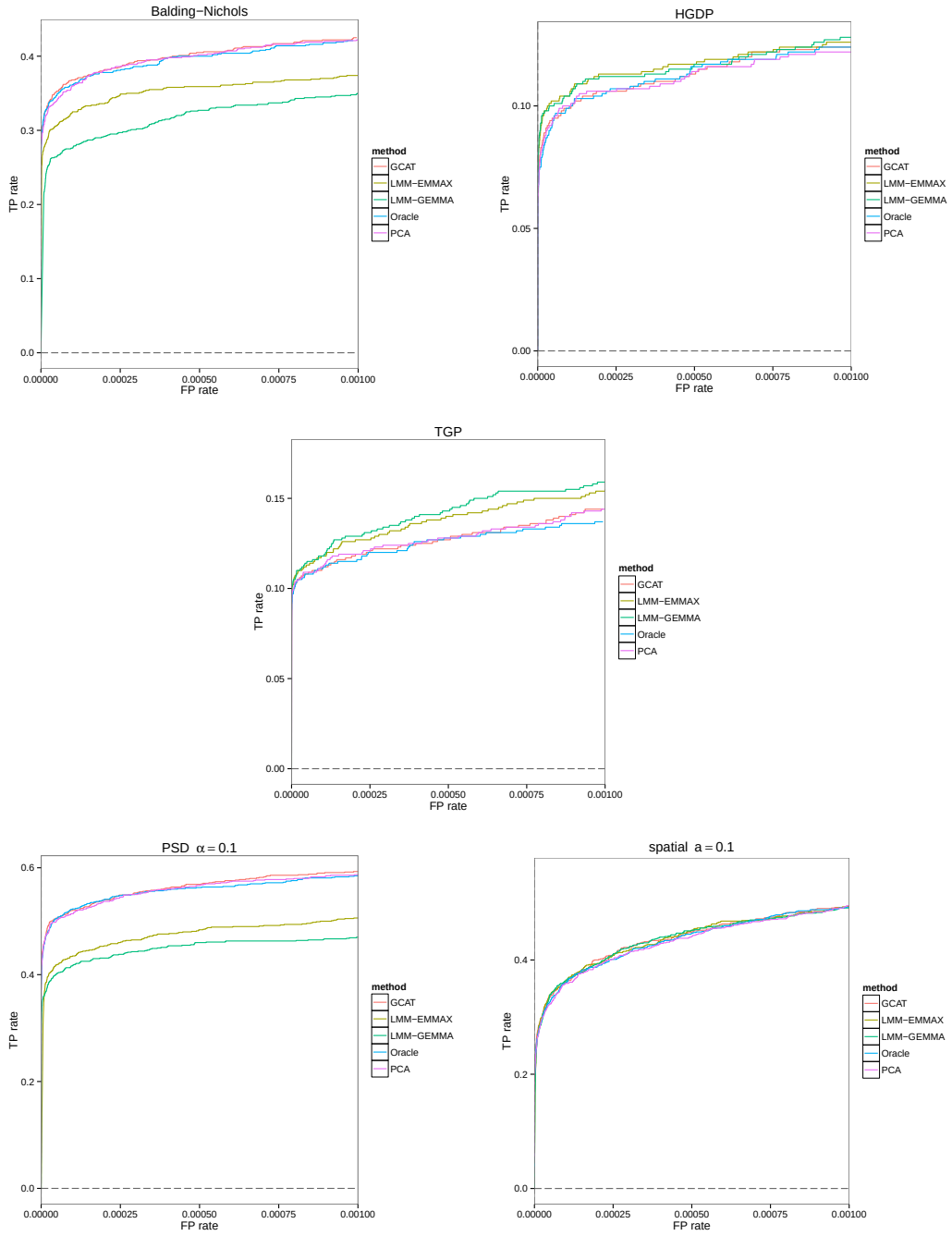

Figure S9: Statistical power of the Oracle, GCAT (proposed), PCA, and both LMM association tests. The results are for the simulated data sets shown in Figures 2. The quantitative traits are based on model (1). The variance contributions to the trait are genetic=5%, environmental=5%, and noise=90%.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

