Biometrika (2019), **106**, 3, *pp.* 651–664
Printed in Great Britain
doi: 10.1093/biomet/asz033
Advance Access publication 13 July 2019

Statistical inference of genetic pathway analysis in high dimensions

BY YANG LIU

Department of Mathematics and Statistics, Wright State University, 3640 Colonel Glenn Highway, Dayton, Ohio 45435, U.S.A.
yang.liu@wright.edu

WEI SUN, ALEXANDER P. REINER, CHARLES KOOPERBERG AND QIANCHUAN HE

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, U.S.A.
wsun@fredhutch.org apreiner@u.washington.edu clk@fredhutch.org qhe@fredhutch.org

SUMMARY

Genetic pathway analysis has become an important tool for investigating the association between a group of genetic variants and traits. With dense genotyping and extensive imputation, the number of genetic variants in biological pathways has increased considerably and sometimes exceeds the sample size n. Conducting genetic pathway analysis and statistical inference in such settings is challenging. We introduce an approach that can handle pathways whose dimension p could be greater than n. Our method can be used to detect pathways that have nonsparse weak signals, as well as pathways that have sparse but stronger signals. We establish the asymptotic distribution for the proposed statistic and conduct theoretical analysis on its power. Simulation studies show that our test has correct Type I error control and is more powerful than existing approaches. An application to a genome-wide association study of high-density lipoproteins demonstrates the proposed approach.

Some key words: Genetic pathway analysis; Genetic variant; High-dimensional inference; Nonsparse signal; Power analysis; Sparse signal.

1. INTRODUCTION

Genetic association analysis plays an important role in identifying genetic variants that are associated with traits. Genetic variants are often analysed by single-variant-based methods, using approaches such as Armitage's trend test. Pathway-based analysis has become a popular tool for analysing genetic variant data (Chen et al., 2011b), whereby multiple genetic variants in the genes in a prespecified pathway are examined. There are several reasons to consider pathway analysis for association studies. First, pathways are generally defined using biological knowledge and thus are more likely to be functionally relevant (Zhong et al., 2010). Second, by analysing multiple variants simultaneously, pathway analysis has the potential to accumulate weak signals into stronger ones, while single-variant-based methods lack power in such a situation. Third, because the number of pathways is much smaller than the number of variants, the multiple-testing burden can be dramatically reduced.

One of the main challenges in pathway analysis is to deal with the high dimensionality. With increasingly dense genotyping and extensive imputation, the number of variants p in genetic pathways has grown so rapidly that it can be larger than the sample size n. This is seen in our real-data example, where the sample size is around 4000 while the number of single nucleotide polymorphisms in a pathway can be as large as 25 000. In such high dimensions, statistical testing methods that were designed for moderate p, such as the likelihood ratio test, tend to have low power or may be inapplicable. To deal with the high dimensionality in pathway analysis, one potential approach is the burden test (Morgenthaler & Thilly, 2007), in which one simply sums the genotypes into a single predictor and then subjects this predictor to regression analysis. The burden test works well if all the variants have similar effect sizes, but this assumption rarely holds in real situations. Another common approach to dealing with high dimensions is to use principal component analysis in the regression modelling. One first derives the principal components from the genetic pathway under consideration, and then uses the leading components for association analysis (Buas et al., 2017). The disadvantages of this approach are that principal components with large variations need not be associated with the traits; it is rarely clear how many principal components to include; the interpretation of the regression coefficients can be difficult; and when $p \to \infty$, the estimated principal components may not be consistent (Shen et al., 2016). Complementary to the aforementioned approaches, kernel machine methods such as the sequence kernel association test (Wu et al., 2011) can also be applied to genetic pathway analysis. However, the latter test has been used primarily to analyse moderate-sized variant sets, and its performance in cases where p is substantially larger is unclear. Other methods that have been developed for testing a group of genetic features in high-dimensional settings (Chen & Qin, 2010; Chen et al., 2011a; Gregory et al., 2015) focus on testing the mean difference between two groups rather than conducting association analysis.

In addition to the high-dimensional challenge, another difficulty in pathway analysis is power maximization under multiple plausible alternative hypotheses. For pathway analysis, the alternative hypothesis concerns both the number and the magnitudes of the nonzero genetic signals, which are generally unknown (Zhang, 2015). A situation often considered for genetic signals is that a pathway harbours potentially many variants with weak effects, called the nonsparse-signal situation. The sequence kernel association test can aggregate multiple signals and is potentially applicable to such a setting. Another possibility is that a genetic pathway contains only a few strong signals, called the sparse-signal situation. Several methods have been proposed to deal with this case, such as the $P_{\min}$ test (Conneely & Boehnke, 2007), which first examines each variant individually and then seeks to obtain the p-value for the maximum of the observed statistics. However, the $P_{\min}$ test has little power in the nonsparse situation, while the sequence kernel association test loses power in the sparse situation.

In this paper, we propose a method for conducting high-dimensional genetic pathway analysis, where the dimension p of the pathway can go to infinity and could exceed the sample size n. Our approach can be used to identify pathways that harbour a large number of weak signals, i.e., nonsparse signals, as well as genetic pathways that contain only a few strong signals, i.e., sparse signals, or a mixture of weak and strong signals. We establish the asymptotic properties of the proposed statistics in high dimensions and conduct theoretical analysis of their power.

2. METHODS

2.1. *Model and statistics*

Suppose that the data consist of a continuous trait vector $y_{n\times 1} = (y_1, \ldots, y_n)^{\mathrm{T}}$, an adjusting covariates matrix $X_{n\times d} = (X_1, \ldots, X_d)$ and a genotype matrix $G_{n\times p} = (G_1, \ldots, G_p)$ for a genetic

pathway; that is, the pathway being considered contains p genetic variants. Suppose that the true regression model is

$$y = X\alpha + G\beta + \varepsilon,$$

where $\alpha_{d\times 1} = (\alpha_1, \ldots, \alpha_d)^{\mathrm{T}}$ is the coefficient vector for X, with α_1 being the intercept, $\beta_{p\times 1} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the coefficient vector for G, and $\varepsilon_{n\times 1} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is a vector of independent Gaussian errors with mean zero and variance σ^2. The design matrices X and G are considered fixed. The dimension d of the adjusting covariates is assumed to be finite, while the dimension p of the genotype matrix can go to infinity.

