Generalization of Rare Variant Association Tests for Longitudinal Family Studies

Li-Chu Chien,[1] Yen-Feng Chiu,[1*] Fang-Chi Hsu,[2] and Donald W. Bowden[3,4,5]

[1]Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Miaoli 35053, Taiwan, ROC.

[2]Department of Biostatistical Sciences, [3]Center for Diabetes Research, [4]Center for Genomics and Personalized Medicine Research, [5]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA.

*Corresponding author: Yen-Feng Chiu, yfchiu@nhri.org.tw

ABSTRACT:

Given the functional relevance of many rare variants, their identification is frequently critical for dissecting disease etiology. Functional variants are likely to be aggregated in family studies enriched with affected members, and this aggregation increases the statistical power to detect rare variants. Longitudinal family studies provide additional information for identifying genetic and environmental factors associated with disease over time. However, methods to analyze rare variants in longitudinal family data remain fairly limited. These methods should be capable of accounting for different sources of correlations and handling large amounts of sequencing data efficiently. To identify rare variants in longitudinal family studies, we extended pedigree-based burden and kernel association tests to genetic longitudinal studies. Generalized estimating equation (GEE) approaches were used to generalize the pedigree-based burden and kernel tests to multiple correlated phenotypes under the generalized linear model framework, adjusting for fixed effects of confounding factors. These tests accounted for complex correlations between repeated measures of the same phenotype (serial correlations) and between individuals in the same family (familial correlations). We conducted comprehensive simulation studies to compare the proposed tests with mixed-effects models and marginal models, using GEEs under various configurations. When the proposed tests were applied to data from the Diabetes Heart Study, we found exome variants of *POMGNT1* and *JAK1* genes were associated with type 2 diabetes.

KEY WORDS: rare variant association test; longitudinal family study; generalized estimating equations; burden test; kernel statistic

Introduction

Because of recent advances in high-throughput sequencing technologies and large studies of many complex traits, great progress has been made in elucidating the genetic basis of complex traits [Mardis, 2008; Metzker, 2010]. However, identifying a variant that influences disease susceptibility can aid in dissecting the disease's etiology only if the variant is causal and functionally relevant. Because the functional effects of rare variants arise mostly from amino acid changes, rare variants are likely to be missense mutations. Thus, very often, any functional effect is due to the rare variant itself [Bodmer and Bonilla, 2008]. Functional variants are likely to be aggregated in family studies enriched with affected members, and such aggregation increases the ability to detect rare variants.

Compared to cross-sectional studies, longitudinal family-based designs provide more information about the genetic and environmental factors associated with traits of interest [Burton et al., 2005]. Longitudinal genetic studies are increasingly being implemented to obtain additional information and to increase statistical power [Smith et al., 2010; Das et al., 2011; Fan et al., 2012; Furlotte et al., 2012; Wu et al., 2014]. Nevertheless, the analysis of data from longitudinal family studies is challenging, due to the need to account for within-family correlations and the computational load resulting from large sequencing datasets. Current statistical methods accounting for correlations within and between subjects and families, such as

generalized estimating equations (GEEs) and mixed models, may not be scalable to whole-genome sequencing (WGS) data [Wu et al., 2014]. Therefore, rare variant association tests, particularly for longitudinal family studies, remain fairly limited despite a pressing need for such approaches.

For population-based designs, many rare variant tests have been developed, which can be classified either into burden tests [Morris and Zeggini, 2010] or variance-component (kernel statistics) tests [Wu et al., 2011]. Burden tests combine information for multiple genetic variants into a single genetic score [Morgenthaler and Thilly, 2007; Li and Leal, 2008; Madsen and Browning, 2009; Morris and Zeggini, 2010; Zawistowski et al., 2010; Asimit et al., 2012] and test for the association between this score and a trait. These methods make a strong assumption that all rare variants in a set are causal. Violation of these assumptions can induce a substantial loss of power [Basu and Pan, 2011; Neale et al., 2011; Lee et al., 2012b]. Another class of methods uses variance-component tests under random-effects models. Instead of aggregating variants, kernel statistics evaluate the distribution of the aggregated score test statistics (probably with weights) of individual variants.

The burden test is simple and powerful when many variants are causal and their effects are in the same direction. The kernel test is powerful in the presence of both protective and risk variants or a small fraction of causal variants, but is less powerful than burden tests when most

variants are causal and the effects are in the same direction. Omnibus tests that combine burden and kernel tests have been proposed. However, although these tests reach robust power by combining two tests, they can be less powerful than either test alone if one of the assumptions of these two tests is true [Schaid et al., 2013; Jiang and McPeek, 2014; Lee et al., 2012a; Lee et al., 2014].

Compared to tests for population-based designs, relatively few rare variant tests have been developed for family-based designs. For example, extending the family-based association test [Laird et al., 2000] to sequence data was recently proposed, based on the population-based burden test [De et al., 2013]. Other authors have extended the kernel association test to quantitative traits of pedigree data by assuming random effects among family members [Schifano et al., 2012; Chen et al., 2013]. By combining variation from random effects with the residual error variation, they constructed a null variance matrix accounting for correlations induced by relationships among family members. However, one important assumption was that the pedigrees were randomly ascertained.

Ionita-Laza et al. [2013] developed a family-based association test for kernel statistics by specifying the distribution of offspring genotypes, conditioned on their phenotypes and parental genotypes (or the sufficient statistic in the absence of parental genotypes). However, the conditioning process and ignorance of between-family information caused a tremendous loss of

statistical power [Schaid et al., 2013]. Therefore, to improve the statistical power of their burden and kernel tests, these authors proposed to include parental data and to account for correlations between families and markers. They treated the genotype data as random given phenotypes, thereby overcoming the challenging problem of ascertainment for highly enriched pedigrees.

Although conditioning on traits in a retrospective likelihood manner tends to be less efficient than treating traits as random variables in a prospective likelihood manner, Kraft and Thomas [2000] found that doing so caused a small loss in efficiency for binary traits. In their simulation study, their proposed kernel test had more power than the proposed burden test, even in situations that seemed to favor the burden test. The great improvement in power, simplicity, and capability of these tests to compute large-scale data suggest that these tests could be useful in more extensive study designs. Given their advantages, it is worthwhile to generalize these tests to multiple phenotypes (different correlated phenotypes or the same phenotype, over time) under generalized linear models (GLMs) for pedigree data.

Material and Methods

Assume that N families are sequenced in a region with p genotyped rare variants. For the r^{th} repeated observation of the j^{th} individual of family i, let Y_{ijr} denote a phenotype variable, where $\boldsymbol{g}_{ij} = (g_{ij1}, g_{ij2}, \ldots, g_{ijp})^T$ are the genotypes (genotype scores) for the p variants ($g_{ijl} = 0, 1,$

or 2 for 0, 1, or 2 copies of the minor allele, respectively), and $\boldsymbol{x}_{ijr} = (x_{ijr1}, x_{ijr2}, \ldots, x_{ijrq})^T$ are the covariates for which we would like to adjust. For simplification of the notation, let $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijR_{ij}})^T$ be the $R_{ij} \times 1$ vector of response variables, $\boldsymbol{x}_{ij} = (\boldsymbol{x}_{ij1}, \boldsymbol{x}_{ij2}, \ldots, \boldsymbol{x}_{ijR_{ij}})^T$ be the $R_{ij} \times q$ matrix of covariate values. Let $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}^T, \boldsymbol{Y}_{i2}^T, \ldots, \boldsymbol{Y}_{in_i}^T)^T$ be the $(\sum_{j=1}^{n_i} R_{ij}) \times 1$ vector of response variables, $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T, \ldots, \boldsymbol{x}_{in_i}^T)^T$ be the $(\sum_{j=1}^{n_i} R_{ij}) \times q$ matrix of covariate values, and $\boldsymbol{g}_i = (\boldsymbol{g}_{i1}, \boldsymbol{g}_{i2}, \ldots, \boldsymbol{g}_{in_i})^T$ be the $n_i \times p$ matrix of genotype scores for the i^{th} family, where $i = 1, 2, \ldots, K$.

To relate genotypes to continuous/categorical phenotypes, we use GLMs [Liang and Zeger, 1986; Lee et al., 2012b], such that each component of $\boldsymbol{Y}_i$ follows an exponential family distribution

$$f(y_{ijr}) = \exp\left[\left\{y_{ijr}\theta_{ijr} - a(\theta_{ijr}) + b(y_{ijr})\right\}\phi\right] \quad (1)$$

with two moments, $\mu_{ijr} = E(Y_{ijr}) = \partial a(\theta_{ijr})/\partial\theta_{ijr}$ and $\text{Var}(Y_{ijr}) = (\partial\mu_{ijr}/\partial\theta_{ijr})/\phi$, and a link function, $h(\mu_{ijr}) = \eta_{ijr}$, where the linear predictor η_{ijr} is given by

$$\eta_{ijr} = \boldsymbol{x}_{ijr}^T\boldsymbol{\alpha} + \boldsymbol{g}_{ij}^T\boldsymbol{\beta}. \quad (2)$$

Here $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_q)^T$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ are the vectors of regression coefficients for the covariates and rare variants, respectively.

To derive the kernel and the burden statistics for longitudinal family data, we took a retrospective view of sampling and treated genotypes as random variables, following the

approach of Schaid et al. [Schaid et al., 2013]. Under the null hypothesis of no association of genotypes with traits, the expectation of the random variables for genotypes $\boldsymbol{G}_i = (\boldsymbol{G}_{i1}, \boldsymbol{G}_{i2}, \ldots, \boldsymbol{G}_{in_i})^T$ has elements $E_0(G_{ijl}) = 2m_l$, where m_l is the minor allele frequency (MAF) of the l^{th} marker. The random variables G_{ijl} have corresponding realized values g_{ijl}, as used in Equation 2. The elements of the null covariance of genotypes $\boldsymbol{G}_i$, $\text{Cov}_0(G_{ijl}, G_{ij'l'})$, are the covariance between the genetic markers l and l' of individuals j and j'. Specifically, the elements $\text{Cov}_0(G_{ijl}, G_{ij'l'})$ describe how individuals within the i^{th} family are related to each other and how the genetic markers are correlated within individuals in the i^{th} family due to linkage disequilibrium (LD).

Let $\boldsymbol{H}$ denote a $p \times p$ correlation matrix of genotype scores, with component $H_{ll'}$ for markers l and l', and let Ω_i denote a $n_i \times n_i$ matrix of genetic correlations for all n_i individuals in the i^{th} family. For autosomes, the elements of Ω_i are twice the kinship coefficients between individuals in the i^{th} family, namely, $\Omega_{ij,ij'} = 2\phi_{ij,ij'}$, where $\phi_{ij,ij'}$ represents the kinship coefficients between individuals j and j' in the i^{th} family. Similar to the approach of Schaid et al. [2013], we define the diagonal elements of Ω_i to be $\Omega_{ij,ij} = 1$ for outbred families and $\Omega_{ij,ij} = 1 + \varphi_{ij}$ for inbred families, where φ_{ij} represents the inbreeding coefficient of the j^{th} individual in the i^{th} family. Elements of the genotype codes in matrix $\boldsymbol{G}_i$ for individuals j and j', and markers l and l' can be expressed as

$$\mathrm{Cov}_0(G_{ijl}, G_{ij'l'}) = 2H_{ll'}\sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})}\Omega_{ij,ij'}.$$

Note that $H_{ll'}$ can be estimated empirically.

Kernel Statistic for Longitudinal Family Data

Let $\tilde{\boldsymbol{G}}_i = (\tilde{\boldsymbol{G}}_{i1}, \tilde{\boldsymbol{G}}_{i2}, \ldots, \tilde{\boldsymbol{G}}_{ip})$ be the $(\sum_{j=1}^{n_i} R_{ij}) \times p$ matrix of genotype scores, where $\tilde{\boldsymbol{G}}_{il} = (G_{i1l} \times \boldsymbol{1}^T_{R_{i1}}, G_{i2l} \times \boldsymbol{1}^T_{R_{i2}}, \ldots, G_{in_il} \times \boldsymbol{1}^T_{R_{in_i}})^T$ is a $(\sum_{j=1}^{n_i} R_{ij}) \times 1$ vector and $\boldsymbol{1}_{R_{ij}}$, $j = 1, 2, \ldots, n_i$, is a column R_{ij}-vector of ones. The quadratic kernel association statistic (KS) without weighting is

$$\kappa^{LF} = \sum_{l=1}^{p}\left[\sum_{i=1}^{K} \tilde{\boldsymbol{G}}_{il}^T \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)\right]^2.$$

Here, $\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i$ is the vector of residuals after adjusting for covariates, $\hat{\boldsymbol{V}}_i$ is the covariance matrix estimate of $\boldsymbol{Y}_i$, $\hat{\boldsymbol{V}}_i^{-1}$ is the inverse matrix of $\hat{\boldsymbol{V}}_i$, and $\hat{\Delta}_i$ and $\hat{\boldsymbol{A}}_i$ are estimates of the two $(\sum_{j=1}^{n_i} R_{ij}) \times (\sum_{j=1}^{n_i} R_{ij})$ diagonal matrices Δ_i and $\boldsymbol{A}_i$, where $\Delta_i = \mathrm{diag}\{\partial\theta_{ijr}/\partial\eta_{ijr}\}$ and $\boldsymbol{A}_i = \mathrm{diag}\{\partial\mu_{ijr}/\partial\theta_{ijr}\}$ for $j = 1, 2, \ldots, n_i$, $r = 1, 2, \ldots, R_{ij}$. All estimates were obtained by the generalized estimating equation method (GEEM) with the linear predictor $\eta_{ijr} = \boldsymbol{x}_{ijr}^T\boldsymbol{\alpha}$ [Liang and Zeger, 1986].

By assuming a weighted linear kernel and letting $(w_1, w_2, \ldots, w_p)^T$ be a $p \times 1$ vector with weights for each marker, the KS can be expressed as

$$\kappa^{LF} = \sum_{l=1}^{p}\left[w_l \sum_{i=1}^{K} \tilde{\boldsymbol{G}}_{il}^T \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)\right]^2 = \left(\boldsymbol{Z}^{LF}\right)^T \boldsymbol{Z}^{LF}, \qquad (3)$$

where $\boldsymbol{Z}^{LF} = (Z_1^{LF}, Z_2^{LF}, \ldots, Z_p^{LF})^T$, with $Z_l^{LF} = w_l \sum_{i=1}^{K} \tilde{\boldsymbol{G}}_{il}^T \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$ for $l = 1, 2, \ldots, p$.

Based on the theory of quadratic forms of normal variables, the null distribution of κ^{LF} can be approximated asymptotically by a mixture of χ^2 distributions (see Appendix A for technical details). As suggested by Schaid et al. [2013], we estimate the null distribution of κ^{LF} by a scaled χ^2 distribution with the scale and degrees of freedom estimated by the first two moments of κ^{LF}. That is, the scale is estimated as $\delta^{LF} = \mathrm{Var}_0(\kappa^{LF})/(2\,E_0(\kappa^{LF}))$ and the degrees of freedom as $d^{LF} = 2(E_0(\kappa^{LF}))^2/\mathrm{Var}_0(\kappa^{LF})$. The *p*-value can be computed by assuming the scaled kernel statistic $\kappa^{LF}_{\text{scaled}} = \kappa^{LF}/\delta^{LF}$, which has an asymptotic χ^2 distribution with d^{LF} degrees of freedom.

When independent working correlations between repeated measures from the same phenotypes (serial correlations) and between individuals within the same family (familial correlations) are assumed, the KS in Equation 3 (κ^{LF}) reduces to

$$\kappa^{LF_l} = \sum_{l=1}^{p}\left[w_l \sum_{i=1}^{K} \tilde{\boldsymbol{G}}_{il}^T \hat{\Delta}_i \left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i \right) \right]^2 = \left(\boldsymbol{Z}^{LF_l} \right)^T \boldsymbol{Z}^{LF_l}. \qquad (4)$$

With the canonical link between θ_{ijr} and η_{ijr}, the KS in Equation 4 (κ^{LF_l}) can be simplified as

$$\kappa^{LF_{lC}} = \sum_{l=1}^{p}\left[w_l \sum_{i=1}^{K} \tilde{\boldsymbol{G}}_{il}^T \left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i \right) \right]^2 = \left(\boldsymbol{Z}^{LF_{lC}} \right)^T \boldsymbol{Z}^{LF_{lC}},$$

which is similar to the kernel statistic derived by Schaid et al. [2013]. Asymptotically, the null distributions of κ^{LF_l} and $\kappa^{LF_{lC}}$, similar to that of κ^{LF}, can be approximated by a mixture of χ^2 distributions.