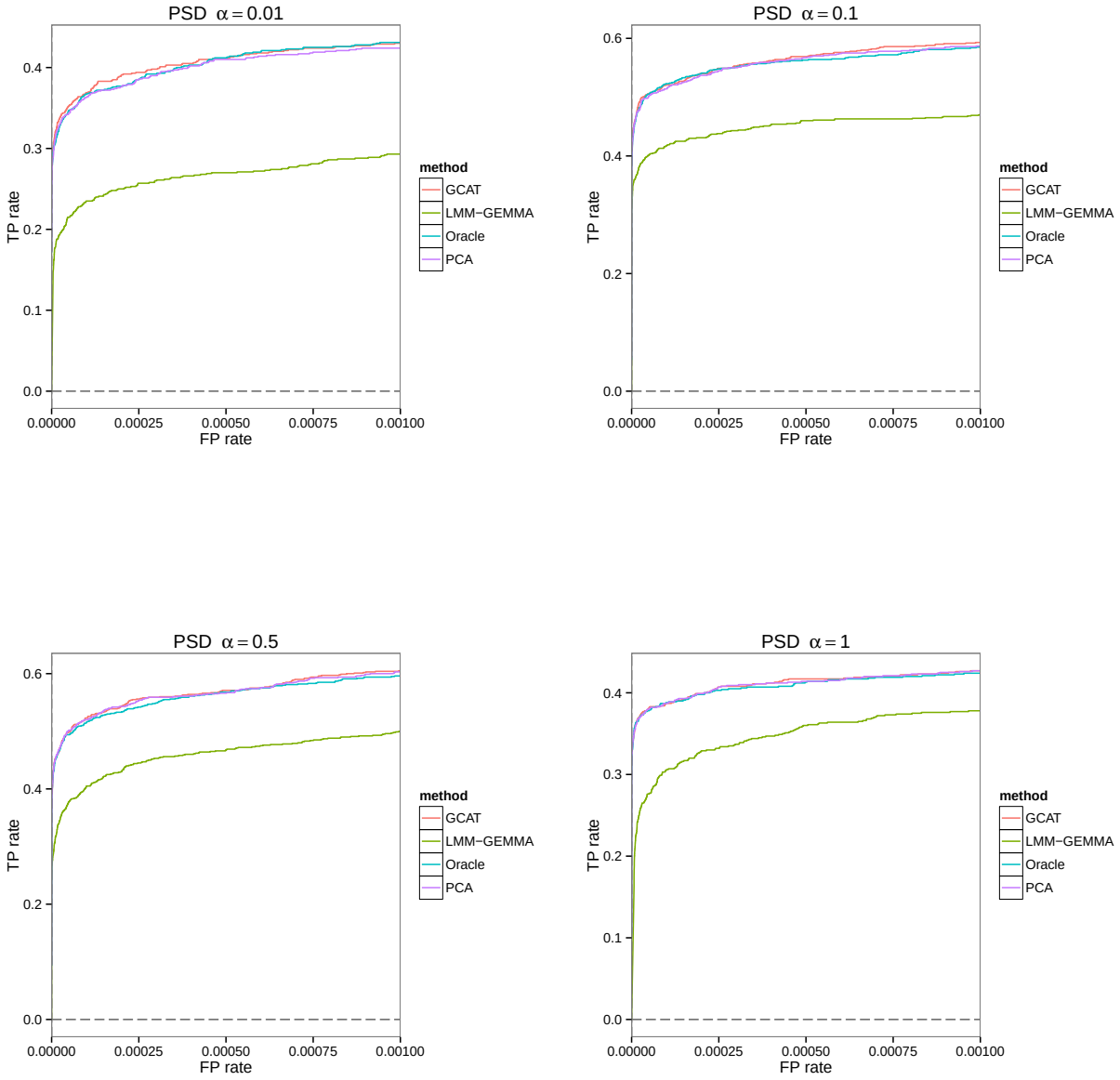

Figure S10: Power analysis for the simulation studies presented in Figure S1.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