We are interested in testing the global null hypothesis $H_0 : \beta = 0$ against the alternative $H_{\mathrm{a}} : \beta \neq 0$. Tests such as the likelihood ratio test and Wald test consider all the p variants jointly and tend to perform poorly when p is large; the statistics may not exist when $p > n$. Marginal statistics are easy to calculate and have been widely used to evaluate the significance of each individual variant. Recall that in a marginal analysis, one first fits a regression model for a given variant, say the jth, by $y = X\alpha + G_j\beta_j + \varepsilon$ $(j = 1, \ldots, p)$ and then obtains the marginal score statistic as

$$b_j = (G_j^{\mathrm{T}} P_X G_j)^{-1/2} G_j^{\mathrm{T}} P_X y$$

with $P_X = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, where I_n is the identity matrix. To conduct a pathway analysis, it is natural to consider the sum of all the squared marginal statistics, $Q_0 = \sum_{j=1}^{p} b_j^2$. In fact, it can be shown that Q_0 is equivalent to the sequence kernel association test statistic, if the estimator $\hat{\sigma}^2$ of σ^2 is ignored in the latter. However, our proposed approach is not focused on Q_0 per se, but rather uses Q_0 to develop a suite of statistics for high-dimensional settings, particularly for the case of $p/n \to \gamma \in (0, \infty)$ for a constant γ.

Under the null hypothesis H_0, it can be shown that $E(Q_0) = p\sigma^2$ and $\mathrm{var}(Q_0) = 2\sigma^4\|A\|_{\mathrm{F}}^2$, where $A = P_X G D^{-1} G^{\mathrm{T}} P_X$, with D a diagonal matrix whose elements are $G_j^{\mathrm{T}} P_X G_j$ $(j = 1, \ldots, p)$, and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. For the moment we assume that σ^2 is known, but later on we will address the practical situation where σ^2 needs to be estimated. We propose to standardize Q_0, which yields

$$T^{*(n,p)} = \frac{\sum_{j=1}^{p} b_j^2 - p\sigma^2}{\sqrt{2\sigma^2}\|A\|_{\mathrm{F}}}, \tag{1}$$

where the superscript (n, p) emphasizes that both n and p can go to infinity; it will be suppressed below for ease of notation. Expression (1) suggests that T^* may converge to normality as p gets large. However, the central limit theorem does not directly apply here because the b_j^2 are correlated. In fact, the correlation matrix for the b_j, Σ, can be shown to have the form

$$\Sigma = D^{-1/2} G^{\mathrm{T}} P_X G D^{-1/2},$$

and it can further be shown that $\|\Sigma\|_{\mathrm{F}} = \|A\|_{\mathrm{F}}$. In Lemma 1 we show that under proper conditions, T^* is standard normal as both n and p go to infinity.

Before presenting Lemma 1, we define some notation. For a vector $a = (a_1, \ldots, a_n)$, let $\|a\|_k = \left(\sum_{i=1}^{n} |a_i|^k\right)^{1/k}$ be the k-norm of the vector for $k = 1, 2, \ldots, \infty$. For any $m \times l$ matrix $M = (m_{ij})_{i=1,\ldots,m;\, j=1,\ldots,l}$, denote the induced k-norm by $\|M\|_k = \sup_{\|x\|_k=1} \|Mx\|_k$. When M is

an $m \times m$ matrix, we denote its maximum and minimum eigenvalues by $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$, respectively.

LEMMA 1. *Let $n, p \to \infty$. If*

$$\|\Sigma\|_2 = o(p^{1/2}) \quad \text{as } p \to \infty, \tag{2}$$

then under H_0, the statistic $T^* = (\sum_{j=1}^p b_j^2 - p\sigma^2)/(\sqrt{2\sigma^2}\|A\|_{\mathrm{F}}) \to N(0,1)$ in distribution.

Remark 1. Here we have no constraint on the order of p with respect to n, providing they both go to infinity. Condition (2) is mild for genetic studies. By Hölder's inequality, $\|\Sigma\|_2 \leqslant (\|\Sigma\|_1\|\Sigma\|_\infty)^{1/2} = \|\Sigma\|_1$, where $\|\Sigma\|_1$ is the maximum absolute column sum of the matrix. When the correlation structure in Σ is not overly strong, as is the case for the power-decay structure, i.e., $\Sigma_{jk} = O(\rho^{|j-k|})$ for some $\rho < 1$, then one can show that $\|\Sigma\|_1 = o(p^{1/2})$. Here Σ_{jk}, the correlation of b_j and b_k, can be interpreted as the linkage disequilibrium of genetic variants after adjusting for covariates X; when there are no adjusting covariates, Σ reduces to the linkage disequilibrium matrix of G. The power decay structure indicates that two distant genetic variants have virtually no linkage disequilibrium, which is indeed what is observed in genetic studies, particularly in the human genome data (International HapMap Consortium, 2005). Similar structures have also been used in other articles on genetic studies, such as Dai et al. (2012). Our proposed statistic naturally takes linkage disequilibrium into account, because $\|A\|_{\mathrm{F}} = \|\Sigma\|_{\mathrm{F}}$. The linkage disequilibrium can influence both the denominator and the numerator of T^*, so the impact of the linkage disequilibrium on the power of the proposed test is influenced by the size and density of the genetic signals. However, the linkage disequilibrium will not affect the validity of the test or its asymptotic properties, because the calculation of $\|\Sigma\|_{\mathrm{F}}$ does not involve inversion of the linkage disequilibrium matrix, and the normality of the proposed statistics requires only that distant variants tend to have linkage disequilibrium approaching zero. In practice, variants in a gene tend to be in linkage disequilibrium, while those for different genes are generally not in linkage disequilibrium; this type of structure is covered in Lemma 1.

So far we have assumed that the noise level σ^2 is known. To make our proposal practical, it is tempting to replace σ^2 with a consistent estimator $\hat{\sigma}^2$. It turns out that the validity of doing so depends on the order of p relative to n. In the following, we elaborate on this and propose different statistics to accommodate different ratios p/n.

We first consider the situation where $p = O(n^{1-\xi})$ for some $0 < \xi < 1$, i.e., p is of smaller order than n. The following lemma shows that if we replace σ^2 with a consistent estimator $\hat{\sigma}^2$, normality still holds.

LEMMA 2. *Suppose that (2) holds. Let $\hat{\sigma}^2$ be a root-n-consistent estimator of σ^2 such that $\hat{\sigma}^2 = \sigma^2 + O_{\mathrm{p}}(n^{-1/2})$. Then under H_0, as $n, p \to \infty$ such that $p = O(n^{1-\xi})$ for $0 < \xi < 1$,*

$$T_{\mathrm{L}} = \frac{\sum_{j=1}^p b_j^2 - p\hat{\sigma}^2}{\sqrt{2\hat{\sigma}^2}\|A\|_{\mathrm{F}}} \to N(0,1)$$

in distribution.