Burden Test for Longitudinal Family Data

For the i^{th} family, a weighted average of all genotype scores can be calculated by $\boldsymbol{S}_i^{LF} = \sum_{l=1}^{p} w_l \tilde{\boldsymbol{G}}_{il}$. Under the null hypothesis that genotype is not associated with traits, a burden test (BT) for longitudinal family data can be constructed as a mean-zero function of $\boldsymbol{y}_i$ and $\boldsymbol{S}_i^{LF}$, where $i = 1, 2, \ldots, K$. Assuming that there are covariates to adjust for, we let $\boldsymbol{L} = (\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots., \boldsymbol{L}_K)^T$ be the vector of residuals, after adjusting for covariates with $\boldsymbol{L}_i = \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$. Under the null hypothesis, the statistic for a BT for longitudinal family data is defined by

$$B^{LF} = \frac{\left[\boldsymbol{L}^T \boldsymbol{S}^{LF}\right]^2}{\mathrm{Var}(\boldsymbol{L}^T \boldsymbol{S}^{LF})} = \frac{\left[\sum_{i=1}^{K} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{A}}_i \hat{\Delta}_i \boldsymbol{S}_i^{LF}\right]^2}{\sum_{i=1}^{K} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{A}}_i \hat{\Delta}_i \mathrm{Cov}_0(\boldsymbol{S}_i^{LF}) \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)},$$

where the null covariance matrix of $\boldsymbol{S}_i^{LF}$ takes the form

$$\begin{aligned}
\mathrm{Cov}_0(\boldsymbol{S}_i^{LF}) &= \mathrm{Cov}_0\left(\sum_{l=1}^{p} w_l \tilde{\boldsymbol{G}}_{il}\right) \\
&= \sum_{l=1}^{p} w_l^2 \mathrm{Cov}_0(\tilde{\boldsymbol{G}}_{il}, \tilde{\boldsymbol{G}}_{il}) + 2\sum_{l=1}^{p} \sum_{l'=l+1}^{p} w_l w_{l'} \mathrm{Cov}_0(\tilde{\boldsymbol{G}}_{il}, \tilde{\boldsymbol{G}}_{il'}) \\
&= \tilde{\Omega}_i \sum_{l=1}^{p} \sum_{l'=1}^{p} 2 w_l w_{l'} H_{ll'} \sqrt{m_l(1-m_l) m_{l'}(1-m_{l'})},
\end{aligned}$$

where $\tilde{\Omega}_i$ is a $(\sum_{j=1}^{n_i} R_{ij}) \times (\sum_{j=1}^{n_i} R_{ij})$ matrix with elements $\tilde{\Omega}_{ijr,ijr'} = \Omega_{ij,ij}$ for any r and r'. Hence,

$$B^{LF} = \frac{\left[\sum_{l=1}^{p} w_l \sum_{i=1}^{K} \tilde{\boldsymbol{G}}_{il}^T \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)\right]^2}{2\sum_{l=1}^{p} \sum_{l'=1}^{p} w_l w_{l'} H_{ll'} \sqrt{m_l(1-m_l) m_{l'}(1-m_{l'})} C^{LF}}, \qquad (5)$$

where $C^{LF} = \sum_{i=1}^{K} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{A}}_i \hat{\Delta}_i \tilde{\Omega}_i \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$. For large samples, B^{LF} approximately

follows a χ^2 distribution with one degree of freedom.

In the special case with an independent working correlation assumption for serial and familial correlations, the BT in Equation 5 (B^{LF}) reduces to

$$B^{LF_I} = \frac{\left[\sum_{l=1}^{p} w_l \sum_{i=1}^{K} \tilde{\boldsymbol{G}}_{il}^T \hat{\Delta}_i \left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i\right)\right]^2}{2\sum_{l=1}^{p}\sum_{l'=1}^{p} w_l w_{l'} H_{ll'} \sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})} C^{LF_I}}, \quad (6)$$

where $C^{LF_I} = \sum_{i=1}^{K} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\Delta}_i \tilde{\Omega}_i \hat{\Delta}_i (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$ .With an independent working structure and a canonical link between θ_{ijr} and η_{ijr} , BT in Equation 6 (B^{LF_I}) can be further simplified to

$$B^{LF_{IC}} = \frac{\left[\sum_{l=1}^{p} w_l \sum_{i=1}^{K} \tilde{\boldsymbol{G}}_{il}^T \left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i\right)\right]^2}{2\sum_{l=1}^{p}\sum_{l'=1}^{p} w_l w_{l'} H_{ll'} \sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})} C^{LF_{IC}}},$$

where $C^{LF_{IC}} = \sum_{i=1}^{K} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \tilde{\Omega}_i (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$. The form of $B^{LF_{IC}}$ is similar to that of the burden test proposed by Schaid et al. [2013]. The asymptotic distributions of B^{LF_I} and $B^{LF_{IC}}$ are the same as that of B^{LF}.

To apply the test statistics κ^{LF} and B^{LF} to more general situations where the genetic relationships between individuals j and j' need to be estimated from genomic data, we suggest an empirical estimate for the genetic correlation matrix Ω_i for autosomes. As proposed by Thornton and McPeek [2010] and Schaid et al. [2013], we estimate the elements of the genetic correlation matrix Ω_i for autosomes by

$$\hat{\Omega}_{ij,ij'} = \frac{1}{p}\sum_{l=1}^{p}\frac{(g_{ijl}-2m_l)(g_{ij'l}-2m_l)}{2m_l(1-m_l)}.$$

The kernel and burden tests can also be applied to X chromosomes (see Appendix B for technical details).

Simulations

Type I error and power were compared between the two commonly used methods for longitudinal data analysis (GEEM and random-effects models [REMs]) and the two proposed tests in the simulation study. In the typical GEEM, each component of $\boldsymbol{Y}_i$ is assumed to have a distribution in the exponential family in Equation 1, with link function $h(\mu_{ijr}) = \boldsymbol{x}_{ijr}^T\boldsymbol{\alpha} + \beta_0\sum_{l=1}^{p} g_{ijl}$ [Diggle et al., 1994]. Associations between genetic variants and phenotypes can be assessed by testing $H_0 : \beta_0 = 0.$ In the REM, given the family-specific random effect $\boldsymbol{U}_i$, we assume that each component of $\boldsymbol{Y}_i$ follows a distribution from the exponential family, with density function $f(y_{ijr} \mid \boldsymbol{U}_i) = \exp\left[\left\{y_{ijr}\theta_{ijr} - a(\theta_{ijr}) + b(y_{ijr})\right\}\phi\right]$ [Diggle et al., 1994]. The conditional moments, $\mu_{ijr} = E(Y_{ijr} \mid \boldsymbol{U}_i) = \partial a(\theta_{ijr})/\partial\theta_{ijr}$ and $\mathrm{Var}(Y_{ijr} \mid \boldsymbol{U}_i) = (\partial\mu_{ijr}/\partial\theta_{ijr})/\phi$, satisfy $h(\mu_{ijr}) = \boldsymbol{x}_{ijr}^T\boldsymbol{\gamma} + \beta_0^*\sum_{l=1}^{p} g_{ijl} + \boldsymbol{d}_{ijr}^T\boldsymbol{U}_i$, where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)^T$ and β_0^* are the regression coefficients for the covariates and genetic variants, respectively, and $\boldsymbol{d}_{ijr} = (\boldsymbol{x}_{ijr}^T, \sum_{l=1}^{p} g_{ijl})^T$. Similar to the GEEM, we use the REM to test $H_0 : \beta_0^* = 0$ and to examine whether the genetic variants

are associated with the phenotype. We conducted GEEM and REM using the R computing packages gee [Carey et al., 2015] and leme4 [Bate, 2010], respectively.

Because LDs among rare variants tend to be small, we first simulated independent genetic variants [Pritchard, 2001; Turkmen and Lin, 2014]. Genetic variants were generated to form two populations. In the first population, the mutation frequency for genetic variant l was generated from Wright's distribution. The frequency distribution for each variant is $f(m) = c_w m^{B_1-1}(1-m)^{B_2-1}e^{\sigma_w(1-m)}$, where $B_1 = 0.001$, $B_2 = B_1/3$, and $\sigma_w = 12$ are the same as those in Madsen and Browning [2009]. The parameter c_w is a constant satisfying the condition $\int_0^1 f(m)dm = 1$. In the second population, the mutation frequency for genetic variant l was generated from the Balding-Nichols model [1995] $BN((1-f_{BN}/f_{BN})p_{BN},\ (1-f_{BN}/f_{BN})(1-p_{BN}))$, with parameters f_{BN} and p_{BN} given by $f_{BN} = 0.0033$ and $p_{BN} = 0.0084$, respectively, which ensures that the expected mutation frequency for variant l is 0.0084.

The homogeneous population is based on the first population only, while the heterogeneous population comprises half of the first population and half of the second population. The number of children per family follows a Poisson distribution with a mean of 2 and a range of 1–4. Parental haplotypes were generated based on the frequencies of the variants, assuming that the variants are independent. Once the parental haplotypes were generated, random transmission was assumed to generate the offspring haplotypes.

In the heterogeneous population, binary and continuous phenotypes for individual j in the i^{th} family were generated through the following logit and linear models, respectively:

$$\text{logit}(P(Y_{ijr}=1)) = \alpha_0 + \log(1.5)x_{ijr1} + \log(1.5)x_{ijr2} + \beta_1 g_{ij1} + \cdots + \beta_s g_{ijs}, \quad (7)$$

$$y_{ijr} = 0.01 + \log(1.5)x_{ijr1} + \log(1.5)x_{ijr2} + \beta_1 g_{ij1} + \cdots + \beta_s g_{ijs} + \varepsilon_{ijr},$$

where $j = 1,2,\ldots,n_i$ and $r = 1,2,\ldots,5$. The intercept α_0 is logit(0.01) = -4.5951. Continuous covariates x_{ij11} for individuals $j = 1,2,\ldots,n_i$ (j =1 and 2 for parents) were generated from a multivariate normal distribution with a mean of 0.5 and covariance matrix $\Sigma_x = (\sigma_{xjj'})$, where $\sigma_{xjj'} = 1$ if $j = j'$; $\sigma_{xjj'} = 0$ if $j = 1$ and $j' = 2$; and $\sigma_{xjj'} = 0.5$, otherwise. Covariates x_{ijr1} are identical for $r = 1,2,\ldots,5$. Covariate x_{ijr2} is an indicator variable equal to 0 or 1 for individuals from the first and second populations, respectively, of a heterogeneous population, and equal to zero and not included for a homogeneous population.

Binary phenotypes for each individual were determined by Equation 7. The serial correlation structure for repeated phenotypes is $\Sigma_y = (\sigma_{yrr'})$ with $\sigma_{yrr'} = \rho^{|r-r'|}$ for $r = 1,2,\ldots,5$ and $r' = 1,2,\ldots,5$, where ρ is 1 or 0.5. For a dichotomous trait, each family has at least one affected child at the first visit as a proband. For a continuous trait, the error terms $(\varepsilon_{ij1}, \varepsilon_{ij2}, \ldots, \varepsilon_{ij5})^T$ follow a multivariate normal distribution with a mean of zero and covariance matrix $\Sigma_\varepsilon = (\sigma_{\varepsilon rr'})$ given by $\sigma_{\varepsilon rr'} = \rho^{|r-r'|}$ for $r = 1,2,\ldots,5$ and $r' = 1,2,\ldots,5$, where ρ is the same as for the dichotomous trait.

In addition, we generated rare variants with linkage disequilibrium (LD) structure to evaluate the performance of the proposed tests. Following the approach of Schaid et al. [2013], we simulated a latent vector $\boldsymbol{Q} = (Q_1, Q_2, \ldots, Q_{200})^T$ from a multivariate normal distribution with a first-order auto-regressive covariance structure. The correlation between any two latent components is $\mathrm{Cor}(Q_l, Q_{l'}) = \rho_{LD}^{|l-l'|}$, where ρ_{LD} is either 0.9 or 0.5. The latent vectors were dichotomized to yield a haplotype with MAFs randomly generated from Wright's distribution. We then combined two independent haplotypes and randomly selected 100 markers with MAF<0.01 to obtain genotype data. The binary and continuous phenotypes for individuals were further generated based on Equation 7.

Under the null hypothesis, regression parameters for genetic variants in Equation 7 are equal to zero. Under the alternative case, 50 or 100 genetic variants were randomly selected, and the numbers of the risk/protective variants are 5/0, 5/5, 15/0, and 15/15 variants. Similar to Lee et al. [2012b], β_l is $0.6\times|\log_{10}(m_l)|$ or $0.6\times\log_{10}(m_l)$ with respect to the risk or protective variant l, $l = 1, 2, \ldots, s$. One thousand replicates were conducted to compare the type I error and power of the two proposed unweighted tests (KS and BT) versus the existing GEEM and REM in the simulations.

Results

For scenarios without LD between markers, type I errors for the KS, BT, GEEM, and REM for dichotomous and continuous traits in the homogeneous population were examined for different numbers of variants (p) and families (K), serial (ρ) and working correlations, and nominal levels (α) (Table 1 and Table S1). Type I errors for the KS, BT, and GEEM were close to the nominal levels under different configurations, regardless of the specified working correlations. For the dichotomous trait, type I errors for the REM were seriously inflated (range: 0.57–0.98) when the serial correlation was 1. Generally, type I errors for the REM were slightly inflated under different configurations when the serial correlation was 0.5. The four approaches showed similar performance in testing for the X chromosome (Table S2).

Because of the similar performance of methods with different sample sizes and nominal levels, we examined type I errors in the presence of heterogeneity for a family size of 450 at the nominal level of 0.05 (Table 2). Type I errors for the KS and GEEM were consistent with the nominal level under all configurations; however, the BT tended to have a deflated type I error in the presence of population stratification. The REM continued to have an inflated type I error when the serial correlation was 1 for a dichotomous trait. These results suggest that the KS and GEEM were robust to population stratification under the null hypothesis.

In the power study, the "independent" and "exchangeable" working correlations were assumed for the KS, BT, and GEEM under different serial correlations (ρ = 1 or 0.5).

Risk/protective variant combinations of 5/0, 5/5, 15/0, and 15/15 variants were tested, and 450 families with 100 or 50 variants at nominal levels of 0.05 or 0.01 were generated for each replicate. Figures 1–2 and Figures 3–4 show the power of the four approaches for dichotomous and continuous traits in the absence and presence of population stratification at a nominal level of 0.05. Figures S1–S2 depict the power under the same configurations without population stratification at a nominal level of 0.01.

Power generally increased with the number of effective variants, particularly when effects were in the same direction. The KS was fairly robust and steadily outperformed the other three tests under different configurations. The BT, GEEM, and REM sometimes achieved power similar to the KS in the presence of causal variants only, but the power of these tests decreased when the variants had mixed effects. Type I error in the REM was inflated for highly corrected dichotomous traits, regardless of the population structure. The power of the KS was always at the highest level for different numbers of variants for continuous traits; for dichotomous traits, it increased with increasing numbers of effective variants. In the absence of prior knowledge, the exchangeable working correlation matrix generally performed better than the independent matrix.

Relative performances of the four approaches were similar under different population structures, serial or working correlations, nominal levels, numbers of variants, and sample sizes

(see results in Figures S3–S6, assuming a sample size of 225). Power for X chromosome analyses is depicted in Figures S7–S14.

For scenarios with LDs between markers, type I errors for each approach for dichotomous and continuous traits in the homogeneous population were studied for numbers of variants (p), families (K), and serial correlation (ρ) given by 100, 450, and 0.5, respectively (Table 3). Table 3 also includes the results from the kernel and burden tests proposed by Schaid et al. [2013] (KSS and BTS) based on the last time point of the dichotomous traits. In addition, Table 3 shows results using famSKAT [Chen et al., 2013] based on the average of all measurements for the continuous traits. The KS, BT, and GEEM all have valid empirical type I error rates under different configurations regardless of the specified working correlations. Overall, the proposed methods are robust to the LD structure. Similar to the results without LD (Table 1), type I errors of REM were slightly inflated for the dichotomous traits when the serial correlation (ρ) is 0.5.

Power for the KS, BT, GEEM, KSS, BTS, and famSKAT with LDs between markers are shown in Table 4. For dichotomized traits, the proposed kernel and burden tests have higher power than the tests by Schaid et al. [2013], assuming an exchangeable working correlation matrix. Power in each scenario was compatible with Schaid et al.s' methods [2013] when assuming an independent working correlation matrix. For continuous traits, the power from the proposed kernel test was slightly lower than that achieved using the average of all measurements

from famSKAT; power was compatible when only the last point was used. However, the proposed kernel and burden are more powerful than famSKAT based on the last time point for scenarios, with a constantly high phenotypic variance over time (Table S4). Given individual phenotypic variances as high as 10, the type I error rates remained valid (Table S3), which suggests that the proposed methods are robust to LD structure.

Application to Real Data

We applied the six approaches (κ^{LF}, B^{LF}, GEEM, REM, KSS and BTS) for dichotomous traits to data from the family-based Diabetes Heart Study (DHS) conducted in Forsyth County, North Carolina. The DHS was designed to better understand the genetic and epidemiological origins of cardiovascular disease in families with type 2 diabetes mellitus (T2DM) [Bowden et al., 2008]. A total of 1,443 European American and African American participants from 564 families were recruited. The study design (including ascertainment and recruitment) and genotyping are reported elsewhere [Bowden et al., 2008]. To enable comparisons of cardiovascular-related phenotypes over time, preliminary data were collected. The final analysis included phenotype and covariate data from 99 families with 220 European American offspring, with 1 to 5 offspring per family. Every individual had two repeated observations at an average interval of 3.9 years. The phenotype of interest was T2DM status at

each time point.

Genotype data were obtained from the Illumina Infinium Human Exome Beadchip v1.0 (San Diego, CA). A total of 8,159 single nucleotide polymorphisms (SNPs) covering 1,683 different genes located on chromosome 1 at positions between 762,320 and 249,211,619 base pairs were studied. We restricted our analyses to genes with more than two SNPs, to obtain a total genotype score greater than zero, as required for constructing a genetic correlation matrix between SNPs.

A total of 7,094 SNPs from 1,036 genes on chromosome 1 located between 865,628 and 249,211,619 base pairs were included in the association analyses with T2DM. There were 5,417 SNPs (76%) with a minor allele frequency (MAF) less than 5% and 5,747 SNPs (81%) with an MAF less than 10%. We adjusted for potential confounding factors, including gender, age, and body mass index (BMI) at baseline.

Assuming an exchangeable working correlation structure for KS, BT, and GEEM and applying KSS and BTS based on the last time point, the analysis was conducted with basic covariate adjustment for age and gender, or with full covariate adjustment for age, gender, and BMI (Table 5). With the basic adjustment, the only gene that remained statistically significant after Bonferroni correction (p-value $< 0.05/1036 \approx 4.83 \times 10^{-5}$) was *POMGNT1*, with a p-value = 2.11×10^{-4}, 2.81×10^{-5}, 1.23×10^{-4}, 0.16, 9.29×10^{-4} and 4.57×10^{-5} for the KS, BT, GEEM, REM,

KSS and BTS, respectively. The *JAK1* (*p*-value = 4.58×10^{-5}, 1.46×10^{-4}, 1.26×10^{-3}, 2.01×10^{-2}, 7.38×10^{-5} and 2.14×10^{-4} for the KS, BT, GEEM, REM, KSS and BTS, respectively) and *POMGNT1* (*p*-value = 1.19×10^{-4}, 1.07×10^{-5}, 4.05×10^{-5}, 0.28, 5.54×10^{-4} and 2.75×10^{-5}, respectively) genes remained significant after full adjustment. These results suggest that the association between T2DM and *JAK1*, but not between T2DM and *POMGNT1*, is likely to be mediated by BMI. The SNPs and their MAF information for *POMGNT1* and *JAK1* are listed in Table S5.

