Received: 26 March 2019 | Accepted: 25 April 2019
DOI: 10.1002/sta4.235

WILEY

ORIGINAL ARTICLE

Set-based differential covariance testing for genomics

Yi-Hui Zhou

Department of Biological Sciences and Bioinformatics Research Center, North Carolina State University, Raleigh, 27695, North Carolina, USA

Correspondence
Yi-Hui Zhou, Department of Biological Sciences and Bioinformatics Research Center, North Carolina State University, Raleigh, 27695, NC, USA.
Email: yihui_zhou@ncsu.edu

Funding information
National Human Genome Research Institute, Grant/Award Number: R01HG009125, R21HG007840 and CFF KNOWLE18XX0

The problem of detecting the changes in covariance for a single pair of genomic features has been studied in some detail but may be limited in importance or general applicability. For testing equality of covariance matrices of a set of features, many methods have been limited to the two-sample problem and involve varying assumptions on the number of features p versus the sample size n. More general covariance regression approaches are appealing but have been insufficiently structured to provide interpretable testing. To address these deficiencies, we propose a simple uniform framework to test association of covariance matrices with an experimental variable, whether discrete or continuous. We describe four different summary statistics, to ensure power and flexibility under various alternatives, including a new "connectivity" statistic that is sensitive to the changes in overall covariance magnitude. For continuous experimental variables, a natural individual "risk score" is associated with several of the statistics. We establish asymptotic results applicable to both continuous and discrete responses, with relatively mild conditions and allowing for situations where $p > n$. We also show that the proposed statistics are permutationally equivalent to some existing methods in the two-sample special case. We demonstrate the power and utility of our approaches via simulation and analysis of real data. The R package *CorDiff* is published on R CRAN.

KEYWORDS
asymptotics, covariance testing, permutation

1 | INTRODUCTION

Methods to detect association of an experimental variable with the changes in mean expression of sets of genes are well established (Barry, Nobel, & Wright, 2005; Goeman, Van De Geer, De Kort, & Van Houwelingen, 2004). In contrast, although tests of the changes in correlations or covariances have received considerable attention in areas such as genomics (McKenzie, Katsyv, Song, Wang, & Zhang, 2016) and finance (Isogai, 2016), set-based methods are less established. Existing methods for genomics have been confined largely to the two-sample problem (Choi & Kendziorski, 2009; Hu, Qiu, & Glazko, 2009; Hu, Qiu, Glazko, Klebanov, & Yakovlev, 2009). Exceptions include the liquid association method of Li (2002), developed to describe ternary relationships among genes, but the underlying motivation is based on the concept of differential covariance. For tests of biological systems, ensemble tests of covariance matrices for sets of genes in a proposed network or pathway can provide insight that might not be apparent in tests of individual gene pairs (Yuan, Deng, Tang, & Li, 2016).

Several existing differential covariance methods for genomics have primarily used permutation resampling for testing (e.g., Hu et al., 2009), partly because classical likelihood approaches require the sample size n to be large compared with the number of features p (Anderson, 1962; John, 1971). The two-sample problem tests equality of $p \times p$ covariance matrices $H_0 : \Sigma_1 = \Sigma_2$ based on samples of sizes n_1 and n_2, where $n_1 + n_2 = n$. In settings where $p > min\{n_1, n_2\}$, likelihood ratio testing may perform poorly or be undefined.

Recently, a number of statistical investigators have reconsidered the two-sample set-based differential covariance problem with an emphasis on behaviour when $p > n$, with little reference to the existing genomics literature. Li and Chen (2012) derived an approximately standard normal statistic for the Frobenius norm of differences in the two $p \times p$ sample covariance matrices, with considerable attention to sources of bias when p is large. A maximum standardized difference statistic was proposed (Cai, Liu, & Xia, 2013) between two sample covariances, with testing based on an extreme value approximation. There has been comparatively little discussion of the fact that these methods are designed for very different

alternatives, ranging from modest but widespread differences in the two sample covariance matrices (Li & Chen, 2012) to large differences in a very few covariance elements (Cai et al., 2013). An approach was reported (Zhu, Lei, Devlin, & Roeder, 2017) for sparse effects using eigenvalues but is also limited to the two-sample problem. Cai and Sun (2017) give a review of the previously cited papers, from the perspective of one-sample and two-sample testing. Additional aspects, such as local testing to discover groups of differentially correlated features, are beyond our scope. Importantly, there are few approaches to test association of covariances with a continuous outcome.