Next, we consider the situation in which $p/n \to \gamma \in (0, \infty)$ for some constant γ. The normality of T_{L} no longer holds because $p(\hat{\sigma}^2 - \sigma^2)$ becomes excessively large; see the proof of

Lemma 2 for more details. In light of this, we propose a new statistic

$$T_{\mathrm{H}} = \frac{\sum_{j=1}^{p} b_j^2 - (n-d)^{-1} p y^{\mathrm{T}} P_X y}{\sqrt{2\hat{\sigma}^2}\|A_1\|_{\mathrm{F}}}, \tag{3}$$

where $A_1 = A - (n-d)^{-1} p P_X$, with d being the number of adjusting covariates as mentioned earlier. The $p\hat{\sigma}^2$ in the numerator of T_{L} is replaced by $(n-d)^{-1} p y^{\mathrm{T}} P_X y$ in T_{H}. The motivation behind this is that $(n-d)^{-1} y^{\mathrm{T}} P_X y$ estimates σ^2 under H_0. We discovered that this replacement of $\hat{\sigma}^2$ enables one to overcome the limitation of T_{L} in high dimensions. The following theorem shows that T_{H} follows a normal distribution for $p/n \to \gamma \in (0, \infty)$.

THEOREM 1. *Suppose that* (2) *holds. For any consistent estimator* $\hat{\sigma}^2$*, as* $n, p \to \infty$ *such that* $p/n \to \gamma \in (0, \infty)$*, if* $p^{-1}\|\Sigma\|_{\mathrm{F}}^2 \geqslant \gamma + \eta$ *for some constant* $\eta > 0$*, then under* H_0 *we have* $T_{\mathrm{H}} \to N(0, 1)$ *in distribution.*

Theorem 1 allows one to conduct statistical inference for pathway analysis when $p > n$, although p should not be excessively larger than n. The condition $p^{-1}\|\Sigma\|_{\mathrm{F}}^2 \geqslant \gamma + \eta$ is necessary to prevent the $\|A_1\|_{\mathrm{F}}$ in the denominator equalling zero, as it can be shown that $\|A_1\|_{\mathrm{F}}^2 = \|\Sigma\|_{\mathrm{F}}^2 - (n-d)^{-1} p^2$. To obtain a consistent estimator for σ^2 under H_{a} in a high-dimensional setting, Fan et al. (2012) proposed a refitted crossvalidation method based on procedures that satisfy the sure screening property. When the sparsity of the model is completely unknown, we can also estimate σ^2 by the moment-based estimators of Dicker (2014), which are root-n consistent when $p > n$.

2.2. *Power loss in the presence of sparse signals*

The proposed statistic T_{H} can handle situations where the association signals are spread out over a large number of genetic variants. However, the power of T_{H} will be relatively low for the sparse-signal situation, in which a few genetic variants carry strong signals while all the others have zero coefficients. Fan et al. (2015) proposed the power-enhancement principle, the fundamental idea of which is to include a screening statistic that goes to zero under H_0, but is nonzero under the sparse alternatives H_{a}. Motivated by this principle, we propose a statistic that strengthens T_{H} and is able to guard against potential power loss in the sparse-signal situation.

We define a screening set $\hat{S} = \{j : |b_j|/\hat{\sigma} > \delta_p\}$, where δ_p is a threshold chosen to be slightly larger than the maximum estimation error of the marginal estimator, i.e., $\max_j |b_j - E(b_j)|/\hat{\sigma}$. Then, a power-enhancement component T_0 is

$$T_0 = \mathrm{sgn}(T_{\mathrm{H}})\, p^{1/2} \sum_{j \in \hat{S}} b_j^2/\hat{\sigma}^2,$$

where $\mathrm{sgn}(T_{\mathrm{H}})$ denotes the sign of T_{H}. Our statistic that is able to detect both nonsparse and sparse signals is $T = T_{\mathrm{H}} + T_0$.

Since T_0 has the same sign as T_{H}, T always has power at least that of T_{H}. The threshold δ_p needs to ensure that the screening set $\hat{S}$ is empty with probability approaching 1 under H_0, so that the size of T will be asymptotically equivalent to that of T_{H}. Then, under H_{a}, if an estimator is large enough that $\hat{S}$ is nonempty, one can gain power. For Gaussian and sub-Gaussian errors, δ_p can be chosen to be $\log\log n (\log p)^{1/2}$, as suggested by Fan et al. (2015). The power-enhancement procedure in Fan et al. (2015) deals with a consistent estimator under H_{a}, which is not available in our procedure, while our approach builds upon marginal estimators which are inconsistent under H_{a}. Nevertheless, the size of our proposed statistic is asymptotically equivalent to that of

T_{H} under H_0; in the next subsection, we will show that under the sparse alternatives T can be powerful even when T_{H} is not.

LEMMA 3. *Under the same conditions as in Theorem 1, if* $\delta_p = a_p(\log p)^{1/2}$ *where* $a_p \to \infty$ *as* $p \to \infty$, *then under the null hypothesis* H_0 *we have* $T \to N(0,1)$ *in distribution. Thus, the sizes of* T *and* T_{H} *are asymptotically equivalent.*

To select δ_p in practice, we propose an adaptive procedure to accommodate different correlation structures. We first generate a vector of n random errors ε^* from the standard normal distribution. Then we compute the maximum of the marginal estimators as $\theta^* = \max_{1 \leqslant j \leqslant p} |(G_j^{\mathrm{T}} P_X G_j)^{-1/2} G_j^{\mathrm{T}} P_X \varepsilon^*|$. Finally, we repeat these two steps many times and set $\delta_p = \max \theta^*$ based on all the replicates. McKeague & Qian (2015) also used an adaptive approach to determine threshold parameters for high-dimensional testing.

2.3. *Power analysis*

In this subsection we investigate the asymptotic power of the proposed tests T_{H} and T for nonsparse and sparse alternatives. Under H_{a}, let $S = \{j : \beta_j \neq 0\}$ be the set of nonzero coefficients, and let $s = |S|$. Define the subvector $\beta_S = \{\beta_j : j \in S\}$ and the submatrix $G_S = \{G_j : j \in S\}$. Let D_S be the diagonal matrix with nonzero elements $G_j^{\mathrm{T}} P_X G_j$ for $j \in S$. Similarly, let $\beta_{S^{\mathrm{c}}}$, $G_{S^{\mathrm{c}}}$ and $D_{S^{\mathrm{c}}}$ denote the corresponding quantities for $S^{\mathrm{c}} = \{j : \beta_j = 0\}$.

The following theorem states that the sum-of-squares type of statistic T_{H} has high power for the nonsparse-signal situation when the accumulated signals are sufficiently large.