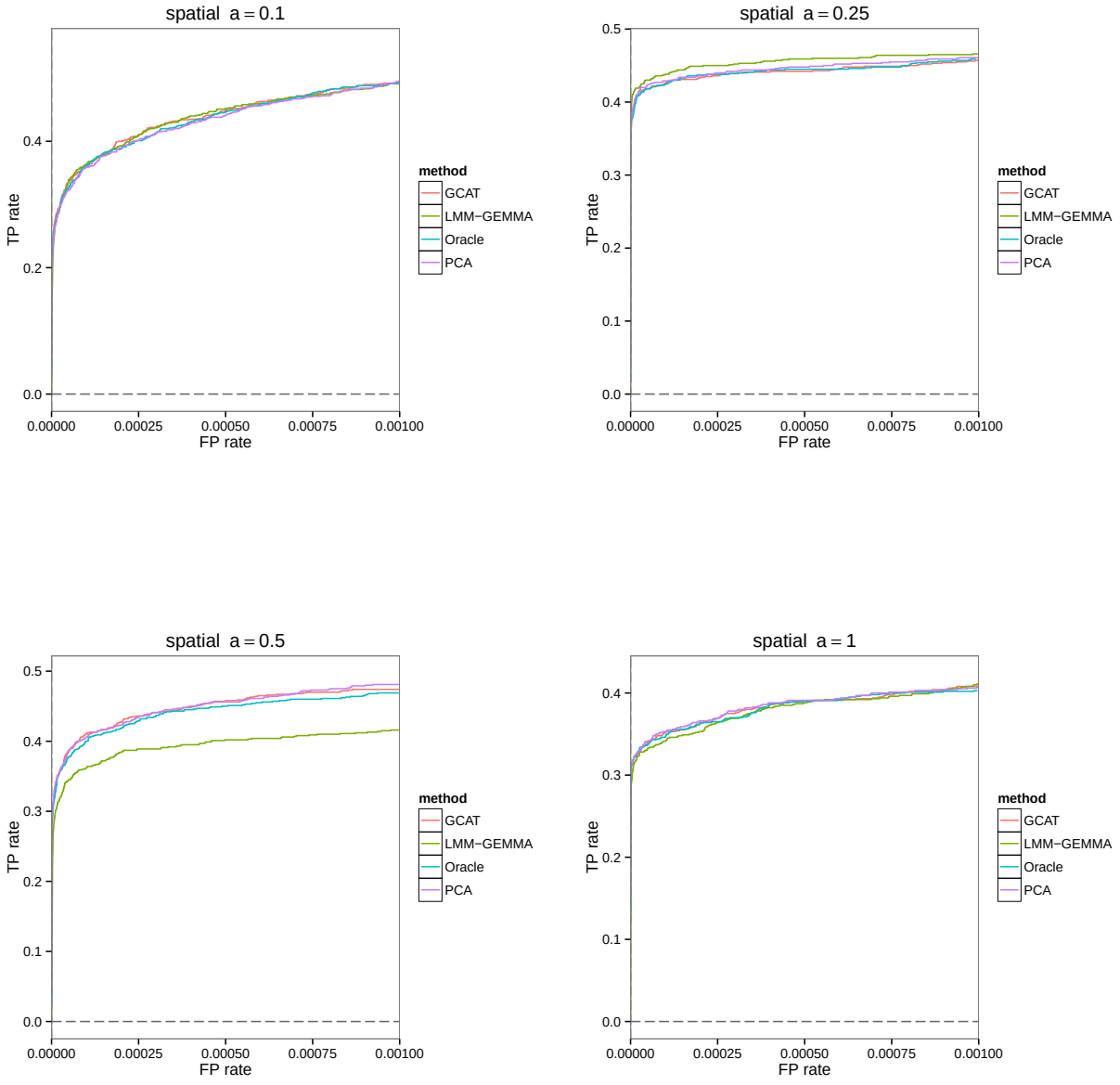

Figure S11: Power analysis for the simulation studies presented in Figure S2.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