After adjusting for common variants, the collapsed rare variants (MAF < 0.05) of *POMGNT1* still showed significant associations with T2DM (*p*-value = 8.96×10^{-3}, 9.72×10^{-3}, 3.80×10^{-3}, 2.36×10^{-3}, 8.73×10^{-3} and 6.04×10^{-3} for the KS, BT, GEEM, REM, KSS and BTS, respectively), but the collapsed rare variants of *JAK1* did not (*p*-value = 0.41, 0.44, 0.06, 0.15, 0.56 and 0.58, respectively) (Table S6).

To clarify the sources of associations for these two genes, we estimated effect sizes with respect to the collapsed rare variants and common variants using GEEM and REM (Table S7). Based on the GEEM, both common variants and the collapsed rare variants from *POMGNT1* were associated with T2DM in the same direction (*p*-value = 1.58×10^{-3} and 3.80×10^{-3}, respectively). The common variant of *JAK1* was positively associated with T2DM (*p*-value = 1.53×10^{-3}), whereas the collapsed rare variants tended to be negatively associated with T2DM

(p-value = 0.06 from GEEM). Results from GEEM and REM were fairly consistent. These observations confirm our expectation that KS can be more powerful than BT if the variants are associated with T2DM in different directions.

Mutations in *POMGNT1*, which encodes a type II transmembrane protein, may be related to muscle-eye-brain disease [Saredi et al., 2012]. The observed association with T2DM is novel and its implications are not clear. JAK1, a human tyrosine kinase protein, is essential for signaling certain cytokines [Gadina et al., 2001]. The *JAK1-STAT1* pathway is the major signaling pathway to mediate the effects of interferon gamma on beta cell apoptosis in type 1 diabetes mellitus and, possibly, in T2DM [Couto et al., 2007; Burke et al., 2013]. This pathway is also a novel target of the T2DM drug exenatide (Ex-4). Thus, Ex-4 treatment may protect beta cells from cytokine-induced cell death by inhibiting the *JAK-STAT1* pathway [Couto et al., 2007]. We found that *JAK1* is associated with T2DM, and recommend further research to understand the relationship between *JAK1,* JAK1 and T2DM.

Discussion

In this study, we extended the kernel statistic and burden tests to examine associations among multiple phenotypes and rare genetic variants in family studies. The GEEM was employed to generalize the pedigree-based kernel and burden tests to correlated phenotypes

under GLMs. The generalized kernel and burden tests are powerful and computationally efficient tests that retain all of the advantages of the tests proposed by Schaid et al. [2013]. They can be applied to study pleiotropic or longitudinal multiple phenotypes under the GLM framework, and to make full use of pedigree data with multiple affected and unaffected individuals, or even with additional related or unrelated subjects from population-based studies.

An important feature of the proposed approaches is that the KS and BT account for complex correlations between repeated measures from the same phenotype (serial correlations) and between individuals within the same family (familial correlations), and for the complex correlations between genetic markers within individuals. Moreover, the two proposed test statistics based on GEEM can be computed easily and efficiently with adjustment for covariates. Therefore, these methods can be used for large-scale data, such as WGS data, from a hybrid of pedigree- and population-based studies.

Compared with the proposed burden test and the existing GEEM and REM approaches, the proposed kernel statistic is advantageous in statistical power regardless of the direction of the variants' effects. The kernel test is robust to population stratification, whereas the burden test has a slight deflation in type I errors in the presence of population stratification. When prior knowledge of the true serial and familial correlation structures is not available, an exchangeable working correlation matrix is suggested for both proposed tests. Both tests performed better

when an exchangeable rather than an independent working correlation matrix was assumed in the simulation study.

Although the longitudinal family-based kernel and burden tests were developed for rare variants, they can also be used for common variants, mixtures of common and rare variants, or genetic scores, which can even be time-varying with an application of the empirical estimate for the genetic correlation matrix. Depending on the research interest, common variants can either be accounted for or collapsed with rare variants in the analyses. Ascertainment bias is not a concern because of the use of conditioning on the phenotypes in pedigree-based designs. The tests allow for the use of covariate adjustments and the incorporation of prior association information into the weight function. External weight functions based on annotation information or functional prediction can be incorporated into the tests to boost statistical power. One possibility is to incorporate weight functions by using Bayesian approaches [Couto et al., 2007]. Further investigations of how to incorporate bioinformatics data into the weight function are necessary.

Due to their low frequencies and potentially moderate effects on disease, rare variants that affect disease susceptibility can be challenging to identify. We have proposed methods focused on longitudinal family-based designs with the same phenotypes, which can easily be extended to designs with multiple different phenotypes. Modeling multiple phenotypes simultaneously may increase power due to the increased likelihood of affected relatives and the

correlations between the multiple disease-related phenotypes. Future studies should continue to develop the proposed methods for identifying rare variants associated with multiple phenotypes in longitudinal family-based designs.

Appendix A

Kernel Statistic for Longitudinal Family Data

The null distributions of $\boldsymbol{Z}^{LF}$

According to the multivariate central limit theorem, $\boldsymbol{Z}^{LF}$ in Equation 3 is asymptotically multivariate normal with a mean of $E(\boldsymbol{Z}^{LF})$ and a covariance matrix of $\mathrm{Cov}(\boldsymbol{Z}^{LF})$. The mean and variance of the KS, κ^{LF}, can be derived from $E(\kappa^{LF}) = \mathrm{tr}(\mathrm{Cov}(\boldsymbol{Z}^{LF}))+(E(\boldsymbol{Z}^{LF}))^T E(\boldsymbol{Z}^{LF})$ and $\mathrm{Var}(\kappa^{LF}) = 2\mathrm{tr}(\mathrm{Cov}(\boldsymbol{Z}^{LF})\mathrm{Cov}(\boldsymbol{Z}^{LF}))+4(E(\boldsymbol{Z}^{LF}))^T \mathrm{Cov}(\boldsymbol{Z}^{LF})E(\boldsymbol{Z}^{LF})$, where $\mathrm{tr}(\cdot)$ represents the trace of a matrix. Thus, the first two moments under the null hypothesis of no associations, $E_0(\kappa^{LF})$ and $\mathrm{Var}_0(\kappa^{LF})$, are derived from $E_0(\boldsymbol{Z}^{LF})$ and $\mathrm{Cov}_0(\boldsymbol{Z}^{LF})$.

Elements of $E_0(Z_l^{LF})$ of the expectation of $\boldsymbol{Z}^{LF}$ under the null hypothesis are $E_0(Z_l^{LF}) = w_l 2m_l \sum_{i=1}^{K} \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$, where $l = 1, 2, \ldots, p$. These elements have zero expectation asymptotically when $\boldsymbol{\mu}_i$, for $i = 1, 2, \ldots, K$, are correctly specified, because $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{V}}_i$ are consistent estimates of $E_0(\boldsymbol{Y}_i)$ and $\mathrm{Cov}_0(\boldsymbol{Y}_i)$, respectively. Estimates $\hat{\Delta}_i$ and $\hat{\boldsymbol{A}}_i$ are consistent estimates of Δ_i and $\boldsymbol{A}_i$ under the null hypothesis, as long as K is sufficiently large and $i = 1, 2, \ldots, K$, are correctly specified. The elements of $\mathrm{Cov}_0(Z_l^{LF}, Z_{l'}^{LF})$ are

$$\mathrm{Cov}_0(Z_l^{LF}, Z_{l'}^{LF}) = w_l w_{l'} \sum_{i=1}^{K} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{A}}_i \hat{\Delta}_i \mathrm{Cov}_0(\tilde{\boldsymbol{G}}_{il}, \tilde{\boldsymbol{G}}_{il'}) \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$$

where $\mathrm{Cov}_0(\tilde{\boldsymbol{G}}_{il}, \tilde{\boldsymbol{G}}_{il'}) = 2H_{ll'}\sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})}\tilde{\Omega}_i$ is a square matrix of order $\sum_{j=1}^{n_i} R_{ij}$ and $\tilde{\Omega}_i$ is a $(\sum_{j=1}^{n_i} R_{ij}) \times (\sum_{j=1}^{n_i} R_{ij})$ matrix with elements $\tilde{\Omega}_{ijr,ijr'} = \Omega_{ij,ij}$ for any r and r'. By writing

$C^{LF} = \sum_{i=1}^{K} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{A}}_i \hat{\Delta}_i \tilde{\Omega}_i \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$, the elements of $\mathrm{Cov}_0(\boldsymbol{Z}^{LF})$ can be expressed in the form $\mathrm{Cov}_0(Z_l^{LF}, Z_{l'}^{LF}) = w_l w_{l'} 2H_{ll'} \sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})} C^{LF}$. Therefore, under the null hypothesis, the mean and variance of κ^{LF} are $E_0(\kappa^{LF}) = \mathrm{tr}(\mathrm{Cov}_0(\boldsymbol{Z}^{LF}))$ and $\mathrm{Var}_0(\kappa^{LF}) = 2\mathrm{tr}(\mathrm{Cov}_0(\boldsymbol{Z}^{LF})\mathrm{Cov}_0(\boldsymbol{Z}^{LF}))$, respectively.

The null distributions of $\boldsymbol{Z}^{LF_I}$ and $\boldsymbol{Z}^{LF_{IC}}$

The null distribution of $\boldsymbol{Z}^{LF_I}$ in Equation 4 is asymptotically multivariate normal with a mean of zero and a covariance matrix of $\mathrm{Cov}_0(\boldsymbol{Z}^{LF_I})$. The elements of $\mathrm{Cov}_0(Z_l^{LF_I}, Z_{l'}^{LF_I})$ of $\mathrm{Cov}_0(\boldsymbol{Z}^{LF_I})$ are $\mathrm{Cov}_0(Z_l^{LF_I}, Z_{l'}^{LF_I}) = w_l w_{l'} 2H_{ll'} \sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})} C^{LF_I}$ with $C^{LF_I} = \sum_{i=1}^{K} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\Delta}_i \tilde{\Omega}_i \hat{\Delta}_i (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$. The null distribution of $\boldsymbol{Z}^{LF_{IC}}$ is asymptotically multivariate normal with a mean of zero and a covariance matrix of $\mathrm{Cov}_0(\boldsymbol{Z}^{LF_{IC}})$. The elements of $\mathrm{Cov}_0(Z_l^{LF_{IC}}, Z_{l'}^{LF_{IC}})$ of $\mathrm{Cov}_0(\boldsymbol{Z}^{LF_{IC}})$ are $\mathrm{Cov}_0(Z_l^{LF_{IC}}, Z_{l'}^{LF_{IC}}) = w_l w_{l'} 2H_{ll'} \sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})} C^{LF_{IC}}$ with $C^{LF_{IC}} = \sum_{i=1}^{K} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \tilde{\Omega}_i (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)$.

Appendix B

Extension to the X Chromosome

To develop our methods in a general way to code male genotypes for the X chromosome, we used *d* to represent the code for men who carry the minor allele. Thus, men are coded as 0 or

d ($d = 1$ or 2), whereas women are coded as 0, 1, or 2 (as for autosomes). Genetic correlations for the X chromosomes between a pair of relatives are given by

$$\Omega_{ij,ij'} = \begin{cases} k_1/2 + k_2 & \text{if female-}j \text{ and female-}j' \text{ pair,} \\ k_1 & \text{if male-}j \text{ and male-}j' \text{ pair,} \\ k_1/\sqrt{2} & \text{if female-}j \text{ and male-}j' \text{ pair.} \end{cases} \quad \text{(B.1)}$$

where k_1 and k_2 represent the probabilities of sharing one and two pairs of identical-by-descent alleles, respectively. The genetic correlations in Equation B.1 have the same form and interpretation as those in Equation 2 in the report by Schaid et al [2013]. Hence, the equation can be computed through the genetic correlations between X chromosomes in relative pairs, using the methods of Schaid et al. [2013].

With the genetic correlations in Equation B.1, the elements of the null covariance of the genotypes in matrix $\boldsymbol{G}_i$ for individuals j and j', and markers l and l', can be expressed as

$$\mathrm{Cov}_0(G_{ijl}, G_{ij'l'}) = \begin{cases} 2H_{ll'}\sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})}\Omega_{ij,ij'} & \text{if female-}j \text{ and female-}j' \text{ pair,} \\ d^2 H_{ll'}\sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})}\Omega_{ij,ij'} & \text{if male-}j \text{ and male-}j' \text{ pair,} \\ d\sqrt{2}H_{ll'}\sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})}\Omega_{ij,ij'} & \text{if female-}j \text{ and male-}j' \text{ pair.} \end{cases}$$

As for the case of autosomes, the elements of the null covariance of $\boldsymbol{G}_i$, $\mathrm{Cov}_0(G_{ijl}, G_{ij'l'})$, can be utilized to express the elements of the null covariance matrix of $\boldsymbol{Z}^{LF}$ as

$$\mathrm{Cov}_0(Z_l^{LF}, Z_{l'}^{LF}) = w_l w_{l'} 2H_{ll'}\sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})}C_{\mathrm{X}}^{LF}$$

where

$$C_{\mathrm{X}}^{LF} = \sum_{i=1}^{K} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{V}}_i^{-1} \hat{\boldsymbol{A}}_i \hat{\Delta}_i \tilde{\Omega}_i \lambda_i \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i).$$

Here, λ_i is a square matrix of order $\sum_{j=1}^{n_i} R_{ij}$, with the $(ijr, ij'r')^{\text{th}}$ element being

$$\lambda_{ijr,ij'r'} = \begin{cases} 2 & \text{if female-}j \text{ and female-}j' \text{ pair,} \\ d^2 & \text{if male-}j \text{ and male-}j' \text{ pair,} \\ d\sqrt{2} & \text{if female-}j \text{ and male-}j' \text{ pair.} \end{cases}$$

From this matrix, the KS for the X chromosome can be derived by using the same procedures for autosomes. Its approximate asymptotic distribution under the null hypothesis is the same as for κ^{LF}. Through similar derivatives for autosomes, the BT for the X chromosome is

$$B_{\text{X}}^{LF} = \frac{\left[\sum_{l=1}^{p} w_l \sum_{i=1}^{K} \tilde{\boldsymbol{G}}_{il}^T \hat{\Delta}_i \hat{\boldsymbol{A}}_i \hat{\boldsymbol{V}}_i^{-1} \left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i\right)\right]^2}{2\sum_{l=1}^{p}\sum_{l'=1}^{p} w_l w_{l'} H_{ll'} \sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})} C_{\text{X}}^{LF}}$$

which, under the null hypothesis, asymptotically has a chi-squared distribution with one degree of freedom.

Extension of the above test statistics to situations where the genetic relationships between individuals j and j' are unknown is straightforward. Using the same arguments as Yang et al. [2011] and Schaid et al. [2013], the elements of the genetic correlation matrix Ω_l for the X chromosome can be estimated by

$$\hat{\Omega}_{ij,ij'} = \begin{cases} \frac{1}{p}\sum_{l=1}^{p} \frac{(g_{ijl} - 2m_l)(g_{ij'l} - 2m_l)}{2m_l(1-m_l)} & \text{if female-}j \text{ and female-}j' \text{ pair,} \\ \frac{1}{p}\sum_{l=1}^{p} \frac{(g_{ijl} - m_l)(g_{ij'l} - m_l)}{m_l(1-m_l)} & \text{if male-}j \text{ and male-}j' \text{ pair,} \\ \frac{1}{p}\sum_{l=1}^{p} \frac{(g_{ijl} - m_l)(g_{ij'l} - 2m_l)}{\sqrt{2}m_l(1-m_l)} & \text{if male-}j \text{ and female-}j' \text{ pair.} \end{cases}$$

Acknowledgments

We thank Ms. Chun-Yi Lee and Mr. Ting-Yan Chang for their kind help with computational questions. We also thank Karen Klein, MA, ELS (Biomedical Research Services and Administration, Wake Forest University Health Sciences) for editing the manuscript. None of the authors has conflict of interest. This project was supported in part by grants from the Ministry of Science and Technology (MOST102-2118-M-400-005) and the National Health Research Institutes in Taiwan (PH-101-PP-04, PH-102-PP-04, PH-103-PP-04, and PH-104-PP-04).

References

Asimit JL, Day-Williams AG, Morris AP, Zeggini E. 2012. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Hum Hered* 73:84-94.

Balding DJ, Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96:3-12.

Basu S, Pan W. 2011. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 35:606-619.

Bates DM. 2010. Lme4: Mixed-Effects Modeling with R.

Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 40:695-701.

Bowden DW, Lehtinen AB, Ziegler JT, Rudock ME, Xu J, Wagenknecht LE, Herrington DM, Rich SS, Freedman BI, Carr JJ and others. 2008. Genetic epidemiology of subclinical cardiovascular disease in the Diabetes Heart Study. *Ann Hum Genet* 72:598-610.

Burke SJ, Goff MR, Lu D, Proud D, Karlstad MD, Collier JJ. 2013. Synergistic expression of the CXCL10 gene in response to IL-1beta and IFN-gamma involves NF-kappaB, phosphorylation of STAT1 at Tyr701, and acetylation of histones H3 and H4. *Journal of*

immunology 191:323-336.

Burton PR, Scurrah KJ, Tobin MD, Palmer LJ. 2005. Covariance components models for longitudinal family data. *Int J Epidemiol* 34:1063-1077.

Carey VJ, Lumley T, and Ripley B. 2015. gee: Generalized Estimation Equation Solver, URL http://CRAN.R-project.org/package=gee, R package version 4.13-18.

Chen H, Meigs JB, Dupuis J. 2013. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* 37:196-204.

Couto FM, Minn AH, Pise-Masison CA, Radonovich M, Brady JN, Hanson M, Fernandez LA, Wang P, Kendziorski C, Shalev A. 2007. Exenatide blocks JAK1-STAT1 in pancreatic beta cells. *Metabolism: clinical and experimental* 56:915-918.

Das K, Li J, Wang Z, Tong C, Fu G, Li Y, Xu M, Ahn K, Mauger D, Li R and others. 2011. A dynamic model for genome-wide association studies. *Hum Genet* 129:629-639.