THEOREM 2. *Suppose that all the conditions in Theorem 1 hold. Consider a nonsparse alternative* $H_{\mathrm{a}}^{(\mathrm{ns})}$ *in which* $\|\beta\|_2 > c_1(p \log p/n)^{1/2}$ *for a sufficiently large constant* c_1. *If* $\lambda_{\min}(n^{-1} G_S^{\mathrm{T}} A_1 G_S) \geqslant c_2$ *and* $\lambda_{\max}(n^{-1} G_S^{\mathrm{T}} G_S) \leqslant c_3$ *for some constants* $c_2, c_3 > 0$, *then as* $n, p \to \infty$, $\mathrm{pr}(|T_{\mathrm{H}}| > q_{1-\zeta/2}) \to 1$, *where* $q_{1-\zeta/2}$ *is the* $(1-\zeta/2)$*-quantile of the standard normal distribution.*

While T_{H} can have high power under nonsparse alternatives, it may lose power under sparse alternatives. In the following theorem we show that T, which adds a power-enhancement term T_0 to T_{H}, can be powerful under both nonsparse and sparse alternatives.

THEOREM 3. *Assume that the conditions in Theorem 2 hold. Consider a sparse alternative* $H_{\mathrm{a}}^{(\mathrm{s})}$ *in which* $\max_{j \in S} (G_j^{\mathrm{T}} P_X G_j)^{1/2} |\beta_j| > c_4 s^{1/2} \delta_p$ *for a sufficiently large constant* c_4. *If* $\lambda_{\min}(D_S^{-1} G_S^{\mathrm{T}} P_X G_S) \geqslant c_5$ *for some constant* $c_5 > 0$, *then under either the nonsparse alternative* $H_{\mathrm{a}}^{(\mathrm{ns})}$ *or the sparse alternative* $H_{\mathrm{a}}^{(\mathrm{s})}$, *as* $n, p \to \infty$, $\mathrm{pr}(|T| > q_{1-\zeta/2}) \to 1$.

In practice, we recommend use of T for detecting both weak and strong signals. However, if one wishes to distinguish between the sparse signals and the nonsparse signals, one can examine the values of T_{H} and T. If $|T|$ is larger than $|T_{\mathrm{H}}|$, then the power-enhancement component T_0 is nonzero and there exist strong signals in the pathway. If $T = T_{\mathrm{H}}$, then there are no strong signals in the pathway and the significance is driven by weak signals.

2.4. *Incorporating biological information into* T_{H} *and* T

The statistics T_{H} and T give equal weight to all the variants. In some applications, one may wish to assign different weights based on prior information. For example, if the effect of a genetic variant is related to its minor allele frequency, one may assign a weight $w_j = 1/\{m_j(1-m_j)\}$

to this variant, where m_j is the minor allele frequency for the jth variant. In other cases, one may assign functional scores to different variants to reflect their biological functions. In lieu of these considerations, we propose incorporating prior biological information into our proposed statistics as follows.

Let w_j $(j = 1, \ldots, p)$ be prespecified positive weights, and let D_w be the diagonal matrix with elements w_j. Next, define $Q_0(w) = \sum_{j=1}^{p} w_j b_j^2$ and $Q_1(w) = Q_0(w) - (n-d)^{-1}(\sum_{j=1}^{p} w_j) y^{\mathrm{T}} P_X y$. Let $A(w) = P_X G D_w D^{-1} G^{\mathrm{T}} P_X$ and $A_1(w) = A(w) - (n-d)^{-1} \sum_{j=1}^{p} w_j P_X$. Similar to T_{H}, we define a statistic $T_{\mathrm{H}}(w) = Q_1(w)/\{\sqrt{2\hat{\sigma}^2}\|A_1(w)\|_{\mathrm{F}}\}$. The following result shows the asymptotic normality of $T_{\mathrm{H}}(w)$.

COROLLARY 1. *Suppose that* (2) *holds and* $\|\Sigma\|_2 \max_{j=1,\ldots,p} w_j/\|w\|_2 \to 0$ *as* $p \to \infty$. *Assume that as* $n, p \to \infty$, $p/n \to \gamma \in (0, \infty)$. *For any consistent estimator* $\hat{\sigma}^2$, *if* $\|w\|_2^{-1}\|D_w \Sigma\|_{\mathrm{F}}^2 \geqslant \gamma + \eta_w$ *for some constant* $\eta_w > 0$, *then under* H_0, $T_{\mathrm{H}}(w) \to N(0,1)$ *in distribution.*

As was done for T_{H}, we can add T_0 to $T_{\mathrm{H}}(w)$ to guard against potential power loss in the presence of strong signals. Thus, our proposed statistics can readily accommodate prior biological information and still preserve their theoretical properties.

2.5. *Edgeworth expansion for extreme significance levels*

Genetic studies sometimes involve a large number of pathways, so the significance level can be much lower than 0.05. For example, in our real-data analysis, the significance level is 0.0003. At such levels, the normal distribution in Lemma 3 may be a poor approximation. We therefore propose a two-term Edgeworth expansion to characterize the tail probability of T_{H} with higher accuracy. Recall that under H_0, $T_{\mathrm{H}} = \varepsilon^{\mathrm{T}} A_1 \varepsilon/(\sqrt{2\hat{\sigma}^2}\|A_1\|_{\mathrm{F}})$. It is known that $\varepsilon^{\mathrm{T}} A_1 \varepsilon/\sigma^2$ follows a mixed chi-squared distribution with weights $\lambda_1, \ldots, \lambda_p$, where λ_j are the eigenvalues of A_1. Using the Edgeworth expansion for independent random variables with varying distributions (Feller, 1971, p. 546), we can derive the following two-term expansion for $\mathrm{pr}\{\varepsilon^{\mathrm{T}} A_1 \varepsilon/(\sqrt{2\sigma^2}\|A_1\|_{\mathrm{F}}) \leqslant t\}$:

$$\Phi(t) + \frac{4\Lambda_3(1-t^2)}{3(2\Lambda_2)^{3/2}}\phi(t) - \left\{\frac{\Lambda_4(t^3-3t)}{96\Lambda_2^2} + \frac{\Lambda_3^2(t^5-10t^3+15t)}{9\Lambda_2^3}\right\}\phi(t) + O\left(\frac{p^3}{\Lambda_2^{9/2}}\right), \quad (4)$$

where $\Lambda_k = \sum_{j=1}^{p} \lambda_j^k$ for $k = 2, 3, 4$. Further, $\Lambda_2 = \|A_1\|_{\mathrm{F}}^2 = \|\Sigma\|_{\mathrm{F}}^2 - p^2/(n-d)$. Then, under the conditions in Theorem 1, the last remainder term in (4) can be shown to be $O(n^{-3/2})$. This expansion tends to be more accurate than the normal approximation, as the remainder term of the normal approximation is typically $O(n^{-1/2})$. Directly calculating (4) involves computing the λ_j, which can be onerous when n and p are large. Instead, we can use the identity $\Lambda_k = \mathrm{tr}(A_1^k)$ for $k = 2, 3, 4$. We call the test that uses (4) to approximate the p-value for T_{H} the $T_{\mathrm{H}}^{\mathrm{e}}$ test. Similarly, we can apply an Edgeworth expansion to T, and we call the resulting test T^{e}.