Covariance regression methods (Hoff & Niu, 2012; Zou, Lan, Wang, & Tsai, 2017) offer a potentially more flexible approach, with the ability to handle continuous predictor variables. Methods such as Zou et al. (2017) are ultimately sensitive to the property that similarity in predictors corresponds to similarity in responses and thus are similar to kernel methods such as SKAT (Ionita-Laza, Lee, Makarov, Buxbaum, & Lin, 2013). The special case of a linear kernel and a restricted set of predictors in Zou et al. (2017) has some similarities to our testing approach, but the authors focus on efficiency and comparison of maximum likelihood and other fitting approaches. The regression method of Hoff and Niu (2012) for a predictor denoted y, which may be continuous, models the covariance as $\Sigma_y = \Psi + Byy^TB^T$ and for a positive definite $p \times p$ matrix Ψ maintains a positive definite covariance function. However, the p-vector B provides in some ways less flexibility compared with the methods we propose, because we wish to be sensitive to any changes in the covariance matrix, but at the same time, the vector B may be difficult to interpret. In addition, the approach does not match with standard statistics used in the two-sample problem and does not lend itself to a simple test of association with y.

The differential covariance literature continues to grow, with little cross-talk among the areas described above (classical two-sample testing, two-sample testing with $p > n$, and covariance regression). Key advances for genomics would be (a) to recognize situations in which analytic p-values are available and accurate even if $p > n$, (b) to move beyond the two-sample scenario without the complications of interpreting the full results of covariance regression, (c) to have a framework for understanding different test statistics, for a variety of alternatives.

1.1 | A new set of statistics

To address these issues, we first note that the two-sample problem can be viewed as an "association" of the covariance matrix with a binary group indicator. More generally, the investigator may be interested in trend association of covariance with an experimental variable y that might be multilevel or on a continuous scale. To our knowledge, no general method is available with the requisite flexibility, without restrictive parametric requirements or assumptions of the feature size p relative to n.

In this manuscript, we use a simple framework to propose four different statistics to test the changes in a covariance matrix of p features. Several of these statistics have been proposed for the two-sample problem but not extended to a continuous predictor. Moreover, existing methods have been published in isolation, providing little opportunity to consider power characteristics for various types of alternatives. Thus, a major contribution of this paper is to provide a common framework and to point out the similarities and differences among covariance-change statistics.

Here, we propose four statistics to test the changes in a covariance matrix of p features, when it is anticipated that a change in y will result in (a) a directional change in many elements of the covariance matrix, (b) a nondirectional change in covariance, (c) a change in the overall magnitude of covariances, or (d) a large change in one or a few elements of the covariance matrix. In contrast to almost all of the comparable methods, the statistics apply naturally whether y is continuous or discrete. Asymptotic results and p-values for most of the statistics are derived under relatively mild conditions, to provide computationally efficient p-values for the two-sample comparisons. Permutation can be used for the remaining statistic and if the researcher does wish to rely on asymptotic theory to ensure control of type I error. Our approach is not limited by the data dimensions and is applicable to situations where $p > n$.

This paper is organized as follows. In Section 2, we introduce the method and test statistics. Section 3 provides asymptotic results for the general methods, and for the special case of the two-sample problem shows that several of our statistics match up with those previously proposed. In Section 4, we compare the proposed statistics with existing methods, in terms of type I error and power. Several different simulation settings are presented for the two-sample problem, comparing our statistics to existing methods. In addition, we compare our proposed methods in the setting with continuous y. Section 5 illustrates with real examples.

2 | METHODS

2.1 | Notation

Let X be the $p \times n$ data matrix consisting of elements x_{ik} and y the n-vector of clinical/experimental data. The sample mean and variance of a vector follow standard notation, for example, $\bar{y}$ and s_y^2. The ith row and kth column of X are denoted $x_{i.}$ and $x_{.k}$, and each column is assumed to have a population p-mean of zero. Random variables are capitalized (e.g., random Y_k vs. observed y_k). We denote the $p \times p$ covariance of X, which may depend on y_k, as Σ_{y_k}. The zero-mean assumption is implicit in most covariance tests, following an intent that the test statistics be sensitive only to the changes in covariance. For a subset of samples ω with at least two samples, the sample covariance is $\hat{\Sigma}_\omega = X_\omega X_\omega^T / n_\omega$. A single i, j element is $\hat{\sigma}_{ij,\omega} = \left(\sum_{k \in \omega} x_{ik} x_{jk}\right) / n_\omega$. We use ξ to denote the operator that sums all elements of a matrix and the superscript "$\circ k$" to denote the element-wise exponent of a matrix to power k.