De G, Yip WK, Ionita-Laza I, Laird N. 2013. Rare variant analysis for family-based design. *PLoS One* 8.

Diggle P, Liang K, Zeger S. 1994. Analysis of Longitudinal Data. *Oxford University Press*.

Fan RZ, Zhang YW, Albert PS, Liu AY, Wang YJ, Xiong MM. 2012. Longitudinal association analysis of quantitative traits. *Genet Epidemiol* 36:856-869.

Furlotte NA, Eskin E, Eyheramendy S. 2012. Genome-wide association mapping with

longitudinal data. *Genet Epidemiol* 36:463-471.

Gadina M, Hilton D, Johnston JA, Morinobu A, Lighvani A, Zhou YJ, Visconti R, O'Shea JJ. 2001. Signaling by Type I and II cytokine receptors: ten years after. *Curr Opin Immunol* 13:363-373.

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. 2013. Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur J Hum Genet* 21:1158-1162.

Jiang D, McPeek MS. 2014. Robust Rare Variant Association Testing for Quantitative Traits in Samples With Related Individuals. *Genet Epidemiol* 38:10-20.

Kraft P, Thomas DC. 2000. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet* 66:1119–1131.

Laird NM, Horvath S, Xu X. 2000. Implementing a unified approach to family-based tests of association. *Genet Epidemiol* 19:S36-S42.

Lee S, Abecasis GR, Boehnke M, Lin XH. 2014. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genets* 95:5-23.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. 2012a.

Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 91: 224-237.

Lee S, Wu MC, Lin X. 2012b. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13:762-775.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83:311-321.

Liang KY, Zeger SL. 1986. Longitudinal data-analysis using generalized linear-models. *Biometrika* 73:13-22.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5:e1000384.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet* 24:133-141.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nature reviews Genetics* 11:31-46.

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat Res* 615:28-56.

Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in

genetic association studies. *Genet Epidemiol* 34:188-193.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet* 7:e1001322.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 69:124-137.

Saredi S, Ardissone A, Ruggieri A, Mottarelli E, Farina L, Rinaldi R, Silvestri E, Gandioli C, D'Arrigo S, Salerno F and others. 2012. Novel POMGNT1 point mutations and intragenic rearrangements associated with muscle-eye-brain disease. *J Neurol Sci* 318:45-50.

Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. 2013. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet Epidemiol* 37:409-418.

Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SLR, Peyser PA, Lin XH. 2012. SNP set association analysis for familial data. *Genet Epidemiol* 36:797-810.

Smith EN, Chen W, Kahonen M, Kettunen J, Lehtimaki T, Peltonen L, Raitakari OT, Salem RM, Schork NJ, Shaw M and others. 2010. Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa heart study. *PLoS Genet* 6:e1001094.

Thornton T, McPeek MS. 2010. ROADTRIPS: case-control association testing with partially or

completely unknown population and pedigree structure. *Am J Hum Genet* 86:172-184.

Turkmen AS, Lin SL. 2014. Blocking Approach for Identification of Rare Variants in Family-Based Association Studies. *PLoS One* 9: e86126.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89:82-93.

Wu Z, Hu Y, Melton PE. 2014. Longitudinal data analysis for genetic studies in the whole-genome sequencing era. *Genet Epidemiol* 38 Suppl 1:S74-80.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88:76-82.

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. 2010. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 87:604-617.

Figure Titles and Legends

Figure 1. Power for the KS, BT, GEEM, and REM without population stratification, by the specified working correlations. Number of families (K) equal to 450 and number of variants (p) equal to 100 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.

Figure 2. Power for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) equal to 450 and number of variants (p) equal to 50 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.

Figure 3. Power for the KS, BT, GEEM, and REM with population stratification by the specified working correlations. Number of families (K) equal to 450 and number of variants (p) equal to 100 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.

Figure 4. Power for the KS, BT, GEEM, and REM with population stratification by the specified working correlations. Number of families (K) equal to 450 and number of variants (p) equal to 50 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.

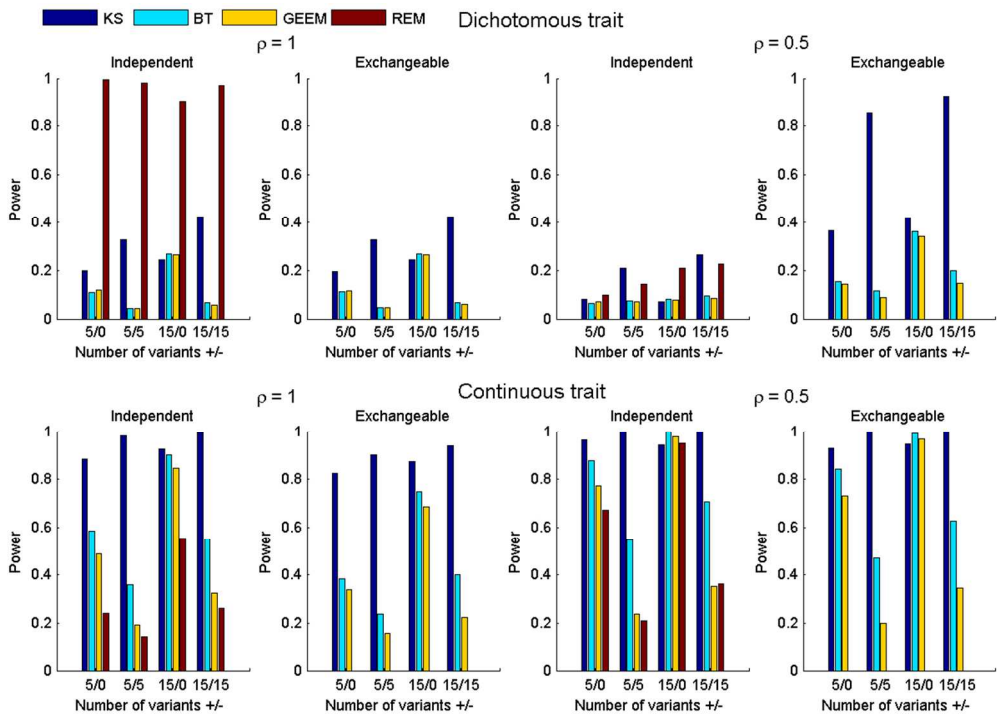

Figure 1

Figure 1. Power for the KS, BT, GEEM, and REM without population stratification, by the specified working correlations. Number of families (K) equal to 450 and number of variants (p) equal to 100 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.
203x152mm (150 x 150 DPI)

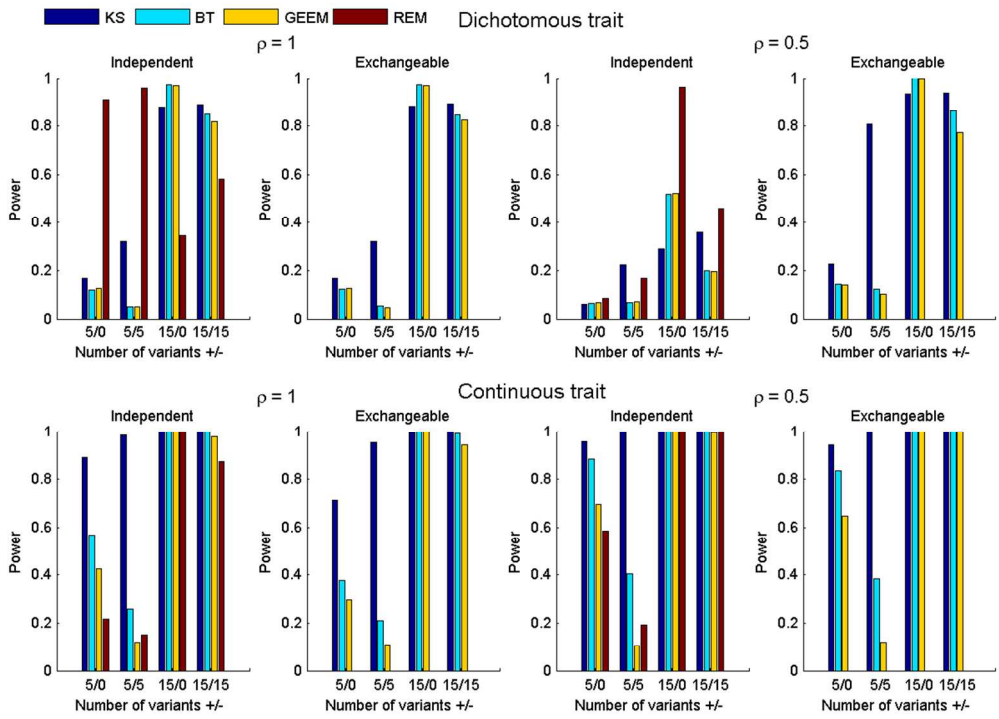

Figure 2

Figure 2. Power for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) equal to 450 and number of variants (p) equal to 50 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.
203x152mm (150 x 150 DPI)

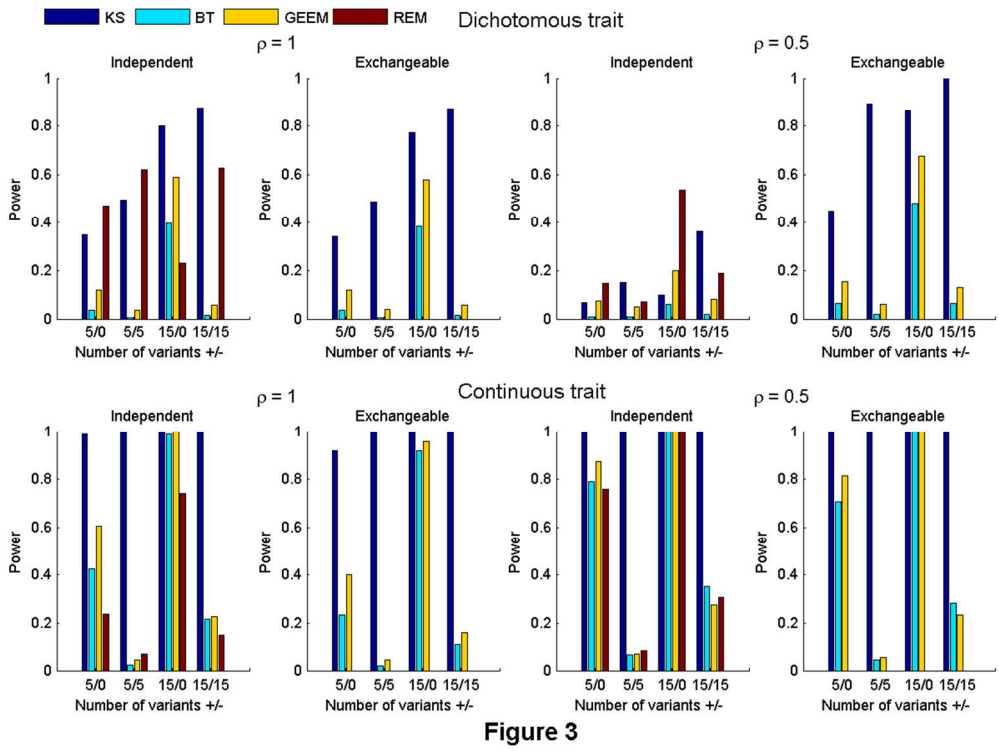

Figure 3

Figure 3. Power for the KS, BT, GEEM, and REM with population stratification by the specified working correlations. Number of families (K) equal to 450 and number of variants (p) equal to 100 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.
203x152mm (150 x 150 DPI)

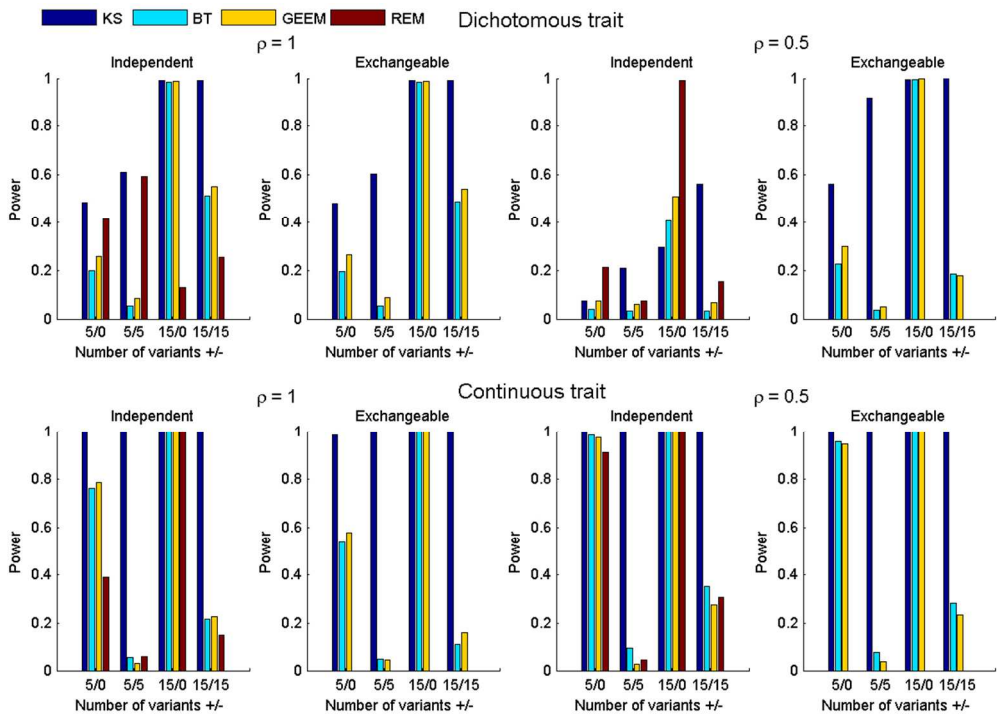

Figure 4

Figure 4. Power for the KS, BT, GEEM, and REM with population stratification by the specified working correlations. Number of families (K) equal to 450 and number of variants (p) equal to 50 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.
203x152mm (150 x 150 DPI)

Table 1. Type I errors for the KS, BT, GEEM, and REM in the absence of population stratification without LD

				Dichotomous trait				Continuous trait			
No. of families	No. of variants	True correlation	Working correlation	KS	BT	GEEM	REM[a]	KS	BT	GEEM	REM[a]
$\alpha = 0.05$											
$K = 450$	$p = 100$	$\rho = 1$	Independent	0.056	0.035	0.038	0.981	0.046	0.054	0.058	0.078
			Exchangeable	0.058	0.034	0.037	-	0.041	0.047	0.058	-
		$\rho = 0.5$	Independent	0.058	0.051	0.051	0.070	0.062	0.049	0.053	0.057
			Exchangeable	0.047	0.043	0.047	-	0.053	0.044	0.047	-
	$p = 50$	$\rho = 1$	Independent	0.056	0.044	0.055	0.959	0.055	0.054	0.061	0.067
			Exchangeable	0.056	0.044	0.057	-	0.055	0.055	0.056	-
		$\rho = 0.5$	Independent	0.052	0.048	0.047	0.072	0.057	0.045	0.050	0.065
			Exchangeable	0.055	0.044	0.045	-	0.052	0.048	0.055	-
$K = 225$	$p = 100$	$\rho = 1$	Independent	0.033	0.043	0.047	0.860	0.039	0.045	0.047	0.068
			Exchangeable	0.037	0.048	0.051	-	0.048	0.041	0.057	-
		$\rho = 0.5$	Independent	0.041	0.039	0.044	0.081	0.050	0.044	0.057	0.066
			Exchangeable	0.048	0.049	0.054	-	0.045	0.055	0.062	-
	$p = 50$	$\rho = 1$	Independent	0.040	0.046	0.053	0.796	0.026	0.067	0.073	0.081
			Exchangeable	0.043	0.047	0.054	-	0.044	0.061	0.078	-
		$\rho = 0.5$	Independent	0.041	0.044	0.050	0.086	0.045	0.053	0.070	0.078
			Exchangeable	0.043	0.052	0.064	-	0.040	0.054	0.071	-

KS, kernel association test statistic; BT, burden test; GEEM, generalized estimating equation method; REM, random-effects model.
[a]A working correlation matrix does not need to be specified for the REM.

Table 2. Type I errors for the KS, BT, GEEM, and REM in the presence of population stratification without LD

No. of families	No. of variants	True correlation	Working correlation	Dichotomous trait KS	Dichotomous trait BT	Dichotomous trait GEEM	Dichotomous trait REM[a]	Continuous trait KS	Continuous trait BT	Continuous trait GEEM	Continuous trait REM[a]
$\alpha = 0.05$											
$K = 450$	$p = 100$	$\rho = 1$	Independent	0.058	0.005	0.035	0.619	0.036	0.009	0.051	0.064
			Exchangeable	0.060	0.005	0.036	-	0.041	0.003	0.044	-
		$\rho = 0.5$	Independent	0.041	0.006	0.047	0.052	0.037	0.015	0.049	0.051
			Exchangeable	0.042	0.009	0.052	-	0.035	0.011	0.051	-
	$p = 50$	$\rho = 1$	Independent	0.061	0.029	0.050	0.621	0.045	0.019	0.051	0.059
			Exchangeable	0.059	0.028	0.051	-	0.049	0.037	0.059	-
		$\rho = 0.5$	Independent	0.043	0.028	0.049	0.060	0.057	0.027	0.043	0.055
			Exchangeable	0.051	0.027	0.056	-	0.049	0.029	0.056	-

KS, kernel association test statistic; BT, burden test; GEEM, generalized estimating equation method; REM, random-effects model.
[a]A working correlation matrix does not need to be specified.