3. SIMULATION STUDIES

Monte Carlo simulations were conducted to evaluate the performance of the proposed tests, T_{H} and T, in high-dimensional settings and to compare them with the Bonferroni test, the burden test, principal component analysis, and the sequence kernel association test.

Table 1. *Type I error* (%) *of the tests at level* 0.05

n	p	Corr.	Bonf.	Burden	PCA	PCA50	SKAT	T_{H}	T
500	300	CS	4.66	4.90	5.00	5.97	2.79	4.84	5.10
		AR	4.47	5.30	4.91	6.40	2.77	5.03	5.13
	500	CS	4.80	4.92	5.06	6.59	2.04	4.89	4.97
		AR	4.64	4.80	5.45	6.59	1.64	4.76	4.84
	1000	CS	4.71	5.02	5.00	7.22	0.75	4.75	4.93
		AR	4.96	4.98	5.10	7.31	0.68	4.83	4.93
1000	800	CS	5.07	4.86	5.20	6.49	2.47	5.03	5.12
		AR	4.78	5.26	5.25	6.49	2.18	4.92	5.01
	1000	CS	4.92	5.08	4.71	6.13	2.02	4.85	5.04
		AR	5.24	5.00	4.90	6.60	1.81	4.88	4.99
	1500	CS	4.92	5.08	5.27	7.07	1.38	4.95	5.14
		AR	5.22	5.08	5.26	6.85	1.04	4.82	4.98

Corr., correlation structure; CS, compound symmetric; AR, autoregressive; Bonf., Bonferroni test; PCA, principal component analysis using the five leading components; PCA50, principal component analysis using components that explain 50% of the variance; SKAT, the sequence kernel association test.

We generated the genotype matrix G similarly to He et al. (2016). For each person, we first generated a block-diagonal covariance matrix with each block being a 5×5 matrix Ω_0. We considered compound symmetric Ω_0 with diagonal elements 1 and off-diagonal elements 0.5 and autoregressive Ω_0 with (i,j)th off-diagonal element $0.6^{|i-j|}$. Then we trichotomized the simulated vector into genotype values of $(0, 1, 2)$ according to the Hardy–Weinberg equilibrium.

We generated the trait by setting $y_i = 1 + x_i + \sum_{j=1}^{p} G_{ij}\beta_j + \varepsilon_i$ $(i = 1, \ldots, n)$, where $x_i \sim N(0.1G_{i1}, 1)$ is an adjusting covariate and $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma = 1$. For the sample size n and the dimension p, we considered both the $p > n$ and the $p < n$ cases by setting $n = 500$ with $p = 300, 500, 1000$ and setting $n = 1000$ with $p = 800, 1000, 1500$.

The null model is $\beta_j = 0$ $(j = 1, \ldots, p)$. To simulate the data under various alternatives, we assume that $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ has s nonzero signals with support set $S = \{1, 6, \ldots, 5s-4\}$. The magnitude of the signal $|\beta_j|$ was set to be $0.4(\log p/n)^{1/2}$, and half of the β_j had positive signs while the other half had negative signs. The magnitude of the signals varied from 0.03 to 0.05 under these set-ups, and the proportion of nonnull variants among all the variants, s/p, was set to 5%, 10%, 15% or 20%.

The tests T_{H} and T were conducted as described in § 2. For variance estimation, we applied the refitted crossvalidation method (Fan et al., 2012) to obtain $\hat{\sigma}^2$. The threshold parameter δ_p in the power-enhancement component T_0 was estimated over 1000 replicates. We used two versions of principal component analysis. In the first, we used the five leading principal components and performed a likelihood ratio test. In the second version, we included the principal components that explain 50% of the total variance for the likelihood ratio test. The reason for considering the second version is that, in practice, a few principal components may not always capture the majority of the variance, as seen in Avery et al. (2011). We also included the sequence kernel association test without any weights.

The Type I errors of the tests were calculated over 10 000 replications, and the power was based on 1000 replications. Table 1 displays the Type I errors of the tests for the models considered. It can be seen that the Bonferroni test, principal component analysis using the five leading components, the sequence kernel association test, and our tests T_{H} and T all have their Type I errors controlled. Principal component analysis using components that explain 50% of the variance appears to have