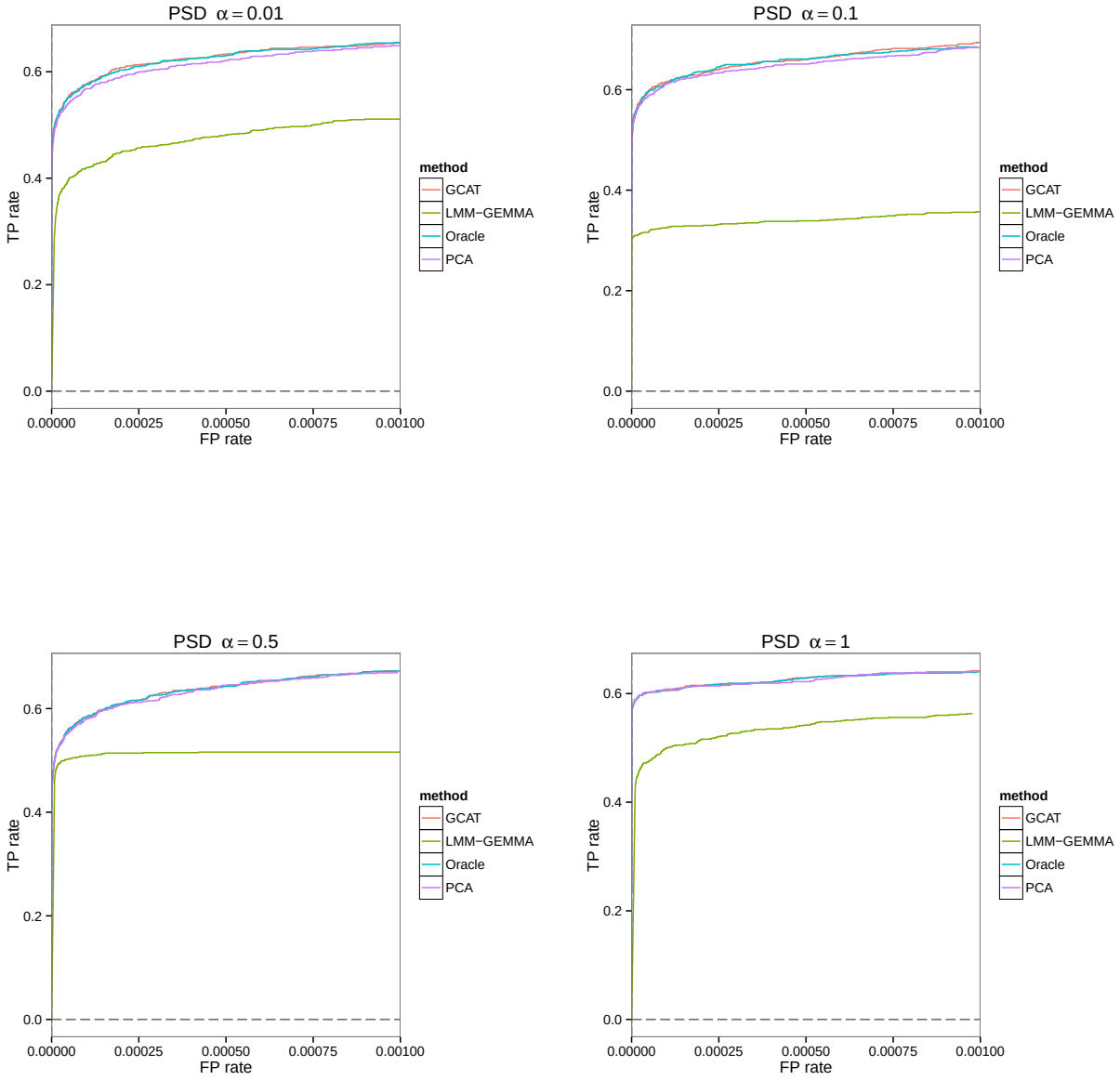

Figure S12: Power analysis for the simulation studies presented in Figure S3.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

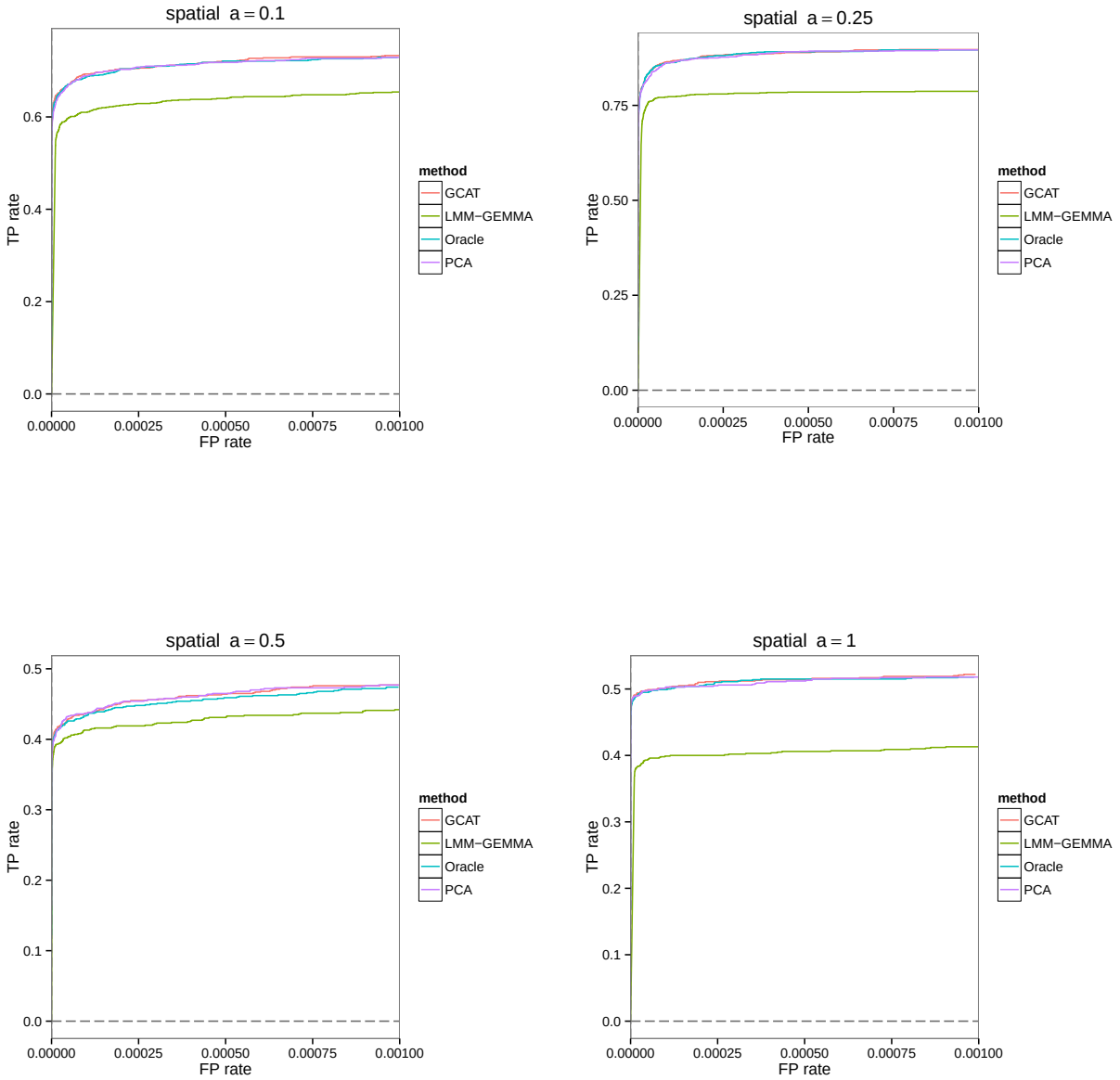

Figure S13: Power analysis for the simulation studies presented in Figure S4.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