Table 3. Type I errors for the KS, BT, GEEM, REM, KSS, BTS and famSKAT in the absence of population stratification with LD

					Dichotomous trait						Continuous trait				
No. of families	No. of variants	True correlation	Linkage disequilibrium	Working correlation	KS	BT	GEEM	REM[a]	KSS[a]	BTS[a]	KS	BT	GEEM	REM[a]	famSKAT[a] (last time point only)
$\alpha = 0.05$															
$K = 450$	$p = 100$	$\rho = 0.5$	$\rho_{LD} = 0.9$	Independent	0.032	0.045	0.050	0.072	0.042	0.044	0.053	0.053	0.060	0.060	0.029 (0.026)
				Exchangeable	0.049	0.056	0.057	-	-	-	0.052	0.053	0.063	-	-
			$\rho_{LD} = 0.5$	Independent	0.041	0.056	0.054	0.073	0.042	0.045	0.052	0.044	0.054	0.050	0.026 (0.028)
				Exchangeable	0.049	0.046	0.051	-	-	-	0.060	0.050	0.058	-	-
$\alpha = 0.01$															
$K = 450$	$p = 100$	$\rho = 0.5$	$\rho_{LD} = 0.9$	Independent	0.009	0.010	0.015	0.020	0.010	0.011	0.009	0.014	0.016	0.016	0.007 (0.008)
				Exchangeable	0.006	0.013	0.012	-	-	-	0.011	0.012	0.015	-	-
			$\rho_{LD} = 0.5$	Independent	0.010	0.011	0.011	0.019	0.007	0.009	0.015	0.007	0.008	0.013	0.007 (0.004)
				Exchangeable	0.015	0.011	0.013	-	-	-	0.012	0.009	0.013	-	-

KS, kernel association test statistic; BT, burden test; GEEM, generalized estimating equation method; REM, random-effects model; KSS, kernel association statistic by Schaid et al. [2013]; BTS, burden test by Schaid et al. [2013]; famSKAT, family-based sequence kernel association test.

[a]A working correlation matrix does not need to be specified.

Table 4. Powers for the KS, BT, GEEM, REM, KSS, BTS and famSKAT in the absence of population stratification with LD

					Dichotomous trait						Continuous trait				
No. of families	No. of variants	True correlation	Linkage disequilibrium	Working correlation	KS	BT	GEEM	REM[a]	KSS[a]	BTS[a]	KS	BT	GEEM	REM[a]	famSKAT[a] (last time point only)
$\alpha = 0.05$															
Number of risk/protective variants 5/0															
$K = 450$	$p = 100$	$\rho = 0.5$	$\rho_{LD} = 0.9$	Independent	0.057	0.064	0.074	0.120	0.051	0.052	0.963	0.855	0.715	0.663	0.996 (0.984)
				Exchangeable	0.377	0.181	0.167	-	-	-	0.953	0.785	0.663	-	-
			$\rho_{LD} = 0.5$	Independent	0.075	0.067	0.071	0.111	0.044	0.056	0.956	0.895	0.800	0.648	0.998 (0.986)
				Exchangeable	0.438	0.198	0.187	-	-	-	0.930	0.895	0.756	-	-
Number of risk/protective variants 5/5															
$K = 450$	$p = 100$	$\rho = 0.5$	$\rho_{LD} = 0.9$	Independent	0.182	0.053	0.059	0.184	0.113	0.058	1.000	0.538	0.238	0.363	0.998 (0.998)
				Exchangeable	0.842	0.116	0.078	-	-	-	1.000	0.512	0.238	-	-
			$\rho_{LD} = 0.5$	Independent	0.202	0.064	0.065	0.131	0.119	0.069	1.000	0.426	0.180	0.223	0.997 (0.997)
				Exchangeable	0.860	0.110	0.080	-	-	-	1.000	0.408	0.173	-	-
$\alpha = 0.01$															
Number of risk/protective variants 5/0															
$K = 450$	$p = 100$	$\rho = 0.5$	$\rho_{LD} = 0.9$	Independent	0.015	0.012	0.020	0.044	0.010	0.013	0.958	0.729	0.473	0.451	0.990 (0.946)
				Exchangeable	0.174	0.063	0.067	-	-	-	0.931	0.634	0.423	-	-
			$\rho_{LD} = 0.5$	Independent	0.018	0.016	0.021	0.040	0.009	0.011	0.951	0.787	0.536	0.407	0.996 (0.969)
				Exchangeable	0.209	0.061	0.052	-	-	-	0.918	0.713	0.498	-	-
Number of risk/protective variants 5/5															
$K = 450$	$p = 100$	$\rho = 0.5$	$\rho_{LD} = 0.9$	Independent	0.061	0.014	0.015	0.060	0.023	0.015	1.000	0.408	0.082	0.170	0.998 (0.998)
				Exchangeable	0.684	0.035	0.020	-	-	-	1.000	0.376	0.100	-	-
			$\rho_{LD} = 0.5$	Independent	0.062	0.012	0.015	0.040	0.028	0.015	1.000	0.279	0.066	0.081	0.997 (0.997)
				Exchangeable	0.683	0.023	0.007	-	-	-	1.000	0.252	0.071	-	-

KS, kernel association test statistic; BT, burden test; GEEM, generalized estimating equation method; REM, random-effects model; KSS, kernel association statistic by Schaid et al. [2013]; BTS, burden test by Schaid et al. [2013]; famSKAT, family-based sequence kernel association test.

[a]A working correlation matrix does not need to be specified.

Table 5. Association tests from the family-based diabetes heart study: Significant genes with a p-value < 4.83×10^{-5} from at least one of the six tests

		P-value					
Covariates	GENE	KS	BT	GEEM	REM	KSS	BTS
Basic adjustment[a]	*POMGNT1*	2.11×10^{-4}	2.81×10^{-5}	1.23×10^{-4}	0.16	9.29×10^{-4}	4.57×10^{-5}
Full adjustment[b]	*POMGNT1*	1.19×10^{-4}	1.07×10^{-5}	4.05×10^{-5}	0.28	5.54×10^{-4}	2.75×10^{-5}
	JAK1	4.58×10^{-5}	1.46×10^{-4}	1.26×10^{-3}	2.01×10^{-2}	7.38×10^{-5}	2.14×10^{-4}

KS, kernel association test statistic; BT, burden test; GEEM, generalized estimating equation method; REM, random-effects model; KSS, kernel association statistic by Schaid et al. [2013]; BTS, burden test by Schaid et al. [2013].

[a]Adjusted for age and sex.

[b]Adjusted for age, sex, and BMI.

Table S1. Type I errors for the KS, BT, GEEM, and REM in the absence of population stratification without LD

No. of families	No. of variant	True correlation	Working correlation	Dichotomous trait KS	BT	GEEM	REM[a]	Continuous trait KS	BT	GEEM	REM[a]
α = 0.01											
K = 450	*p* = 100	ρ = 1	Independent	0.007	0.009	0.011	0.936	0.012	0.014	0.015	0.021
			Exchangeable	0.007	0.008	0.010	-	0.010	0.008	0.010	-
		ρ = 0.5	Independent	0.009	0.008	0.008	0.017	0.008	0.010	0.011	0.017
			Exchangeable	0.007	0.007	0.008	-	0.006	0.011	0.010	-
	p = 50	ρ = 1	Independent	0.010	0.008	0.010	0.886	0.012	0.012	0.013	0.022
			Exchangeable	0.010	0.010	0.010	-	0.008	0.013	0.015	-
		ρ = 0.5	Independent	0.008	0.008	0.013	0.016	0.011	0.010	0.013	0.016
			Exchangeable	0.007	0.006	0.005	-	0.012	0.012	0.012	-
K = 225	*p* =100	ρ = 1	Independent	0.005	0.008	0.009	0.656	0.004	0.011	0.013	0.016
			Exchangeable	0.008	0.008	0.010	-	0.011	0.009	0.015	-
		ρ = 0.5	Independent	0.009	0.008	0.012	0.022	0.009	0.011	0.017	0.020
			Exchangeable	0.008	0.008	0.012	-	0.010	0.009	0.014	-
	p = 50	ρ = 1	Independent	0.007	0.012	0.011	0.571	0.007	0.014	0.022	0.018
			Exchangeable	0.009	0.011	0.012	-	0.008	0.017	0.024	-
		ρ = 0.5	Independent	0.014	0.006	0.012	0.026	0.010	0.010	0.017	0.017
			Exchangeable	0.012	0.012	0.013	-	0.009	0.012	0.020	-

KS, kernel association statistic; BT, burden test; GEEM, generalized estimating equation method; REM, random-effects model.
[a]A working correlation matrix does not need to be specified for the REM.

Table S2. Type I errors for the X chromosome for the KS, BT, GEEM and REM in the absence of population stratification without LD

				Dichotomous trait				Continuous trait			
No. of families	No. of variants	True correlation	Working correlation	KS	BT	GEEM	REM[a]	KS	BT	GEEM	REM[a]
$\alpha = 0.05$											
$K = 450$	$p = 100$	$\rho = 1$	Independent	0.056	0.037	0.038	0.982	0.054	0.053	0.061	0.080
			Exchangeable	0.059	0.037	0.039	-	0.047	0.047	0.060	-
		$\rho = 0.5$	Independent	0.056	0.053	0.050	0.071	0.062	0.049	0.052	0.057
			Exchangeable	0.050	0.041	0.046	-	0.052	0.044	0.047	-
	$p = 50$	$\rho = 1$	Independent	0.057	0.046	0.056	0.960	0.055	0.055	0.064	0.067
			Exchangeable	0.056	0.045	0.058	-	0.055	0.057	0.058	-
		$\rho = 0.5$	Independent	0.051	0.046	0.046	0.075	0.058	0.047	0.050	0.065
			Exchangeable	0.056	0.042	0.045	-	0.053	0.049	0.055	-
$K = 225$	$p = 100$	$\rho = 1$	Independent	0.037	0.043	0.046	0.861	0.040	0.043	0.049	0.068
			Exchangeable	0.035	0.047	0.050	-	0.043	0.042	0.057	-
		$\rho = 0.5$	Independent	0.044	0.038	0.044	0.079	0.052	0.047	0.057	0.066
			Exchangeable	0.049	0.050	0.053	-	0.047	0.055	0.062	-
	$p = 50$	$\rho = 1$	Independent	0.038	0.047	0.052	0.795	0.025	0.071	0.072	0.080
			Exchangeable	0.039	0.043	0.052	-	0.039	0.063	0.077	-
		$\rho = 0.5$	Independent	0.042	0.045	0.049	0.086	0.045	0.052	0.070	0.078
			Exchangeable	0.045	0.053	0.064	-	0.038	0.054	0.071	-
$\alpha = 0.01$											
$K = 450$	$p = 100$	$\rho = 1$	Independent	0.007	0.009	0.010	0.935	0.011	0.015	0.015	0.021
			Exchangeable	0.007	0.008	0.009	-	0.011	0.008	0.011	-
		$\rho = 0.5$	Independent	0.011	0.008	0.008	0.017	0.009	0.010	0.011	0.017
			Exchangeable	0.006	0.007	0.008	-	0.007	0.011	0.011	-

	p = 50	ρ = 1	Independent	0.008	0.008	0.010	0.887	0.011	0.013	0.013	0.022
			Exchangeable	0.009	0.010	0.010	-	0.007	0.013	0.016	-
		ρ = 0.5	Independent	0.009	0.008	0.012	0.017	0.010	0.010	0.013	0.016
			Exchangeable	0.008	0.006	0.007	-	0.013	0.012	0.012	-
K= 225	p =100	ρ = 1	Independent	0.007	0.008	0.009	0.662	0.004	0.009	0.014	0.016
			Exchangeable	0.009	0.008	0.010	-	0.011	0.010	0.015	-
		ρ = 0.5	Independent	0.009	0.009	0.012	0.021	0.008	0.011	0.017	0.020
			Exchangeable	0.009	0.006	0.012	-	0.010	0.010	0.014	-
	p = 50	ρ = 1	Independent	0.007	0.011	0.011	0.572	0.006	0.013	0.022	0.018
			Exchangeable	0.008	0.010	0.012	-	0.004	0.016	0.025	-
		ρ = 0.5	Independent	0.012	0.006	0.012	0.025	0.009	0.010	0.017	0.017
			Exchangeable	0.013	0.011	0.014	-	0.011	0.012	0.020	-

KS, kernel association statistic; BT, burden test; GEEM, generalized estimating equation method; REM, random-effects model.
[a]A working correlation matrix does not need to be specified for the REM.

Table S3. Type I errors for the KS, BT, GEEM, REM and famSKAT in the absence of population stratification with LD and a constantly high phenotypic variance over time

					Continuous trait				
No. of families	No. of variants	True correlation	Linkage disequilibrium	Working correlation	KS	BT	GEEM	REM[a]	famSKAT[a] (last time point only)
α = 0.05									
K = 225	p = 100	ρ = 0.5	ρ_{LD} = 0.9	Independent	0.046	0.050	0.081	0.054	0.045 (0.037)

				Exchangeable	0.040	0.051	0.083	-	-
			ρ_{LD} = 0.5	Independent	0.055	0.043	0.054	0.043	0.064 (0.059)
				Exchangeable	0.054	0.046	0.054	-	-
α = 0.01									
K = 225	*p* = 100	ρ = 0.5	ρ_{LD} = 0.9	Independent	0.005	0.009	0.017	0.008	0.008 (0.007)
				Exchangeable	0.005	0.007	0.018	-	-
			ρ_{LD} = 0.5	Independent	0.011	0.012	0.009	0.009	0.014 (0.012)
				Exchangeable	0.011	0.012	0.009	-	-

KS, kernel association statistic; BT, burden test; GEEM, generalized estimating equation method; REM, random-effects model; famSKAT, family-based sequence kernel association test.
[a]A working correlation matrix does not need to be specified for the REM and famSKAT.

Table S4. Powers for the KS, BT, GEEM, REM and famSKAT in the absence of population stratification with LD and a constantly high phenotypic variance over time

					Continuous trait				
No. of families	No. of variants	True correlation	Linkage disequilibrium	Working correlation	KS	BT	GEEM	REM[a]	famSKAT[a] (last time point only)
α = 0.05									
The numbers of the risk/protective variants are 5/0									
K = 225	*p* = 100	ρ = 0.5	ρ_{LD} = 0.9	Independent	0.544	0.252	0.224	0.233	0.590 (0.145)
				Exchangeable	0.536	0.253	0.227	-	-
			ρ_{LD} = 0.5	Independent	0.539	0.243	0.215	0.196	0.611 (0.196)

				Exchangeable	0.531	0.241	0.214	-	-
The numbers of the risk/protective variants are 5/5									
$K = 225$	$p = 100$	$\rho = 0.5$	$\rho_{LD} = 0.9$	Independent	0.940	0.195	0.133	0.127	0.960 (0.418)
				Exchangeable	0.942	0.194	0.131	-	-
			$\rho_{LD} = 0.5$	Independent	0.896	0.139	0.096	0.083	0.958 (0.430)
				Exchangeable	0.899	0.141	0.098	-	-
$\alpha = 0.01$									
The numbers of the risk/protective variants are 5/0									
$K = 225$	$p = 100$	$\rho = 0.5$	$\rho_{LD} = 0.9$	Independent	0.345	0.108	0.095	0.088	0.402 (0.048)
				Exchangeable	0.339	0.109	0.091	-	-
			$\rho_{LD} = 0.5$	Independent	0.353	0.112	0.085	0.087	0.417 (0.068)
				Exchangeable	0.346	0.114	0.089	-	-
The numbers of the risk/protective variants are 5/5									
$K = 225$	$p = 100$	$\rho = 0.5$	$\rho_{LD} = 0.9$	Independent	0.858	0.072	0.040	0.038	0.903 (0.207)
				Exchangeable	0.854	0.069	0.042	-	-
			$\rho_{LD} = 0.5$	Independent	0.802	0.059	0.033	0.023	0.900 (0.233)
				Exchangeable	0.798	0.058	0.036	-	-

KS, kernel association statistic; BT, burden test; GEEM, generalized estimating equation method; REM, random-effects model; famSKAT, family-based sequence kernel association test.
[a]A working correlation matrix does not need to be specified for the REM and famSKAT.

Table S5. Information for *POMGNT1* and *JAK1*

	SNP	Position (base-pair)	MAF
a. *POMGNT1*			
	exm56140	46655158	0.0023

	exm56153	46655645	0.016
	exm56170	46657769	0.0023
	exm56174	46657799	0
	exm56222	46660037	0
	exm56246	46661749	0.0023
	exm56263	46662669	0
	exm56278	46669276	0.011
	exm56301	46685792	0.051
b. *JAK1*			
	exm66205	65305287	0.0045
	exm66228	65312368	0.0023
	exm66238	65321324	0
	exm66249	65330499	0
	exm2252624	65339122	0
	exm66280	65348981	0
	exm-rs310199	65350122	0.32

Table S6. Association test of *POMGNT1* and *JAK1* adjusting for common variants

GENE	*P*-value					
	KS	BT	GEEM	REM	KSS	BTS
POMGNT1	8.96×10^{-3}	9.72×10^{-3}	3.80×10^{-3}	2.36×10^{-1}	8.73×10^{-3}	6.04×10^{-3}
JAK1	4.14×10^{-1}	4.44×10^{-1}	6.01×10^{-2}	1.54×10^{-1}	5.62×10^{-1}	5.77×10^{-1}

KS, kernel association statistic; BT, burden test; GEEM, generalized estimating equation method;
REM, random-effects model; KSS, kernel association statistic by Schaid et al. [2013]; BTS, burden test by Schaid et al. [2013].