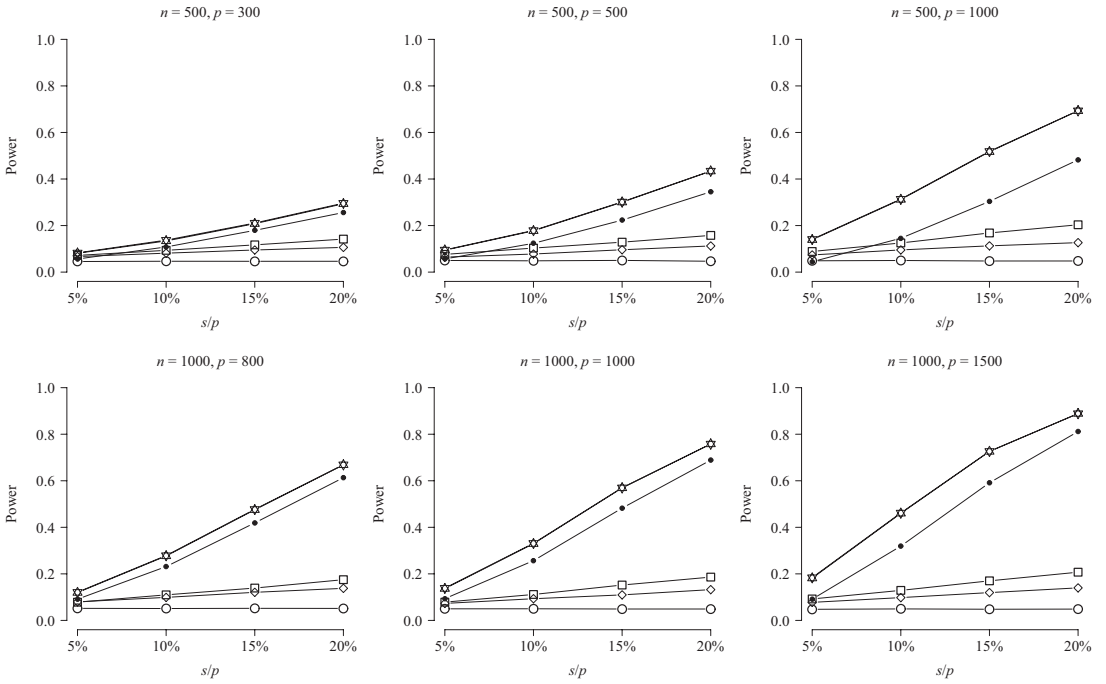

Fig. 1. Power of the Bonferroni test (◊), the burden test (○), principal component analysis (□), the sequence kernel association test (·), T_{H} (▽) and T (△) at level 0.05 for different sample sizes and dimensions plotted against the proportion of nonzero signals, s/p. The compound symmetric dependence structure is considered.

an inflated Type I error; this is likely due to the fact that many principal components are needed to account for the 50% of variance, and hence the likelihood ratio test has a large degree of freedom. Because of its inflated Type I error, this method was excluded from the subsequent experiments.

For the compound symmetric dependence structure, plots of the power for different sample sizes and dimensions against the proportion s/p of nonzero signals are shown in Fig. 1. When the ratio s/p is small, all the methods have low power. This indicates that when signals are sparse and weak, it is highly difficult to detect the association for the pathway considered. As s/p increases, the power improves for all methods, because more variants carry association signals in the studied pathway. However, the tests T_{H} and T always have higher power than the other approaches in these settings. The results for the autoregressive structure show a similar pattern.

We then considered the situation in which a genetic pathway contains both weak and strong signals. We simulated weak signals as described earlier, and then simulated a strong signal with $\beta_1 = 2(\log p/n)^{1/2}$. Table 2 shows that both T_{H} and T compete favourably with the other statistics, and T has higher power than T_{H}.

We conducted simulation studies to examine the performance of the T^{e} test, which involves a two-term Edgeworth expansion and is expected to be more accurate than T in controlling Type I error at extreme significance levels. We set the significance level to 0.0001 and evaluated the Type I error with 1 000 000 simulations. The threshold parameter δ_p in the power enhancement was estimated over 50 000 replicates. Table 3 shows that at level 0.0001, the T statistic tends to have inflated Type I error due to the less accurate characterization of the tail probability. In contrast, T^{e} can control the Type I error well at 0.0001 when the sample size and dimension are sufficiently large.

Table 2. *Power* (%) *of the tests under mixed signals at level* 0.05

n	p	Corr.	Bonf.	Burden	PCA	SKAT	T_H	T
500	300	CS	36.0	5.7	15.6	30.8	35.1	39.8
		AR	35.7	5.4	12.0	24.0	28.6	34.1
	500	CS	34.8	5.3	15.7	33.6	43.5	46.7
		AR	35.2	5.5	12.6	25.0	35.6	39.4
	1000	CS	32.4	5.2	18.5	39.2	61.8	63.6
		AR	31.1	5.0	14.8	26.8	51.0	53.5
1000	800	CS	38.4	5.2	16.0	54.0	60.1	62.4
		AR	36.1	5.0	12.3	42.4	50.1	53.3
	1000	CS	36.7	5.0	17.0	59.2	67.2	69.0
		AR	36.0	4.8	12.7	45.8	56.2	59.2
	1500	CS	35.4	5.1	18.1	67.3	79.4	80.9
		AR	34.5	4.6	13.1	52.6	69.1	71.5

Corr., correlation structure; CS, compound symmetric; AR, autoregressive; Bonf., Bonferroni test; PCA, principal component analysis using the five leading components; SKAT, the sequence kernel association test.

Table 3. *Type I error* ($\times 10^4$) *of the tests at level* 0.0001

n	p	Corr.	Bonf.	Burden	PCA	SKAT	T	T^e
500	500	CS	1.01	1.08	1.31	0.01	5.06	1.04
		AR	0.96	1.16	1.16	0.00	4.08	0.97
1000	1000	CS	1.34	1.01	1.20	0.00	3.19	1.08
		AR	1.30	1.08	1.10	0.00	2.86	1.05
1500	1500	CS	1.20	1.11	1.00	0.00	2.29	0.95
		AR	1.11	1.04	1.34	0.00	2.23	0.95

Corr., correlation structure; CS, compound symmetric; AR, autoregressive; Bonf., Bonferroni test; PCA, principal component analysis using the five leading components; SKAT, the sequence kernel association test.

4. REAL-DATA ANALYSIS

We analysed the high-density lipoprotein cholesterol data from the Genomics and Randomized Trials Network in the Women's Health Initiative (Coviello et al., 2012). The overall goal of the study is to identify novel genetic factors that contribute to the incidence of myocardial infarction, stroke and diabetes. DNA samples were genotyped on the HumanOmni-Quad platform, and genotypes were imputed with reference panels. Genetic variants that have imputations $R^2 > 0.99$ and minor allele frequency greater than 5% were included. We focused on the 3990 samples of Caucasian ancestry.

We first tested whether our approach can capture existing genetic pathways that are known to be involved in high-density lipoprotein metabolism. Assmann & Gotto (2004) listed a pathway involved in the generation and conversion of high-density lipoprotein. The pathway includes 11 genes: APOA1, APOE, LCAT, LIPC, CETP, PLTP, SCARB, LRP1, LDLR, ABCA1 and ABCF1. We mapped the genetic variants to these genes and obtained 629 variants for this pathway. We adjusted for the following covariates: age, hormone replacement therapy arm, smoking status, body mass index, and the first two principal components for ancestry (Asselbergs et al., 2012). The p-values for the pathway analysis are displayed in Table 4. Several methods yielded low p-values, including the Bonferroni test, the sequence kernel association test and the proposed

Table 4. *Real-data analysis: p-values of the tests for the known lipid pathway*

	Bonf.	Burden	PCA	SKAT	T^{e}_{H}	T^{e}
p-value	3.35×10^{-14}	0.252	0.034	4.09×10^{-6}	3.23×10^{-11}	$< 1 \times 10^{-300}$

Bonf., Bonferroni test; PCA, principal component analysis using the five leading components; SKAT, the sequence kernel association test.