2.2 | A conceptual trend model

The existing methods are limited to two-sample comparison with binary y. To motivate our statistics for binary/continuous y, we adopt a conceptual trend model for the covariance dependence of X on y: $\Sigma_y = \beta_0 + \beta_1 y$ for $p \times p$ matrices β_0, β_1. Thus, for sample k, according to our assumptions, $E(X_{ik}|y_k) = 0$ for each i and $\text{cov}(X_{ik}, X_{jk}|y_k) = E(X_{ik}X_{jk}|y_k) = \beta_{0,ij} + \beta_{1,ij}y_k$ for the ith and jth features. Letting $z_{ijk} = x_{ik}x_{jk}$, the model immediately suggests linear regression of z on y, for which the least squares slope solution is $\hat{\beta}_{1,ij} = (\Sigma_k z_{ijk}y_k/n - \bar{z}\bar{y})/\left(s_y^2 \frac{n-1}{n}\right)$, where $\bar{z} = \sum_k z_{ijk}$. Although the trend assumption is simple, $\hat{\beta}_1$ can be viewed as an approximate score statistic for additive models of the form $E(Z_{ijk}|y_k) = \eta_{0,ij} + \eta_{1,ij}f(y_k)$ for a strictly monotone smooth f, where $\eta_{0,ij}$ and $\eta_{1,ij}$ are the unknown coefficients in this "true" model, and our $\beta_{1,ij}$ is thus locally powerful for detecting departures from the null $\eta_{1,ij} = 0$. We make two further observations: (a) We do not consider $\beta_{0,ij}$ to be of interest for detecting the covariance changes, and (b) linear rescaling of y will not meaningfully change our results, because it results in constant changes in the proposed statistics. Thus, without loss of generality, we assume $\bar{y} = \sum_k y_k/n = 0$, so $\hat{\beta}_{1,ij} = \frac{1}{ns_y^2}\sum_k x_{ik}x_{jk}y_k = \frac{1}{ns_y^2}\sum_k z_{ijk}y_k$. These least squares solutions are not intended to be used directly but serve to motivate global test statistics described below. For the two-sample special case, the trend model results in $\hat{\Sigma}_y$ matrices that are sample covariance matrices for each of the two samples and thus non-negative definite. For continuous y, the trend model does not guarantee non-negative definiteness throughout the range of y. However, the clear interpretation remains for $\hat{\beta}_1$ and its use for each of the summary covariance-change statistics.

2.3 | Four statistics

In this subsection, we propose four statistics as a unified framework for covariance testing for sets of genes, which are equally applicable to a binary (the two-sample problem) or continuous y . In different scenarios, each of them has a role to promote biological/medical discoveries. A summation statistic S is sensitive to the covariance changes in the same direction, whereas a quadratic form statistic Q is sensitive to the changes in either direction. Several existing methods are essentially two-sample special cases of S and Q, as we later show. The connectivity statistic C is completely novel and has a relationship to overall covariance magnitude, which may be useful in network analysis. Statistics similar to the maximum statistic M have been well studied (Cai et al., 2013) but had been previously limited to the two-sample problem.

2.3.1 | A summation statistic

To effectively measure the covariance changes, we propose $S = \sum_i \sum_j \hat{\beta}_{1,ij}$ as a summation statistic to detect the global changes in covariances that are concordantly associated with the experimental variable y (i.e., in the same direction). A simplification for S is

$$S = \sum_i \sum_j \sum_k x_{ik}x_{jk}y_k = \sum_k y_k \sum_i x_{ik} \sum_j x_{jk} = \sum_k y_k \left(\sum_i x_{ik}\right)^2 = \sum_k w_k y_k = y^T w,$$

for $w_k = (\sum_i x_{ik})^2$. In datasets where the null for S can be rejected, the value w_k represents a natural "risk score" for sample k, with extreme w values corresponding to extreme y.

Although the initial motivation for S was based on $p \times p$ covariance terms, the restated statistic is ultimately based on an inner product of n-vectors, and thus, we may use a large sample normal approximation as $n \to \infty$ for rescaled S to obtain p-values.

2.3.2 | A quadratic form statistic

In contrast, $Q = \sum_i \sum_j \hat{\beta}_{1,ij}^2$ is sensitive to the changes that are not directionally concordant. Similar to S, Q can also be represented by n-vectors and $n \times n$ matrices.

$$\begin{aligned} Q &= \sum_i \sum_j \left(\sum_k x_{ik}x_{jk}y_k\right)^2 = \sum_i \sum_j \sum_k \sum_l x_{ik}x_{jk}x_{il}x_{jl}y_k y_l \\ &= \sum_k \sum_l y_k y_l \sum_i x_{ik}x_{il} \sum_l x_{jk}x_{jl} = \sum_k \sum_l y_k y_l a_{kl}, \end{aligned}$$

where $a_{kl} = \left(\sum_i x_{ik}x_{il}\right)^2$. The matrix with elements a_{kl} can be simplified to $A = (X^T X)^{\circ 2}$. Finally, we have the quadratic form $Q = y^T A y$.