Table S7. Regression analysis for the common and collapsed rare SNPs of *POMGNT1* and *JAK1* genes by GEEM and REM

		GEEM		REM	
GENE	Regression coefficient	Common SNP	Collapsed rare SNPs	Common SNP	Collapsed rare SNPs
POMGNT1	Estimate	-1.52	-1.41	-2.14	-1.51
	P-value	1.58×10^{-3}	3.80×10^{-3}	6.18×10^{-4}	2.36×10^{-1}
JAK1	Estimate	1.29	-0.71	1.62	-1.06
	P-value	1.53×10^{-3}	6.01×10^{-2}	5.90×10^{-6}	1.54×10^{-1}

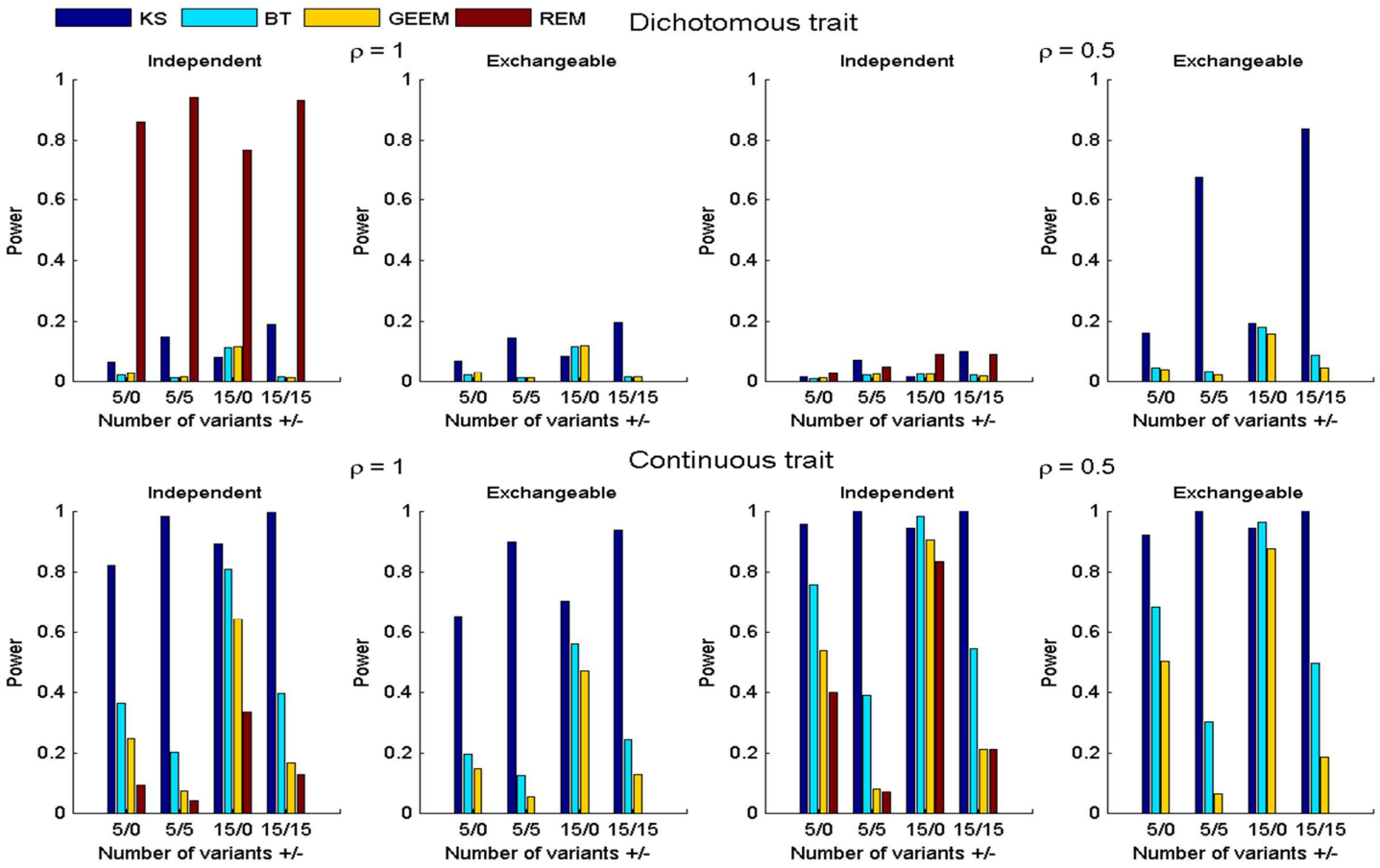

Figure S1. Powers for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 450 and number of variants (p) is 100 at a nominal level of 0.01. A working correlation matrix does not need to be specified for the REM.

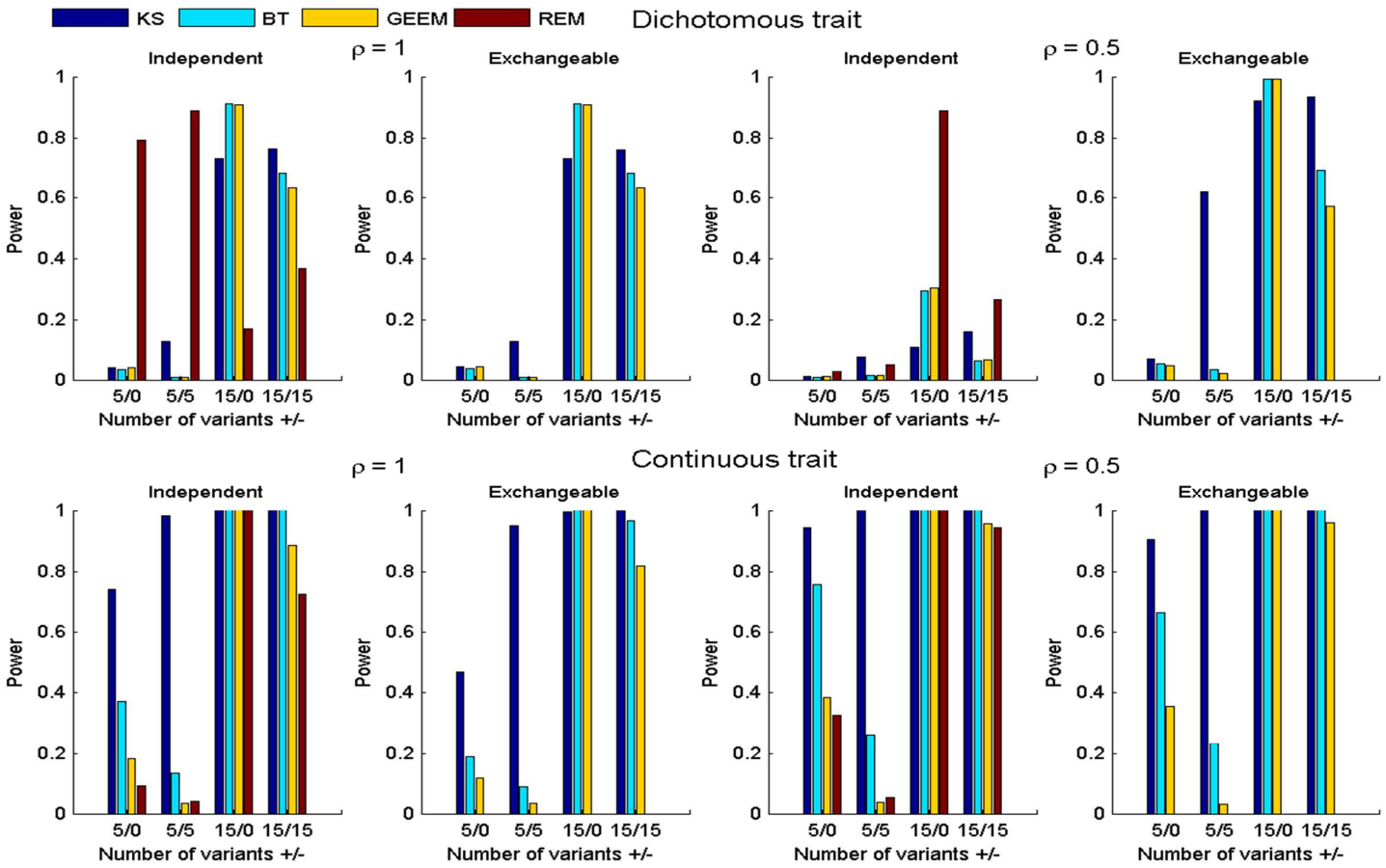

Figure S2. Powers for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 450 and number of variants (p) is 50 at a nominal level of 0.01. A working correlation matrix does not need to be specified for the REM.

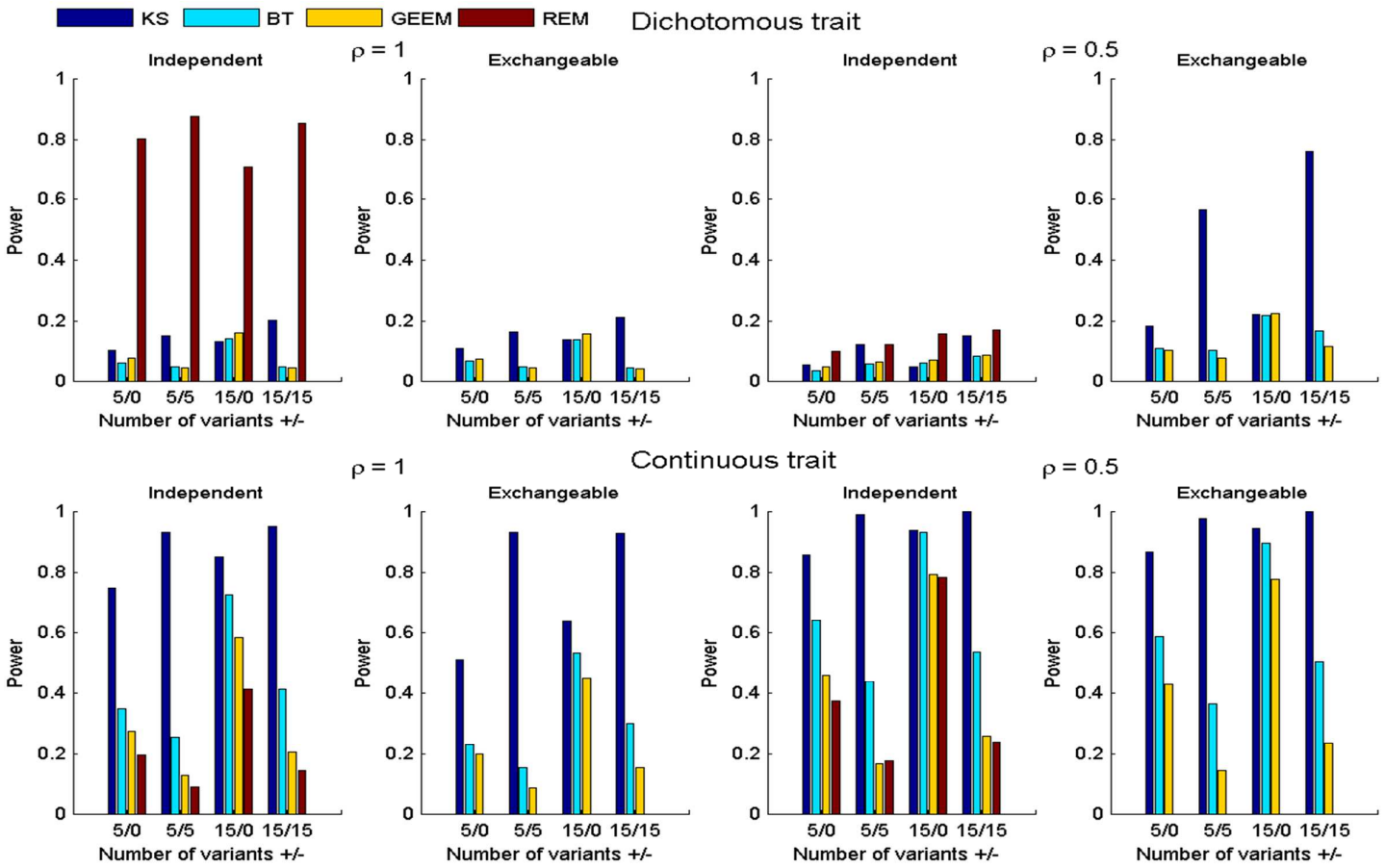

Figure S3. Powers for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 225 and number of variants (p) is 100 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.

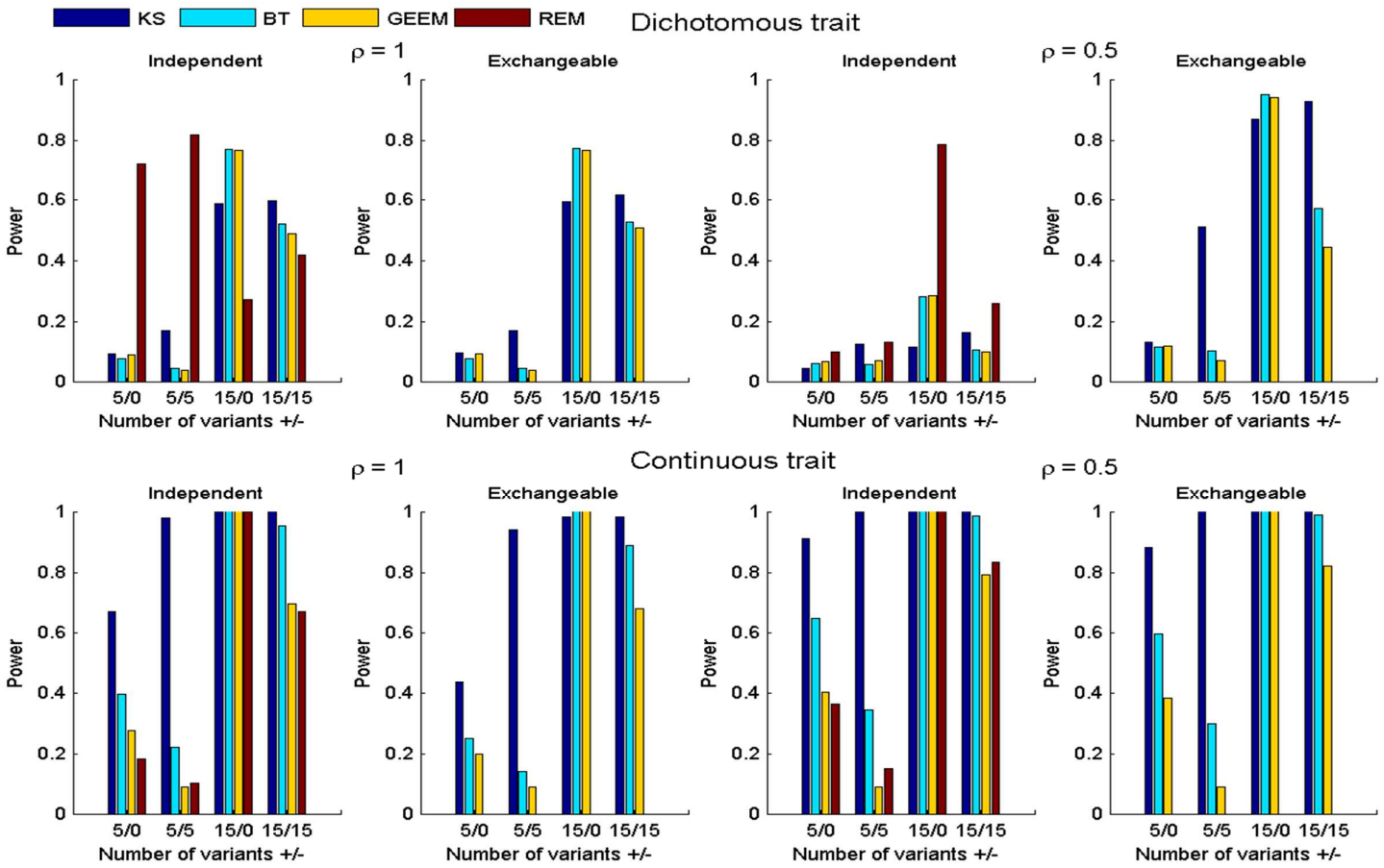

Figure S4. Powers for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 225 and number of variants (p) is 50 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.

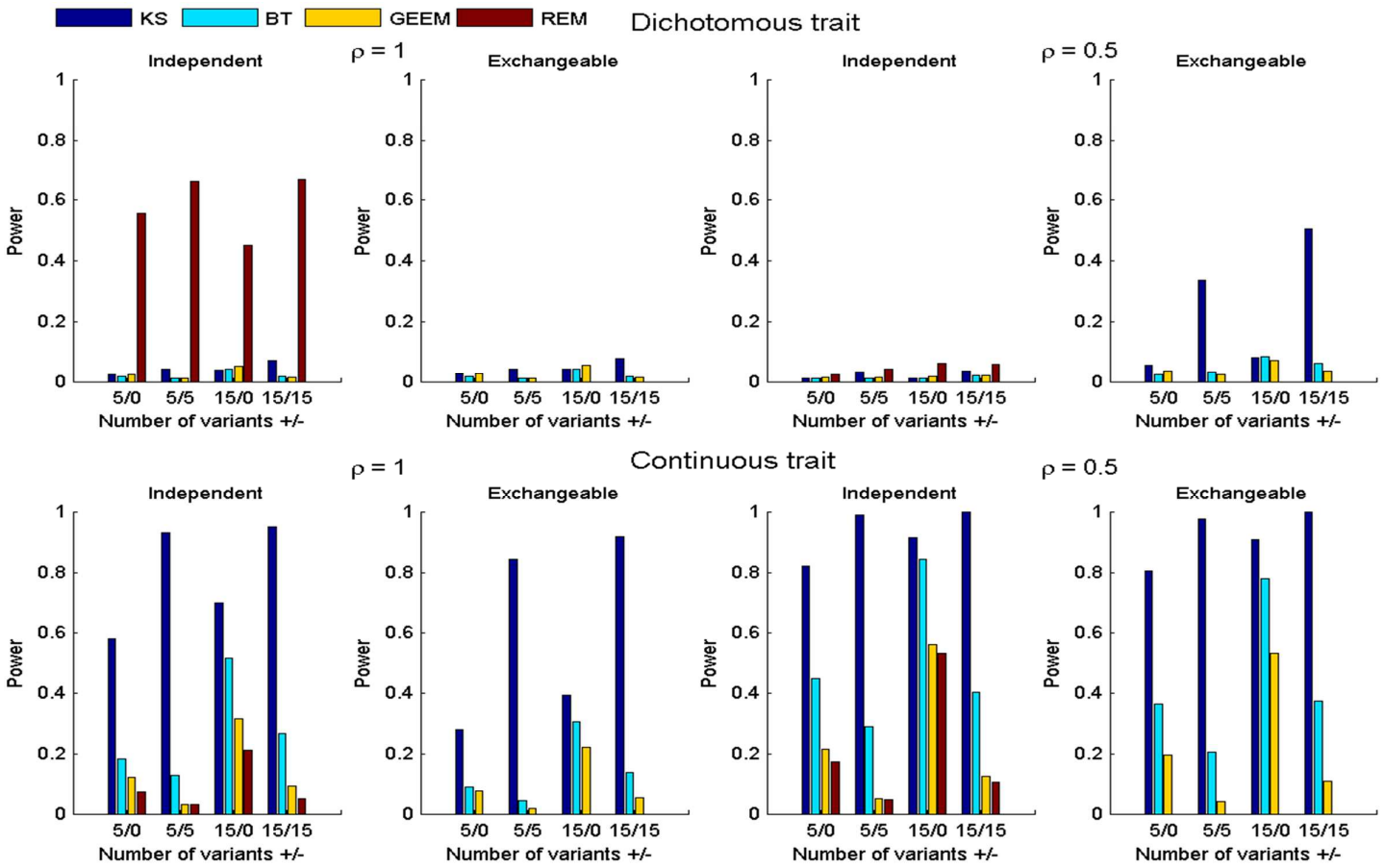

Figure S5. Powers for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 225 and number of variants (p) is 100 at a nominal level of 0.01. A working correlation matrix does not need to be specified for the REM.