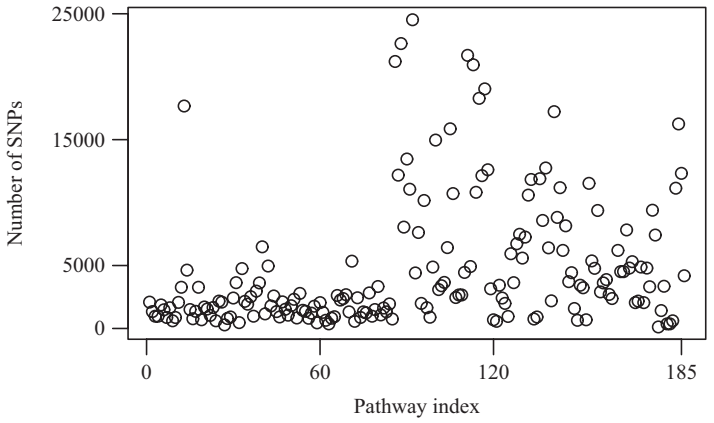

Fig. 2. The number of single-nucleotide polymorphisms, SNPs, in each of the 185 KEGG pathways.

tests T^{e}_{H} and T^{e}. The test T^{e} yielded the lowest *p*-value. The *p*-value of the test T^{e} is lower than that of T^{e}_{H} because a number of variants in the CETP and LIPC genes were observed to carry strong association signals that exceed the power-enhancement threshold.

Next, we investigated the associations between the KEGG pathways and high-density lipoprotein. The KEGG database contains 186 pathways, which represent a wide variety of cellular processes and molecular functions; for more details see `http://www.genome.jp/kegg/pathway.html`. We excluded one pathway from our analysis due to overlapping, so our real-data analysis includes 185 pathways. Figure 2 provides an overview of the number of variants in each of the 185 pathways. The median number of variants in these pathways is around 3000. A number of pathways have more than 10 000 variants, with some containing nearly 25 000.

To control for the familywise Type I error, the threshold of significance was set to $0.05/185 \approx 0.00027$, i.e., a Bonferroni correction. Table 5 shows the pathways that pass the significance threshold in any of the tests. The T^{e} approach identified three pathways: arachidonic acid metabolism, metabolism of xenobiotics by cytochrome P450, and drug metabolism by cytochrome P450. The T^{e}_{H} statistic yielded the same values as T^{e}, indicating that no signal exceeds the power-enhancement threshold in the studied pathways. The sequence kernel association test detected only the arachidonic acid metabolism pathway, while the other methods identified no significant pathway.

The arachidonic acid metabolism pathway contains 2590 variants in 55 genes. A recent biological study suggested that this pathway is an important regulator of cholesterol metabolism (Demetz et al., 2014). The linkage disequilibrium plot of the genetic variants of this pathway in Fig. 3(a) shows that variants in proximity to each other tend to have strong correlations, while those far apart have barely detectable correlations. To gain more insight, we plot the marginal *p*-values for all 2590 variants in Fig. 3(b). There are a number of variants with *p*-values between 10^{-2} and 10^{-4}, but none of them reaches genome-wide significance. Instead, the proposed T^{e}

Table 5. *The p-values* (%) *of the tests for the three significant KEGG pathways; p-values lower than* 0.05/185 *are indicated by* *

		p-value					
	#SNPs	Bonf.	Burden	PCA	SKAT	$T_{\mathrm{H}}^{\mathrm{e}}$	T^{e}
Arach. acid metab.	2590	5.11	0.57	0.07	0.02*	1.85×10^{-3}*	1.85×10^{-3}*
Metab. xenobio.	2254	6.30	0.46	0.18	0.07	2.29×10^{-2}*	2.29×10^{-2}*
Drug metab.	2385	7.86	0.39	0.16	0.04	6.00×10^{-3}*	6.00×10^{-3}*

#SNPs, number of single-nucleotide polymorphisms; Bonf., Bonferroni test; PCA, principal component analysis using the five leading components; SKAT, the sequence kernel association test; Arach. acid metab., arachidonic acid metabolism pathway; Metab. xenobio., metabolism of xenobiotics by cytochrome P450; Drug metab., drug metabolism by cytochrome P450.

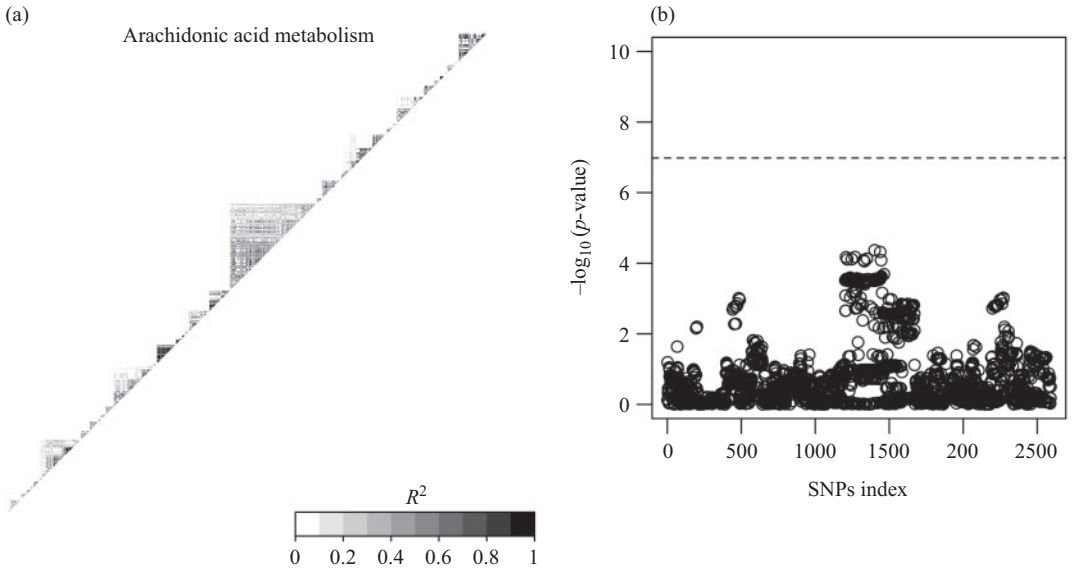

Fig. 3. Analysis of the arachidonic acid metabolism pathway: (a) linkage disequilibrium plot; (b) marginal *p*-values of the 2590 single-nucleotide polymorphisms, SNPs, in the pathway, where the dashed line represents the Bonferroni threshold.

statistic was able to aggregate these relatively mild signals into a stronger one, which leads to the detection of the arachidonic acid metabolism pathway. The variants that contribute to the significance of this pathway, the linkage disequilibrium plots and marginal *p*-values of variants in the other two pathways, metabolism of xenobiotics by cytochrome P450 and drug metabolism by cytochrome P450, are given in the Supplementary Material.