The nature of Q makes it difficult derive a risk score analogue and also difficult to justify closed-form limiting approximations to its null distribution. An exception is in extreme cases, such as dominance of a single eigenvalue in A (approximately chi-square), or with a large number of eigenvalues of similar magnitude (approximately normal). As explored in Zhou et al. (2013), for small to moderate sample sizes, a weighted beta approximation can be more accurate than standard approximations for sums of independent chi-square distributions. However, the procedure can be somewhat computationally intensive, and here, we opt for direct permutation of y to obtain p-values, for a general A. For certain special cases, solutions for the first four permutation moments (Zhou et al., 2013) may be used to obtain p-values but require restrictive assumptions on the form of A.

2.3.3 | A connectivity statistic

Each element a_{kl} of A has the form of a squared inner product between samples k and l, and so $b_k = \sum_l a_{k,l}$ reflects broad-scale association (a "connectivity index") of sample k with remaining samples. Accordingly, we propose the connectivity statistic $C = y^T b$ for risk scores $\{b_k\}$ to reflect correlation between y and the connectivity index. Correlations between samples are ultimately driven by correlation between features, and C reflects the tendency for the aggregate magnitude of feature–feature correlations to be associated with y, which is quite different from the type of alternative envisioned for S and Q.

2.3.4 | A maximum statistic

Our fourth statistic is similar to that of Cai et al. (2013) for the two-sample special case. We use a test for the maximum element difference, scaled by an appropriate standard error. For our generalization of the statistic, we define $M_{ij} = (n-1)r_{ij}^2$, where r_{ij} is the Pearson correlation between y and z_{ij}. Defining $\{i', j'\} = argmax_{i,j} M_{ij}$, we propose the maximum statistic $M = M_{i',j'}$, with a risk score $z_{i',j'}$.

Approximate p-values for M use an extreme value approximation for χ_1^2 variates as $n \to \infty$. Beyond standard assumptions that elements of X and Y have appropriate tail behaviour (e.g., sub-Gaussian), there are modest restrictions on Σ and that $\log(p)$ grows more slowly than $n^{1/5}$. With these assumptions, approximate p-values are obtained using $P(M - 4\log p + \log\log p \le t) \approx \exp\left(-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{t}{2}\right)\right)$.

2.4 | Multiple testing

For practical data analysis, we use the standard Benjamini–Hochberg false discovery rate (FDR) control for each of the covariance-change statistics across all the gene sets examined, in order to examine all findings comprehensively. This approach may be viewed as a form of stratified FDR control (Sun, Craiu, Paterson, & Bull, 2006), with the aggregated FDR (for all gene sets and all four statistics) expressible as a weighted average of the stratum-specific FDRs. In this manner, we are able to focus on top findings for each of the statistics in turn, highlighting the most significant findings. Note that, according to this perspective, we can choose the most significant findings across the statistics while maintaining the stratified FDR, and whichever statistics appear in the FDR-significant set are driven by those that are most powerful for the problem at hand.

3 | THEORETICAL RESULTS

3.1 | Asymptotic theorems to obtain *p* values

We have established asymptotic results for statistics S, C, and M for general y, which is non-trivial because p may be large relative to n. Suitable rescaling provides p-values using normal (S and C) or extreme value (M) distributions.

3.1.1 | Theorem 1 (*S* and *C*)

This theorem implies that S and C are asymptotically normal, as will be explained following the proof. The proof is a triangular array version of the central limit theorem tailored for our setting.

Theorem 1. *The Lindeberg condition for the sequence of random variables $\{U_{k,n}\}$ states that for any $\varepsilon > 0$,*

$$\frac{1}{\sum_k \mathrm{var}(U_{k,n})} \sum_k E\left((U_{k,n} - E(U_{k,n}))^2 I(|U_{k,n} - E(U_{k,n})| > \varepsilon \sqrt{\sum_k \mathrm{var}(U_{k,n})} \right) \to 0,$$

as $n \to \infty$. Assume Y and $W_{k,n}$ are random variables such that $E(Y^4) < \infty$ and $E(W_{k,n}^4) < \infty$ for all k, n, and $k = 1, \ldots, n$. Define $V_{k,n} = W_{k,n} Y_k$ and $S_n = \sum_k V_{k,n}$ and assume the Lindeberg condition holds for $\{W_{k,n}^2\}$ and $\{V_{k,n}\}$. Further, define

$$R = \frac{S_n/n - \bar{W}_n \bar{Y}}{(n-1) s_W s_Y},$$

where $\bar{W}_n$, $\bar{Y}$, s_W, and s_Y are the sample means and standard deviations of $W_{\cdot,n}$ and Y, and $Z = \sqrt{n}R$. Then, under H_0 that $W_{\cdot,n}$ and Y are independent, for any t, $P(Z \le t) \to \Phi(t)$ as $n \to \infty$.