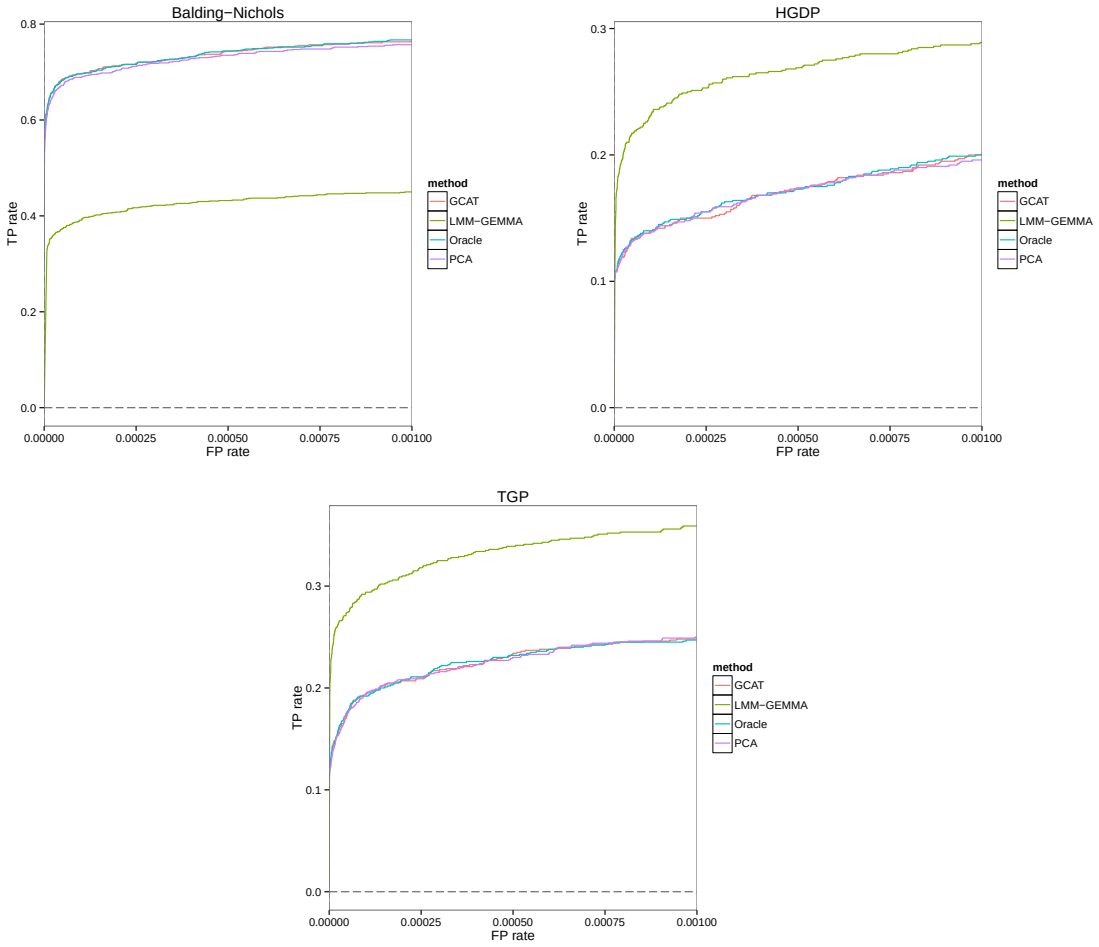

Figure S14: Power analysis for the simulation studies presented in Figure S5.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