5. DISCUSSION

Our approach can be extended to deal with non-Gaussian errors as long as the errors satisfy the moment condition $\{E(|\varepsilon|^k)\}^{1/k} \leqslant Ck^{1/2}$ for some constant $C > 0$ and $k \geqslant 1$. In such a situation, we can adjust the denominator of T_{H} in (3) from $\sqrt{2\hat{\sigma}^2}\|A_1\|_{\mathrm{F}}$ to $\{2\hat{\sigma}^4\|A_1\|_{\mathrm{F}}^2 + (\kappa - 3)\hat{\sigma}^4 \sum_{i=1}^n A_{1ii}^2\}^{1/2}$, where κ is the kurtosis of the errors and A_{1ii} is the ith diagonal entry of A_1. Then, using the results in Bhansali et al. (2007), we can show the asymptotic normality of the

adjusted test statistic accordingly. Our approach can be also extended to accommodate genetic interactions.

Screening techniques have been used in genetic association studies to filter out irrelevant variants; see, for example, Li et al. (2014) and Cui et al. (2015). However, these screening procedures are typically used as a variable-selection step to reduce dimensions, not for statistical testing. In contrast, our screening statistic is directly integrated into the test statistic and is designed for statistical testing. Our approach has focused on the fixed design, which is commonly considered in genetic studies. It will be interesting to develop similar methods under the random design, although it remains challenging to establish the asymptotic properties of the proposed statistics in high dimensions.

Acknowledgement

This research was supported by the U.S. National Institutes of Health. We thank the Women's Health Initiative investigators for sharing the data. The Women's Health Initiative programme is funded by the National Heart, Lung and Blood Institute. Correspondence should be addressed to QH. We thank the editor, associate editor and reviewers for helpful comments.

Supplementary material

Supplementary material available at *Biometrika* online includes technical proofs, together with additional simulation results and real-data analysis.

References

Asselbergs, F. W., Guo, Y., Van Iperen, E. P., Sivapalaratnam, S., Tragante, V., Lanktree, M. B., Lange, L. A., Almoguera, B., Appelman, Y. E., Barnard, J. et al. (2012). Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am. J. Hum. Genet.* **91**, 823–38.

Assmann, G. & Gotto, A. M. (2004). HDL cholesterol and protective factors in atherosclerosis. *Circulation* **109**, III8–14.

Avery, C. L., He, Q., North, K. E., Ambite, J. L., Boerwinkle, E., Fornage, M., Hindorff, L. A., Kooperberg, C., Meigs, J. B., Pankow, J. S. et al. (2011). A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. *PLoS Genet.* **7**, e1002322.

Bhansali, R., Giraitis, L. & Kokoszka, P. (2007). Convergence of quadratic forms with nonvanishing diagonal. *Statist. Prob. Lett.* **77**, 726–34.

Buas, M. F., He, Q., Johnson, L. G., Onstad, L., Levine, D. M., Thrift, A. P., Gharahkhani, P., Palles, C., Lagergren, J., Fitzgerald, R. C. et al. (2017). Germline variation in inflammation-related pathways and risk of Barrett's oesophagus and oesophageal adenocarcinoma. *Gut* **66**, 1739–47.

Chen, L. S., Paul, D., Prentice, R. L. & Wang, P. (2011a). A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *J. Am. Statist. Assoc.* **106**, 1345–60.

Chen, M., Cho, J. & Zhao, H. (2011b). Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genet.* **7**, e1001353.

Chen, S. X. & Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38**, 808–35.

Conneely, K. N. & Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. *Am. J. Hum. Genet.* **81**, 1158–68.

Coviello, A.D., Haring, R., Wellons, M., Vaidya, D., Lehtimaki, T., Keildson, S., Lunetta, K.L., He, C., Fornage, M. & Lagou, V. et al. (2012). A genome-wide association meta-analysis of circulating sex hormone–binding globulin reveals multiple Loci implicated in sex steroid hormone regulation. *PLoS Genet.* **8**, e1002805.

Cui, H., Li, R. & Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *J. Am. Statist. Assoc.* **110**, 630–41.

Dai, J. Y., Kooperberg, C., Leblanc, M. & Prentice, R. L. (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* **99**, 929–44.

DEMETZ, E., SCHROLL, A., AUER, K., HEIM, C., PATSCH, J. R., ELLER, P., THEURL, M., THEURL, I., THEURL, M., SEIFERT, M. et al. (2014). The arachidonic acid metabolome serves as a conserved regulator of cholesterol metabolism. *Cell Metab.* **20**, 787–98.

DICKER, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101**, 269–84.

FAN, J., GUO, S. & HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Statist. Soc.* B **74**, 37–65.

FAN, J., LIAO, Y. & YAO, J. (2015). Power enhancement in high dimensional cross-sectional tests. *Econometrica* **83**, 1497–541.

FELLER, W. (1971). Expansions in the case of varying components. In *An Introduction to Probability Theory and Its Applications*, vol. 2. New York: Wiley, pp. 546–8.

GREGORY, K. B., CARROLL, R. J., BALADANDAYUTHAPANI, V. & LAHIRI, S. N. (2015). A two-sample test for equality of means in high dimension. *J. Am. Statist. Assoc.* **110**, 837–49.

HE, Q., ZHANG, H. H., AVERY, C. L. & LIN, D. (2016). Sparse meta-analysis with high-dimensional data. *Biostatistics* **17**, 205–20.

INTERNATIONAL HAPMAP CONSORTIUM (2005). A haplotype map of the human genome. *Nature* **437**, 1299–320.

LI, J., ZHONG, W., LI, R. & WU, R. (2014). A fast algorithm for detecting gene–gene interactions in genome-wide association studies. *Ann. Appl. Statist.* **8**, 2292–318.

MCKEAGUE, I. W. & QIAN, M. (2015). An adaptive resampling test for detecting the presence of significant predictors. *J. Am. Statist. Assoc.* **110**, 1422–33.

MORGENTHALER, S. & THILLY, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test. *Mutat. Res.* **615**, 28–56.

SHEN, D., SHEN, H. & MARRON, J. S. (2016). A general framework for consistency of principal component analysis. *J. Mach. Learn. Res.* **17**, 1–34.

WU, M. C., 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.

ZHANG, G. (2015). Genetic architecture of complex human traits: What have we learned from genome-wide association studies? *Curr. Genet. Med.* **3**, 143–50.

ZHONG, H., YANG, X., KAPLAN, L. M., MOLONY, C. & SCHADT, E. E. (2010). Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet.* **86**, 581–91.

[*Received on* 3 *July* 2017. *Editorial decision on* 4 *January* 2019]