The proof is in the Supporting Information.

Both S and C can be placed within this framework. To apply the theorem to S, we define $A_{k,n} = \sum_i X_{ik}$ and $W_{k,n} = A_{k,n}^2$. Under the null hypothesis that X is independent of Y, clearly, W is also independent of Y. It is clear from the definition that it allows for dependence of p on n. For C, we define $W_{k,n} = B_{k,n}$ where B_k is the random variable risk score given in Section 2.3.3.

Note that the theorem is stated generally enough that the columns of X, which correspond to the k subscript in W, be nonidentically distributed. Verifying the conditions requires a model for X with increasing n but will be satisfied for a wide variety of "typical" assumptions. For example, previous work has often assumed multivariate normality of X and either a binary or normal Y, for which the conditions can be shown to be satisfied (see below for example) under very general conditions.

3.1.2 | An example of establishing conditions for Theorem 1 for *S* and *C* for multivariate normal *X*

Suppose each of the columns of X is p-variate multivariate normal $MVN(0,\Sigma)$, where both p and Σ may depend on n. We have from the above definitions $A_{k,n}=\sum_i X_{ik}$, $W^2_{k,n}=A^4_{k,n}$, and $\xi^2_n=\mathrm{var}(A_{k,n})=E(A^2_{k,n})$. For MVN X, $A^4_{k,n}=Z^4_{k,n}\xi^4_n$, where $Z_{k,n}\sim N(0,1)$ (and we drop the n subscript from Z). As k is arbitrary, Theorem 1 requires establishing Lindeberg conditions for $W^2_{1,n}$, for which the stronger Lyapunov condition is sufficient, and we compute

$$L_n=\frac{1}{(n\mathrm{var}(W_{1,n}))^{1+\delta/2}}nE\left(\left|W^2_{1,n}-E\left(W^2_{1,n}\right)\right|^{2+\delta}\right)=\frac{1}{n^{\delta/2}\xi_n^{8+4\delta}\mathrm{var}(Z^4)}E\left(\left(\xi^4_n\left|Z^4_1-E\left(Z^4_{1,n}\right)\right|\right)^{2+\delta}\right),$$

for $\delta>0$, and the terms involving ξ_n cancel whereas other terms do not depend on n. Also, p does not appear because the sum across rows of X is normal and already considered in ξ^2_n. It is thus clear that $L_n\to 0$ as $n\to\infty$, satisfying the condition. Applying the same approach to $\sum_k W_{k,n}Y_k$ (i.e., establishing the condition for the $V_{k,n}$ terms) similarly yields cancellation of ξ_n and $L_n\to 0$. Specifically, p does not appear, and the denominator of the right-hand side of the above equation will include $\xi^{4+2\delta}E(Y^2)\mathrm{var}(Z^2)$ and $E\left(\left(\xi^2_nE(Y^2)\left|Z^2_1-E\left(Z^2_{1,n}\right)\right|\right)^{2+\delta}\right)$, because under the null hypothesis, $A^2_{k,n}$ and Y_k are independent.

For C, we apply the theorem conditions to $W_{k,n}=B_k=\sum_k\left(\sum_i X_{ik}X_{il}\right)^2$, and the Lyapunov condition is easy to establish from independence of columns of X.

3.1.3 | Theorem 2 (*M*)

Theorem 2. *Suppose that the correlation condition C1, tail condition C2, and the moment condition C3 shown below hold. Then, under H_0, for any t,*

$$P(M-4\log(p)+\log(\log(p))\le t)\to\exp\left(-\frac{1}{\sqrt{8\pi}}\exp(-t/2)\right),$$

as $n\to\infty$.

The proof is in the Supporting Information.

In Cai et al. (2013), several conditions (denoted C1–C3) were assumed to hold, and our analogues are required. In our notation, consistent with Cai et al. (2013), X_{ik} is the data value for feature i and sample k, with mean μ_i. Key terms in the proof and conditions include the pair $\{U_{i[j]},V_{j[i]}\}$, where $U_{i[j]}=(X_i-\mu_i),V_{j[i]}=W_j$ if $i\le j$, and $U_{i[j]}=W_i,V_{j[i]}=(X_j-\mu_j)$ if $i>j$. Terms ρ_{ij} are feature–feature correlations, and $s_j(\alpha_0)=\mathrm{card}\{i:|\rho_{ij}|\ge(\log p)^{-1-\alpha_0}\}$, $\sigma_{ij}=E\left(U_{i[j]}V_{j[i]}\right)$, $\theta_{ij}=\mathrm{var}\left(U_{i[j]}V_{j[i]}\right)$, and under the null hypothesis $\sigma_{ij}=0$. Roughly, condition C1 controls the proportion of large feature–feature correlations, C2 limits the tail behaviour for X, and C3 provides bounds on four-way cross moments among features.