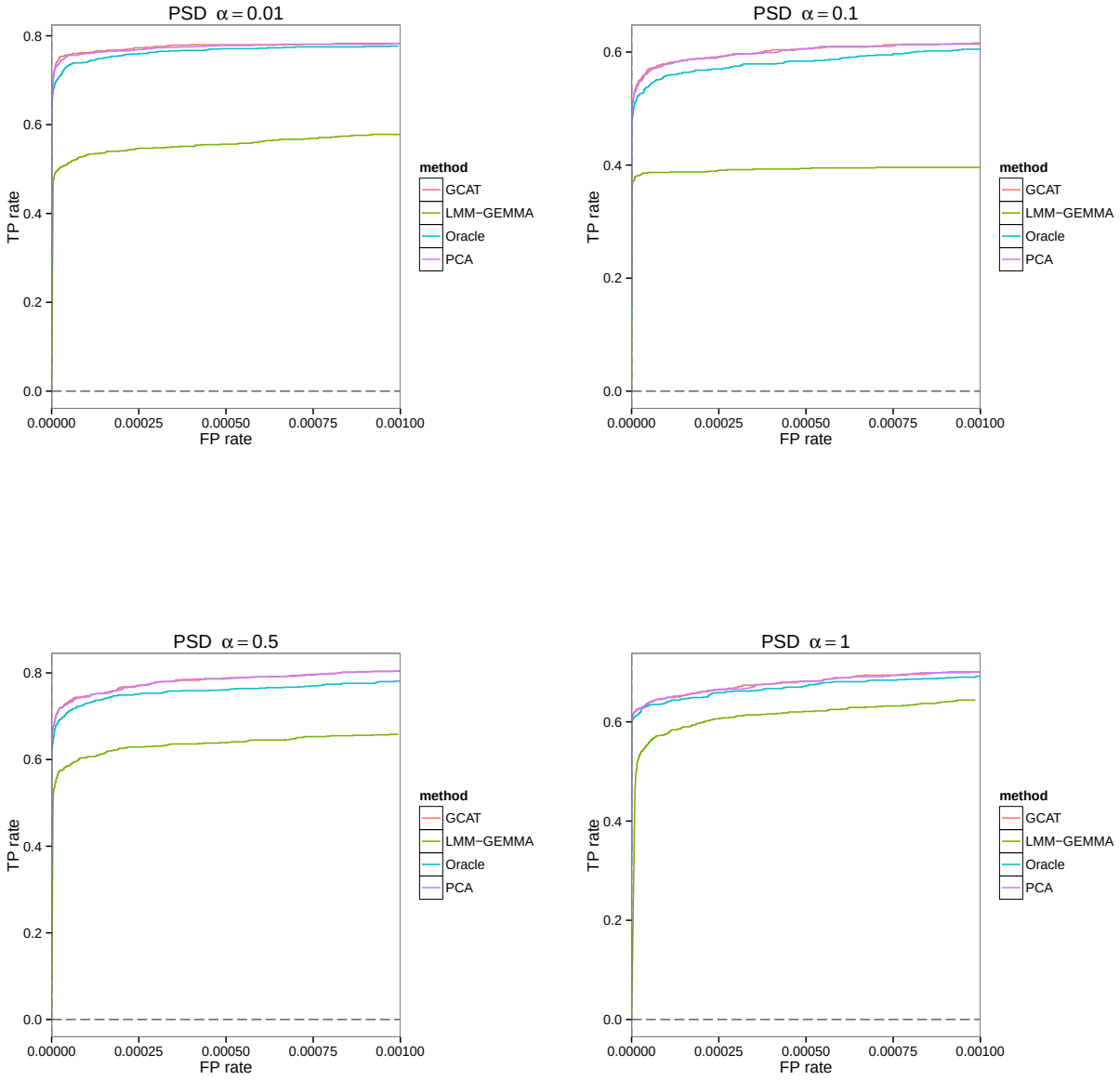

Figure S15: Power analysis for the simulation studies presented in Figure S6.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