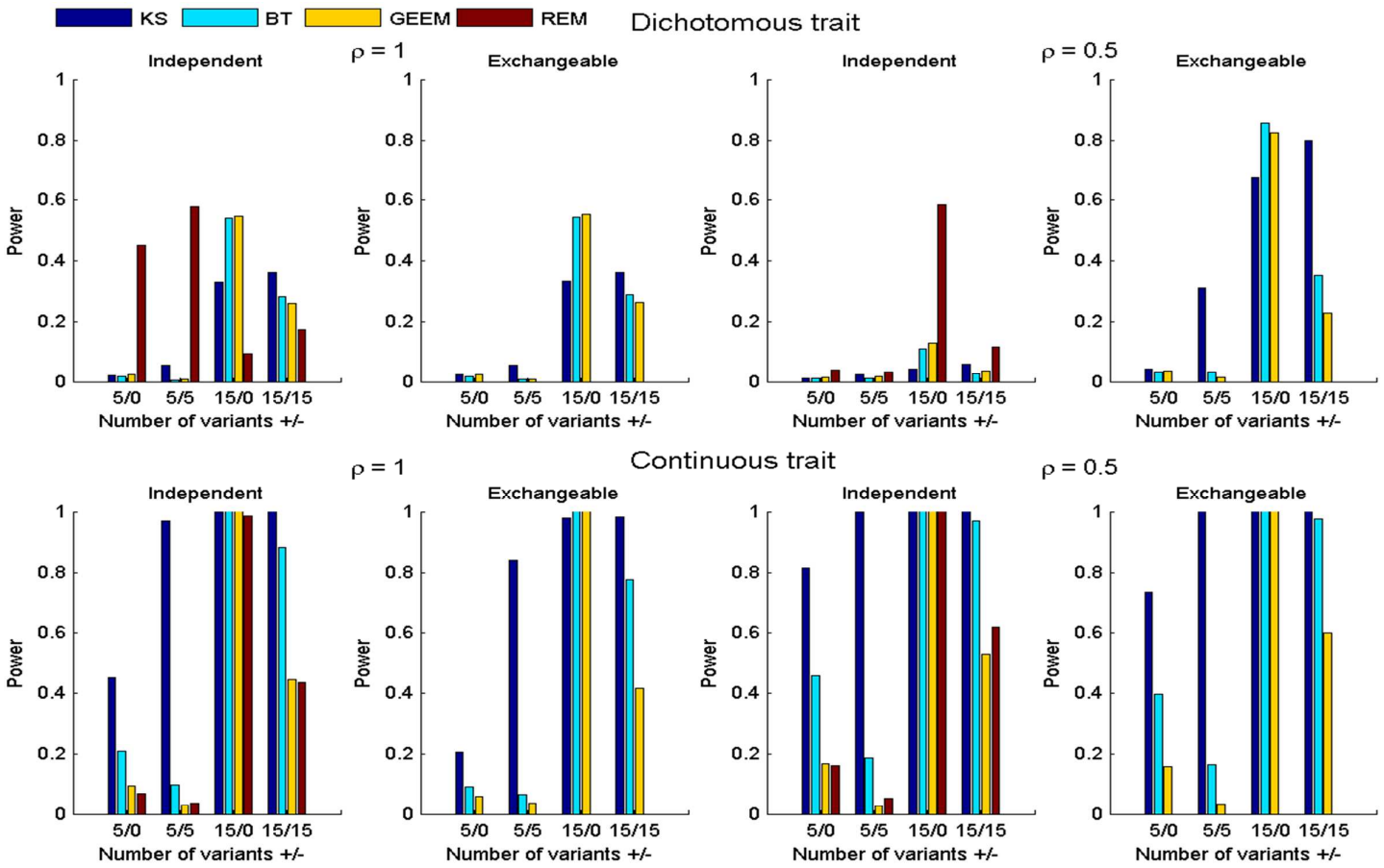

Figure S6. Powers for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 225 and number of variants (p) is 50 at a nominal level of 0.01. A working correlation matrix does not need to be specified for the REM.

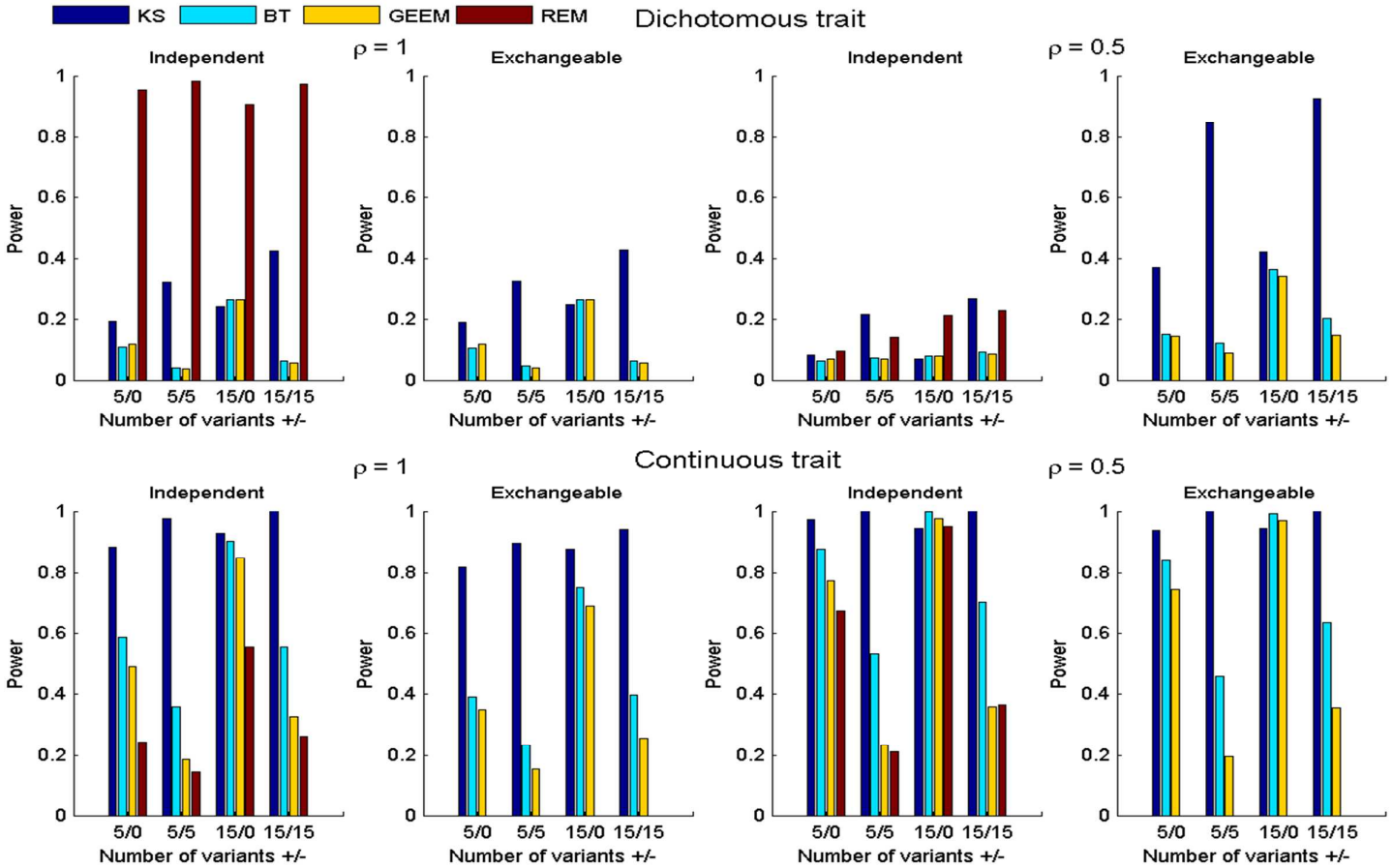

Figure S7. Powers for the X chromosome for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 450 and number of variants (p) is 100 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.

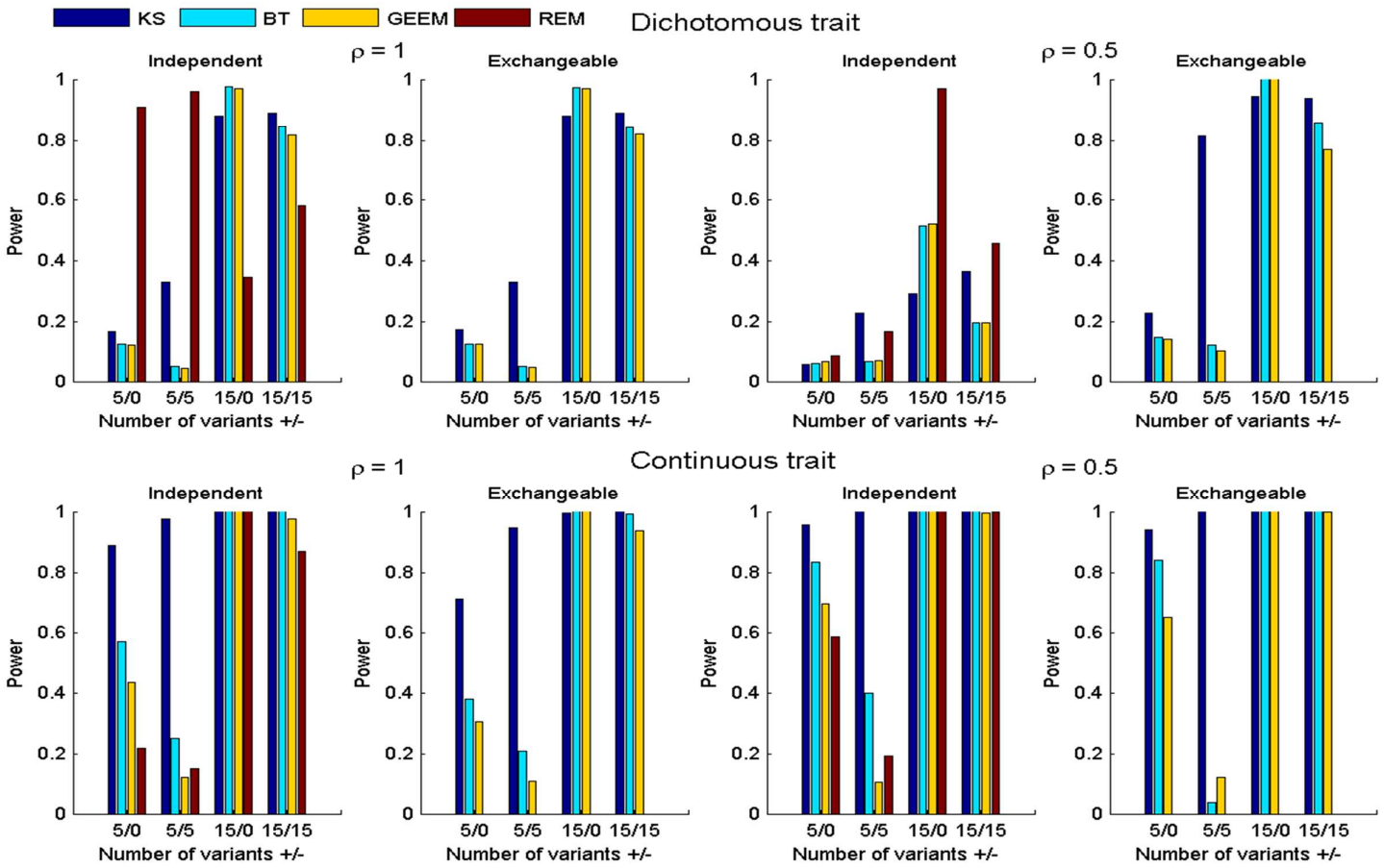

Figure S8. Powers for the X chromosome for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 450 and number of variants (p) is 50 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.

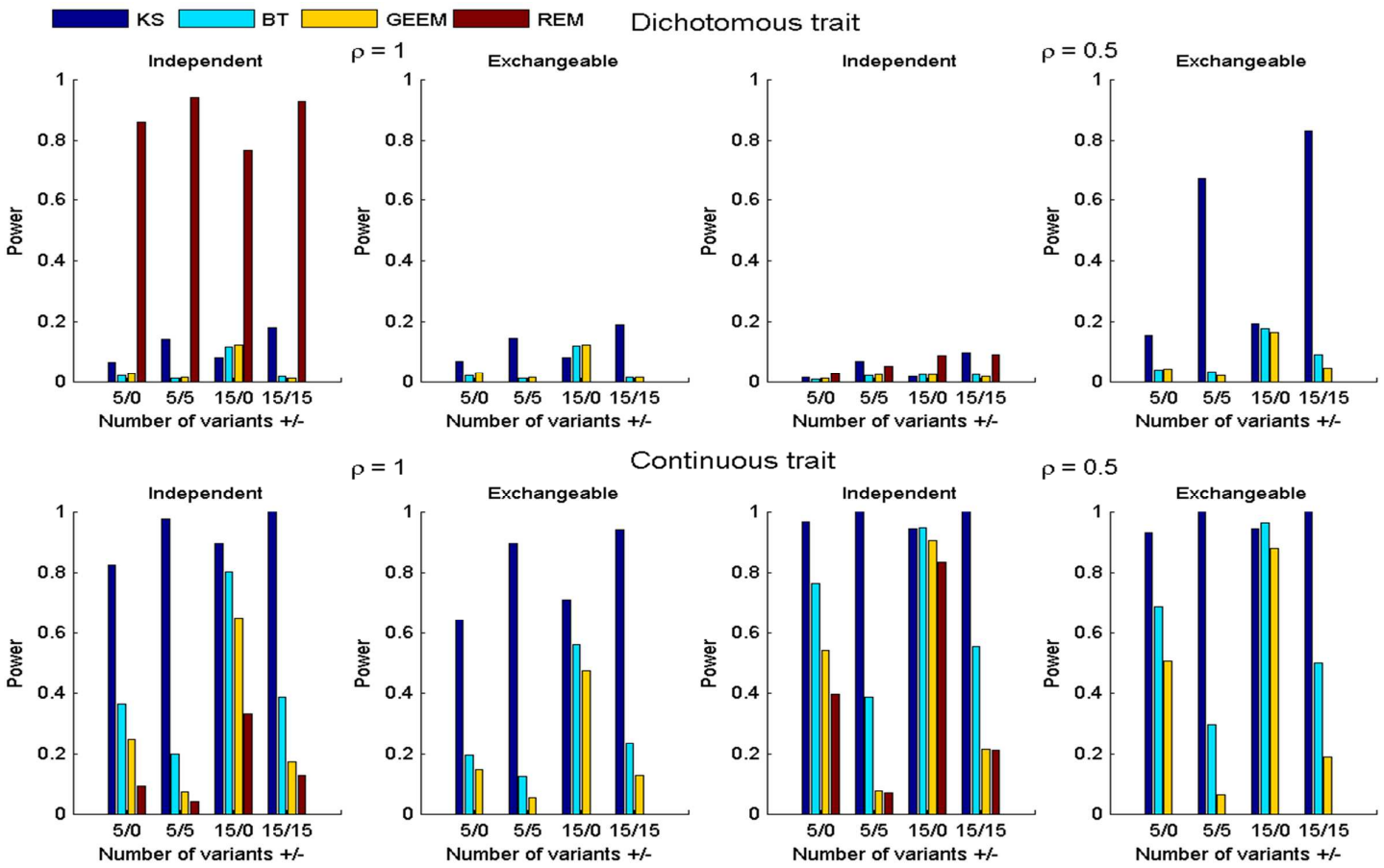

Figure S9. Powers for the X chromosome for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 450 and number of variants (p) is 100 at a nominal level of 0.01. A working correlation matrix does not need to be specified for the REM.