Condition C1. Cai et al. (2013). For $r\in(0,1)$, define $\Lambda(r)=\{1\le i\le p:|\rho_{ij}|>r\}$ for some $j\ne i$. Suppose there exists a subset $\Upsilon\subset\{1,2,..,p\}$ with cardinality $o(p)$ and a constant $\alpha_0>0$ such that for all $\gamma>0$, $\max_{1\le j\le p,j\notin\Upsilon}s_j(\alpha_0)=o(p^\gamma)$ and there exists $r>1$ and a sequence $\Lambda_{p,r}$ such that the cardinality of $\mathbf{\Lambda}(r)\le\Lambda_{p,r}=o(p)$.

Condition C2. Sub-Gaussian tail and polynomial tail conditions, analogue of C2 and C2* in Cai et al. (2013). Suppose that $\log(p)=o(n^{1/5})$. There exist constants $\eta>0$ and $K>0$ such that $E(e^{\eta(X_i-\mu_i)^2/\mathrm{var}(X_i)})\le K, E(e^{\eta(W_i)^2/\mathrm{var}(W_i)})\le K$ for all i. Alternatively, we assume that for some $\gamma_0,c_1>0$, $p\le c_1n^{\gamma_0}$, and $\epsilon>0$, $E(|(X_i-\mu_i|/\mathrm{var}(X_i)^{1/2})^{4\gamma_0+4+\epsilon}\le K$ and $E(|(W_i-E(W_i))|/\mathrm{var}(W_i)^{1/2})^{4\gamma_0+4+\epsilon}\le K$ for all i.

Also, in either case, we assume $\min_{1\le i\le j\le p}\theta_{ij}/(\mathrm{var}(U_{i[j]})\mathrm{var}(V_{j[i]}))\ge\tau$ for some $\tau>0$.

Condition C3. Analogue of C3 in Cai et al. (2013). For any collection i,j,k, and $l\in\{1,2,\ldots,p\}$, we assume without loss of generality that $i\le j$ and $k\le l$, and we suppose there exists $\kappa\ge\frac{1}{3}$ such that

$$E\left(U_{i[j]}V_{j[i]}U_{k[l]}V_{l[k]}\right)=\kappa\left(\sigma_{ij}\sigma_{kl}+E\left(U_{i[j]}V_{j[i]}\right)E\left(V_{j[i]}V_{l[k]}\right)+E\left(U_{i[j]}V_{l[k]}\right)E\left(V_{j[i]}U_{k[l]}\right)\right).$$

We use Lemmas 1 and 2 from Cai et al. (2013) and the following lemma.

Analogue of Lemma 3 of Cai et al. (2013). Under the conditions of C2, there exists some constant $C>0$ such that

$$P\left(\max_{i,j}|\hat\theta_{ij}-\theta_{ij}|/\left(\mathrm{var}\left(U_{i[j]}\right)\mathrm{var}\left(V_{j[i]}\right)\right)\right)\ge C\frac{\epsilon_n}{\log(p)}=O\left(p^{-1}+n^{-\epsilon/8}\right),$$

where $\epsilon_n=\max\left((\log(p)^{1/6}/n^{1/2},\log(p)^{-1}\right)\to 0$ as $n,p\to\infty$.

We have $\hat\theta_{ij}=s^2_{Z_{ij}}s^2_Y$. We first note that Lemma 3 from Cai et al. (2013) applies directly to $s^2_{Z_{ij}}$ as an estimator of $\mathrm{var}(Z_{ij})$ or alternatively to $s^2_{Z_{ij}Y}$ as an estimator of the same quantity if Z is centred. Furthermore, $s^2_Y/\sigma^2_Y=O_P(n^{-1})$ and so

$$P\left(\max_{i,j}\left|s_{Z^2_{ij}}s^2_Y-\sigma^2_{Z_{ij}}\sigma^2_Y\right|\right)\ge C\frac{\epsilon_n}{\log(p)}=O\left(p^{-1}+n^{-\epsilon/8}\right),$$

using variance scaling wlog as in the Supplemental Appendix of Cai et al. (2013)

With the above assumptions in hand, the proof follows from the proof of Theorem 1 in Cai et al. (2013), where in each instance, our $\hat\sigma_{ij}$ is substituted for (say) $\hat\sigma_{ij1}$ from Cai et al. (2013), and zero substituted for $\hat\sigma_{ij2}$, and a single denominator $\hat\theta_{ij}/n$ in place of the denominator in Cai et al.

(2013), which gives the correlation coefficient. Condition C1 prevents excessive correlation of features, whereas C2 ensures that large deviations in the data do not prevent limiting convergence of extreme $\hat{\sigma}_{ij}$.