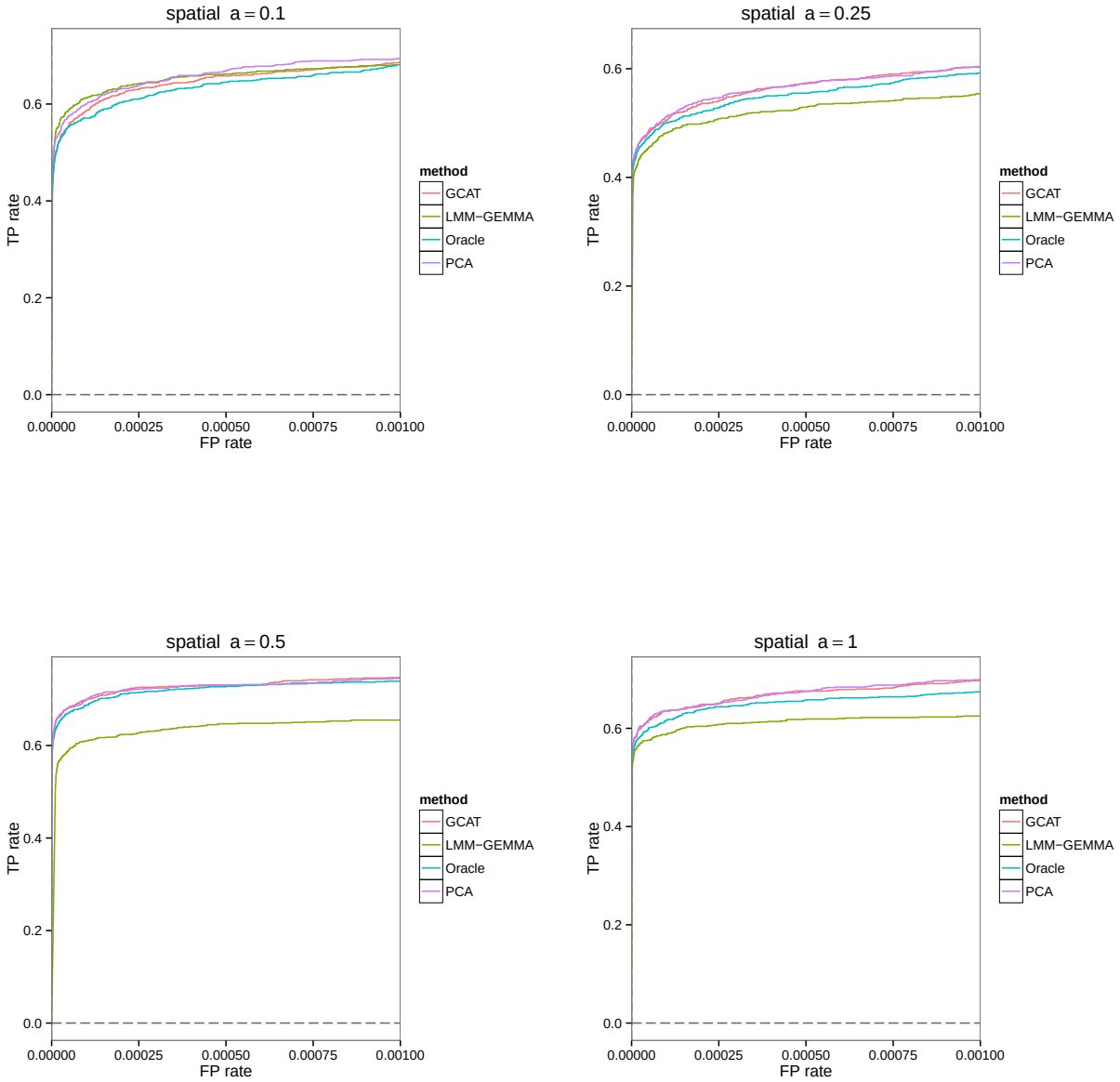

Figure S16: Power analysis for the simulation studies presented in Figure S7.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

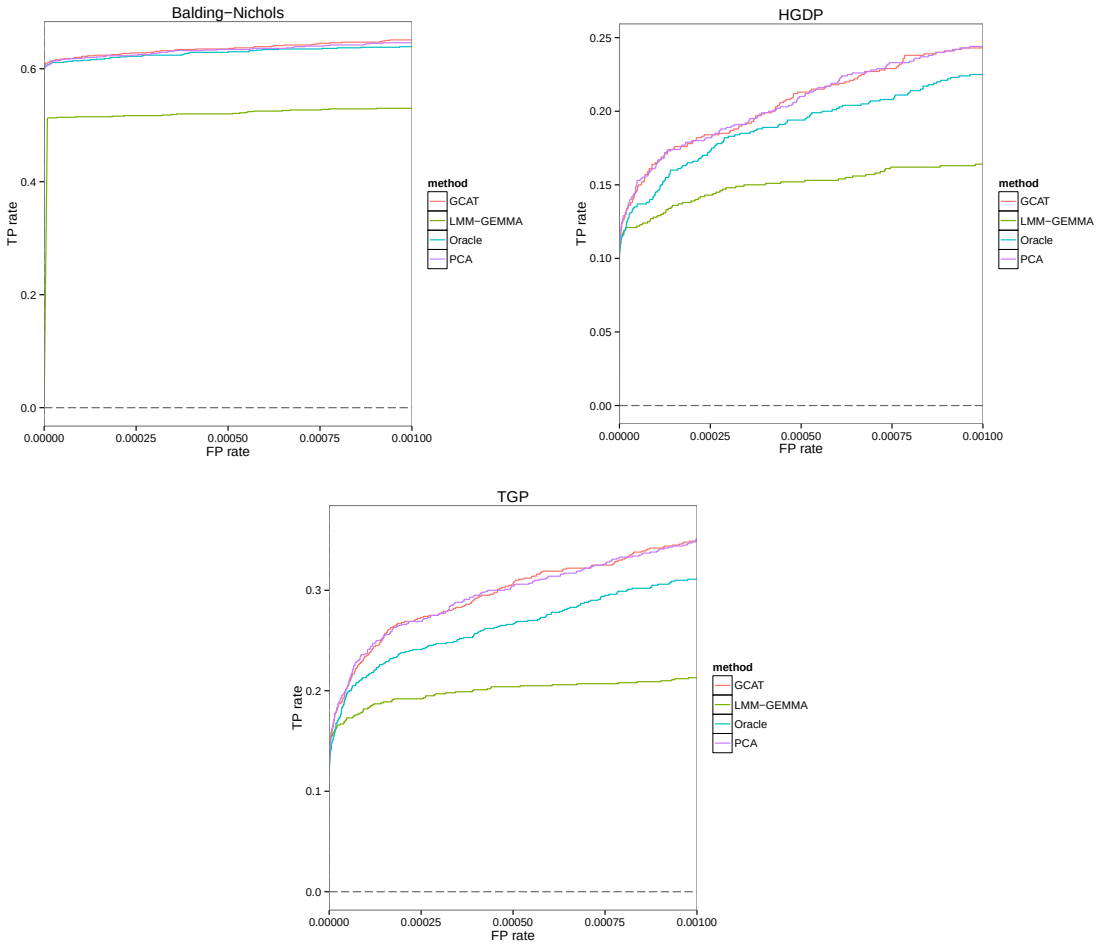

Figure S17: Power analysis for the simulation studies presented in Figure S8.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