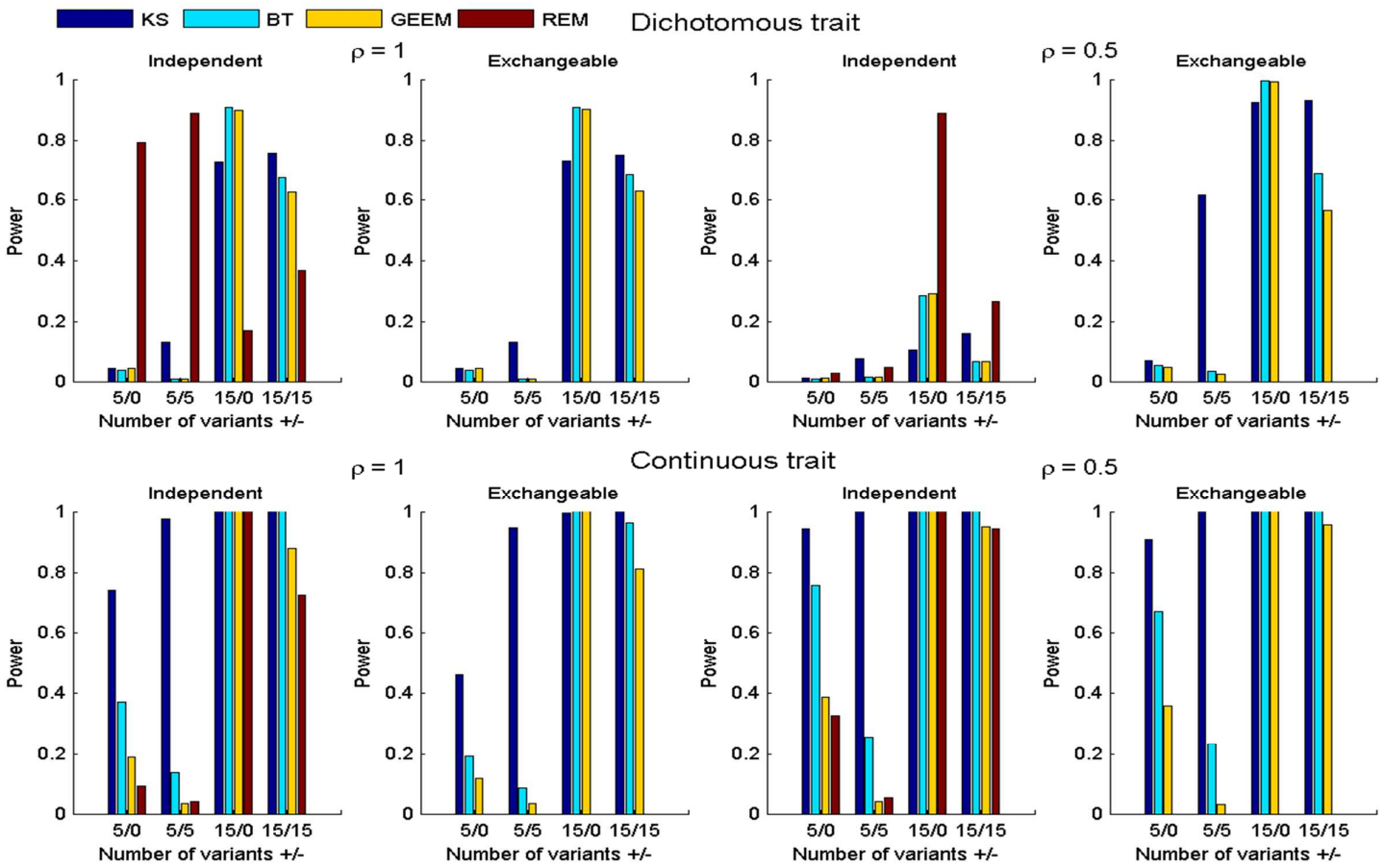

Figure S10. Powers for the X chromosome for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 450 and number of variants (p) is 50 at a nominal level of 0.01. A working correlation matrix does not need to be specified for the REM.

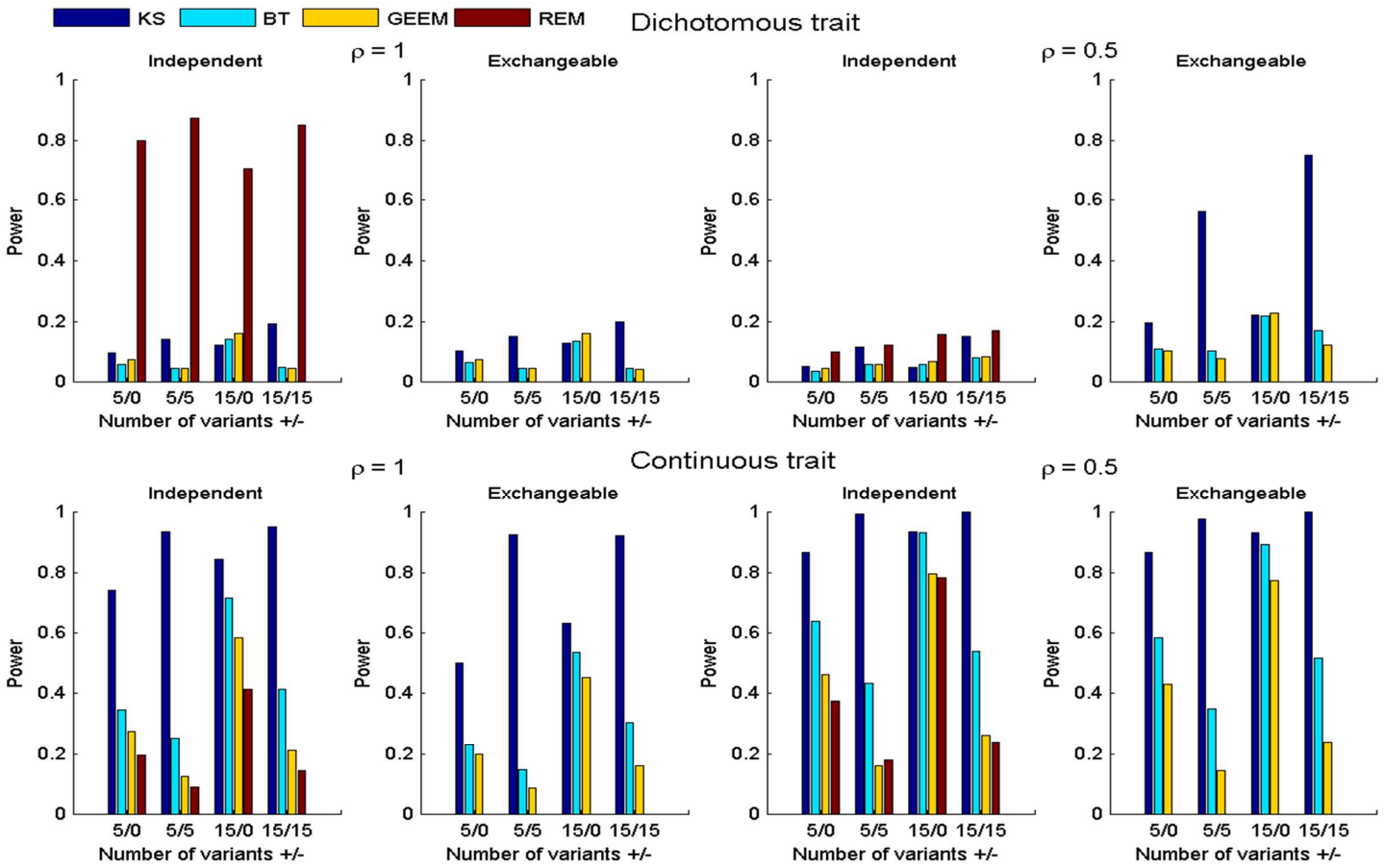

Figure S11. Powers for the X chromosome for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 225 and number of variants (p) is 100 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.

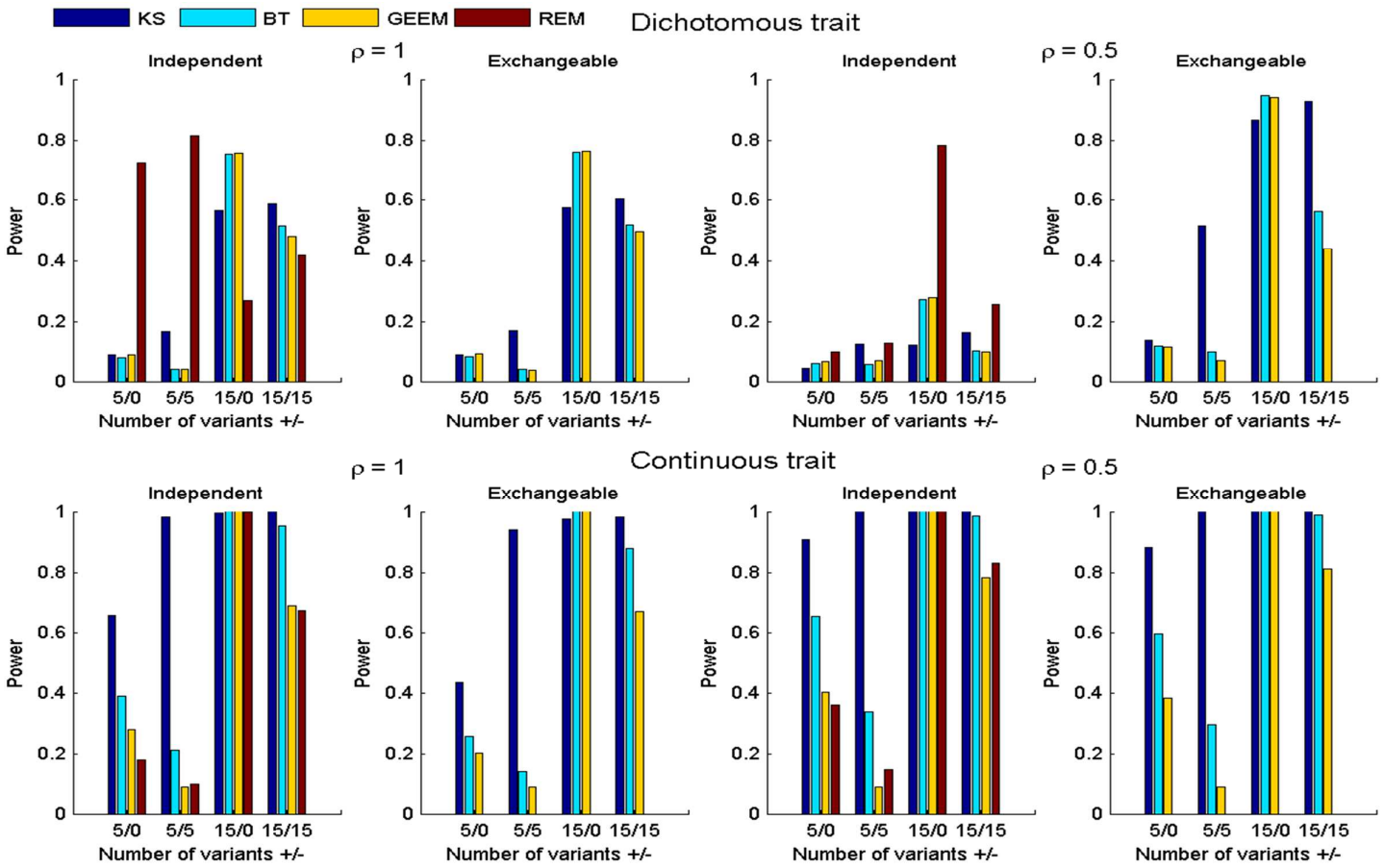

Figure S12. Powers for the X chromosome for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 225 and number of variants (p) is 50 at a nominal level of 0.05. A working correlation matrix does not need to be specified for the REM.

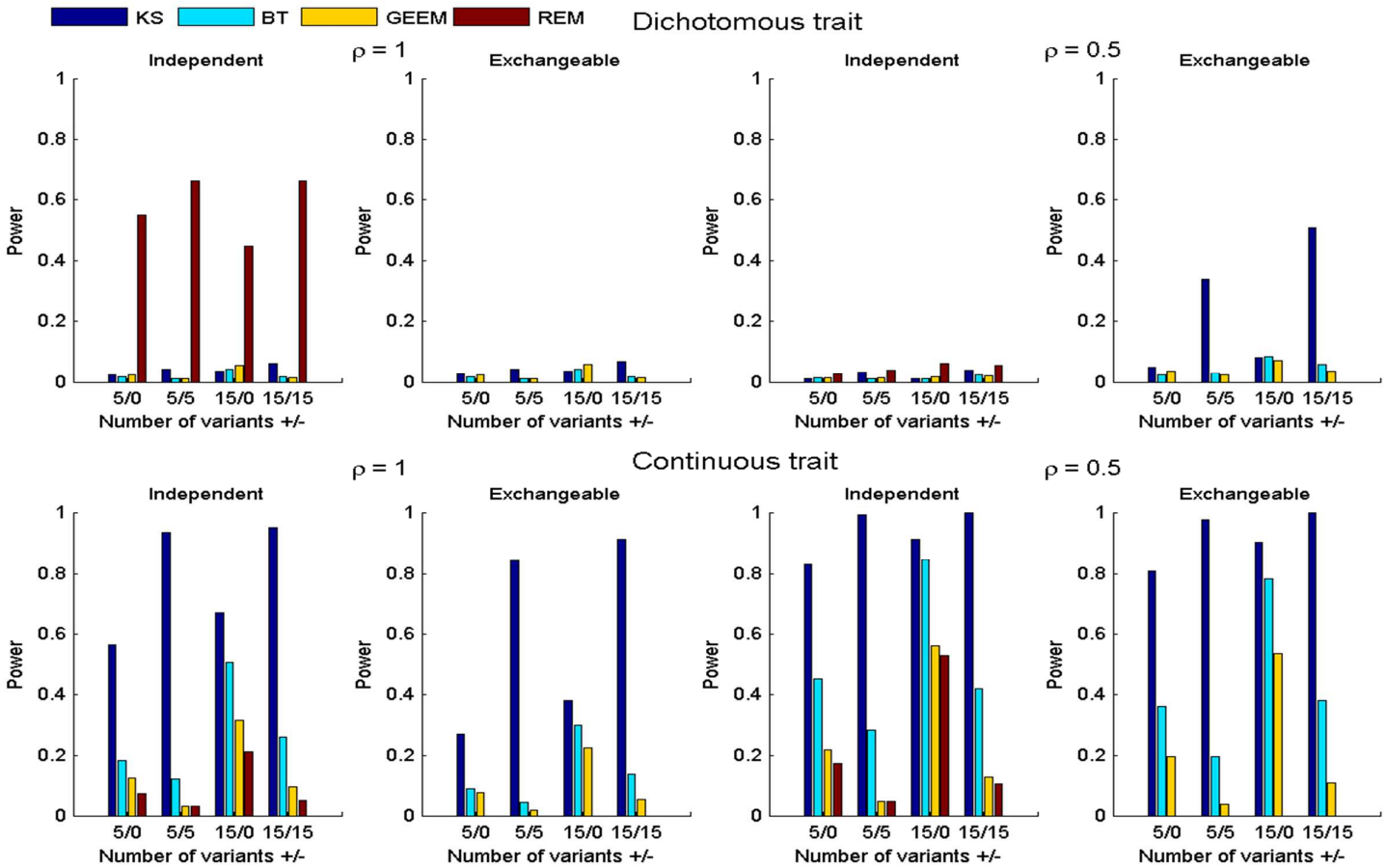

Figure S13. Powers for the X chromosome for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 225 and number of variants (p) is 100 at a nominal level of 0.01. A working correlation matrix does not need to be specified for the REM.

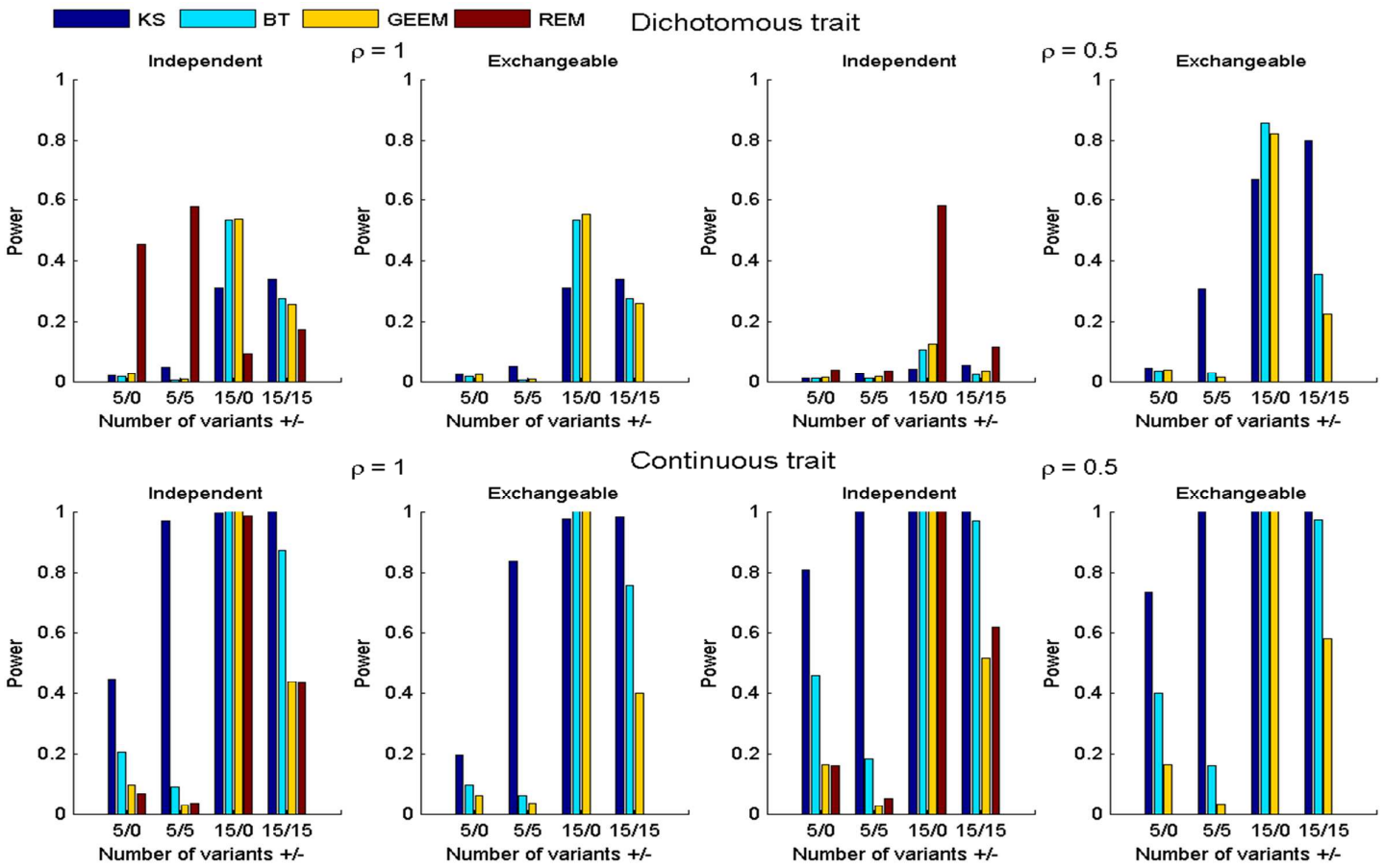

Figure S14. Powers for the X chromosome for the KS, BT, GEEM, and REM without population stratification by the specified working correlations. Number of families (K) is 225 and number of variants (p) is 50 at a nominal level of 0.01. A working correlation matrix does not need to be specified for the REM.