3.2 | Permutation testing

Although for computational speed we typically rely on asymptotic *p*-values, permutation testing can be useful, as a means of both performing small sample analysis and informing interpretation of our statistics, as we show in the next subsection. Letting Π denote a random permutation of n elements from among the $n!$ possibilities (realized value π), the statistics for permutation π are $S_\pi = y_\pi^T w$, $Q_\pi = y_\pi^T A y_\pi$, $C_\pi = y_\pi^T b$, and M_π (which require computation of the $\hat{\beta}_1$ values and standard errors for each permutation). S and C are subjected to two-sided testing, with *p*-values based on both right and left tails, whereas Q and M are one-tailed, rejecting for large values. For example, with H random permutations and $\pi[h]$ denoting the hth permutation, the empirical *p* value for S is $p_S = \sum_{h=1}^{H} I[|S_{\pi[h]}| \geq |S_{observed}|]/H$, whereas the *p* value for Q is $p_Q = \sum_{h=1}^{H} I[Q_{\pi[h]} \geq Q_{observed}]/H$.

The null hypothesis is that the relationships of columns of X to the elements of y are exchangeable (Good, 2002), which holds if X and y are drawn from independent distributions. A primary advantage of permutation testing is that, aside from slight issues due to discreteness or tied outcomes, type I error rates are controlled without requiring parametric assumptions (Zhou & Wright, 2015). Note that, because permutation testing is conditional on the observed data, the dimension p is immaterial in terms of the ability to maintain appropriate false positive control. However, the sample size n should be sufficient that *p*-values $< \alpha$ can be achieved. For example, if y is continuous, then the minimum *p* value achievable for two-sided statistics is $2/n!$.

3.3 | Special case: Two group comparisons and permutation equivalence

As stated earlier, all of our statistics apply for a general y. Here, we show that in the two-sample special case, our proposed statistics S and Q are equivalent (in a permutation sense) to natural summaries of sample covariance matrices. In addition, M closely matches the Cai et al. (2013) statistic. These concepts are illustrated in Figure 1 (left panel), which summarizes the properties that each statistic has in terms of alternatives to which it is sensitive.