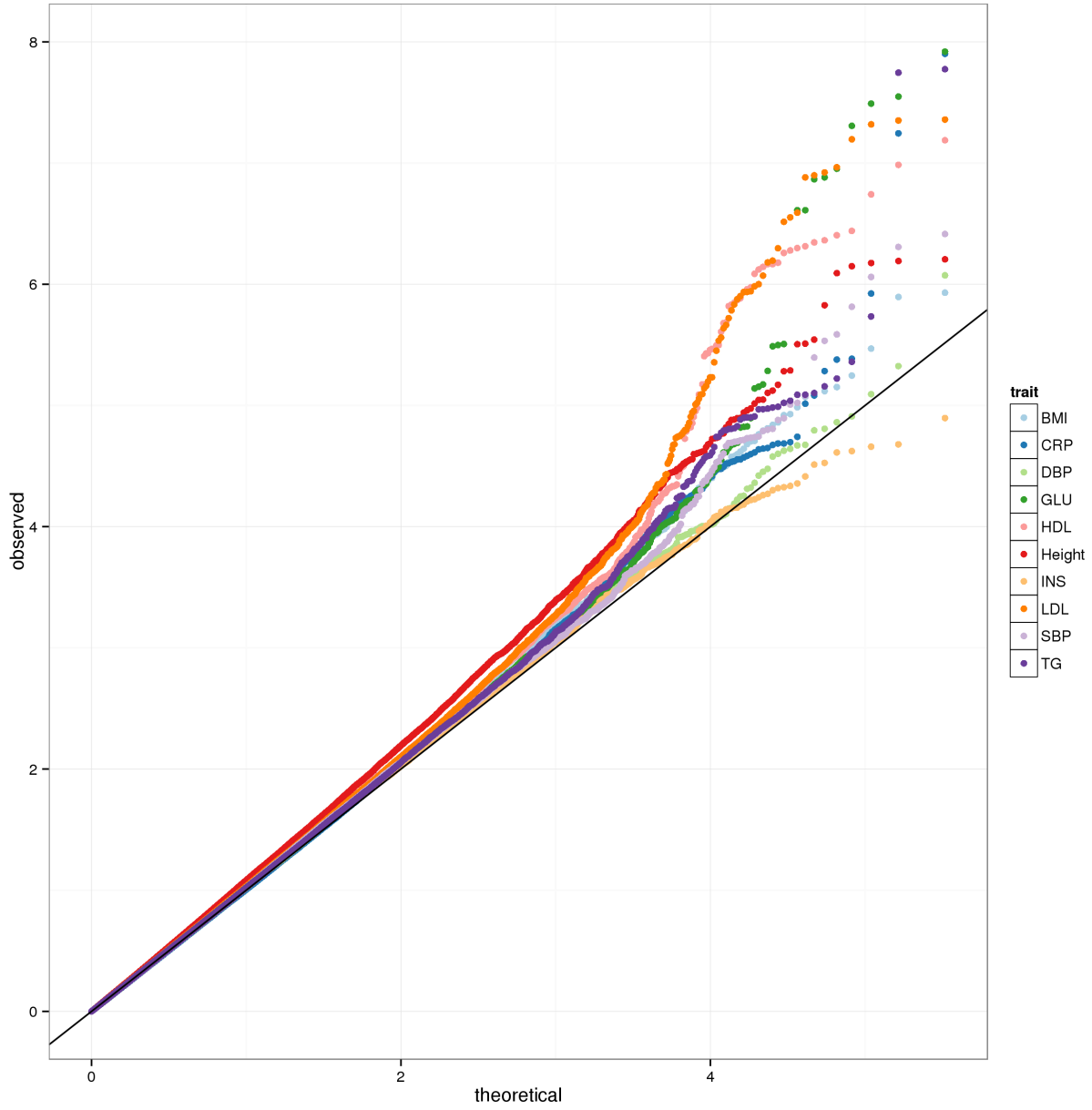

Figure S18: Theoretical versus observed quantiles of $-\log_{10}$(p-value) from the GCAT association tests on the Northern Finland Birth Cohort traits. The y-axis was truncated at p-value $< 10^{-8}$; see Table S1 for the smallest p-values for each trait.

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S1: 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) is 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

DBT

	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

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S1 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

bioRxiv preprint doi: https://doi.org/10.1101/012682; this version posted December 12, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S2: 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