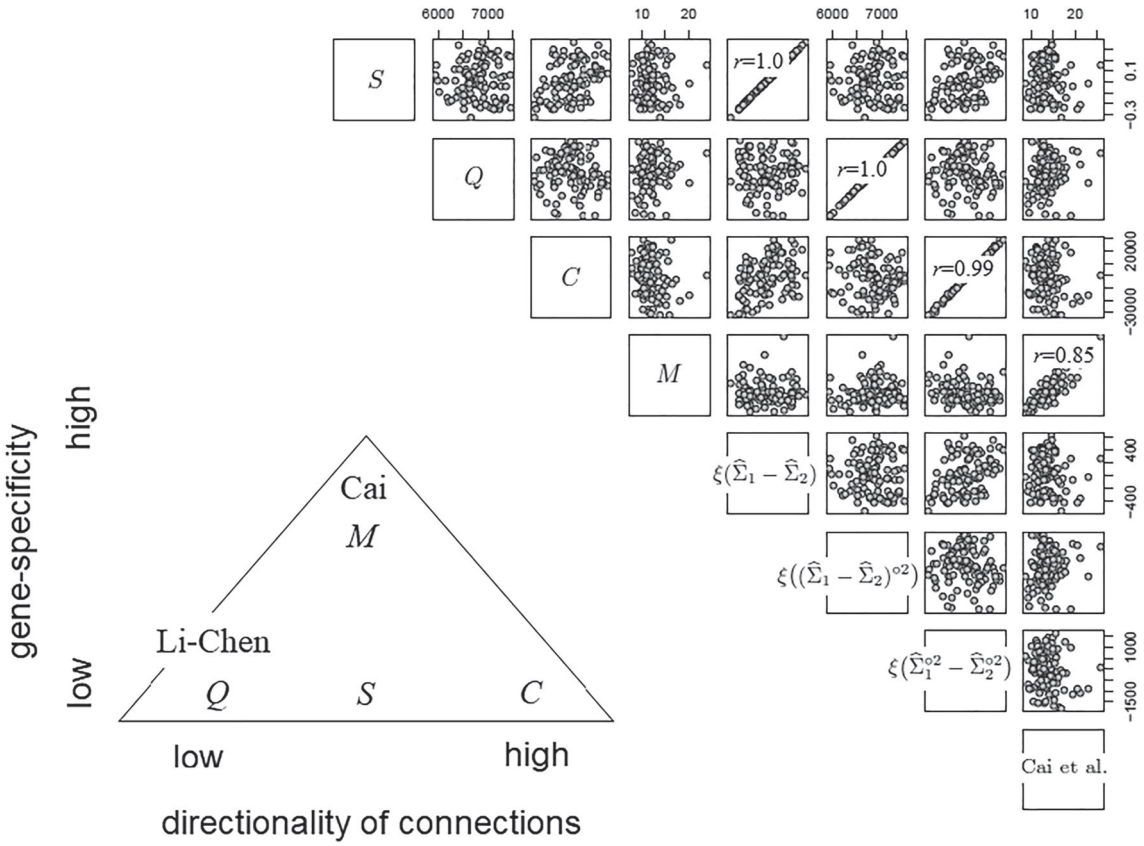

FIGURE 1 Left panel: A visual summary of the proposed covariance-change statistics and are grouped with previously proposed statistics for the two-sample problem special case. Right panel: Comparison of the four proposed statistics to various existing statistics for the two-sample problem special case, for a single-simulated dataset, and for 100 permutations. Pearson correlations illustrate the exact and approximate correspondence of some pairs of statistics

For the two-sample problem, an obvious summation-like statistic would be $\xi\left(\hat{\Sigma}_1 - \hat{\Sigma}_2\right)$. A statistic $\xi\left(\left(\hat{\Sigma}_1 - \hat{\Sigma}_2\right)^{\circ 2}\right)$ is sensitive to the covariance changes in either direction that might otherwise cancel in the previous statistic. A statistic $\xi\left(\hat{\Sigma}_1^{\circ 2} - \hat{\Sigma}_2^{\circ 2}\right)$ is directional, but for the magnitude of covariances, and by squaring the covariance elements provides additional weight to gene–gene pairs with high covariance magnitude. Finally, a maximum statistic would identify the maximum covariances differences $(\hat{\sigma}_{ij,1} - \hat{\sigma}_{ij,2})^2$ but appropriately scaled by a standard error for each gene pair.

Result 1. *Let ω_1 and ω_2 be the indexes for samples in groups 1 and 2, respectively, and the subscripts 1 and 2 will be used for simplicity. We assign the experimental variable $y_k = \frac{1}{n_1}$ if $k \in \omega_1$ and $y_k = \frac{-1}{n_2}$ if $k \in \omega_2$. Then*

(i) the directional statistic S is equivalent to $\xi\left(\hat{\Sigma}_1 - \hat{\Sigma}_2\right)$;
(ii) the nondirectional statistic Q is equivalent to $\xi\left(\left(\hat{\Sigma}_1 - \hat{\Sigma}_2\right)^{\circ 2}\right)$.

The proof is in the Supporting Information.

Figure 1 (right panel) shows the results from 100 random permutations of y for the two-sample problem with $n_1 = n_2 = 20$, $p = 50$. A single X was generated using the null version of Model 2 described in the next section, but the qualitative results hold regardless of the choice of X. As we showed above, S and Q are equivalent to $\xi(\hat{\Sigma}_1 - \hat{\Sigma}_2)$ and $\xi\left((\hat{\Sigma}_1 - \hat{\Sigma}_2)^{\circ 2}\right)$, respectively. Under the permutations, C has a high Pearson correlation over permutations with $\xi\left(\hat{\Sigma}_1^{\circ 2} - \hat{\Sigma}_2^{\circ 2}\right) = \xi(\hat{\Sigma}_1^{\circ 2}) - \xi(\hat{\Sigma}_2^{\circ 2})$, supporting the perspective that C reflects a contrast in the overall magnitude of covariances. Finally, our M is correlated under permutation with the statistic from Cai et al. (2013), although they differ modestly due to the differences in the standard errors used.

This permutation example underscores the correspondence between our statistics and those that seem "natural" for the two-sample problem, but we emphasize that our statistics apply for either discrete or continuous y.

4 | SIMULATION MODELS FOR TYPE I ERROR AND POWER

Initial comparisons follow the simulation settings from Li and Chen (2012), for which feature covariances were described using auto-regressive notation. More compactly than the original articles, we describe their simulation settings in terms of the covariance matrices.

4.1 | Simulation Model 1 with a continuous y

For this simulation model, values in y are drawn iid $N(0, 1)$ in each simulation and converted to the rescaled experimental variable $y^* = \frac{y - min(y)}{max(y) - min(y)} \in [0, 1]$. X is drawn as multivariate $N(0, \Sigma_{y^*})$, with $\Sigma_{y^*} = (1 - y^*)\gamma_1 + y^*\gamma_2$. We assume γ_1 is the identity matrix and γ_2 is the compound symmetric matrix,

$$\gamma_{2ij} = \begin{cases} 1, & \text{if} \quad i = j \\ \rho, & \text{if} \quad i = j + 1, \quad j - 1 \\ 0, & \text{if} \quad i \neq j - 1, j, j + 1 \end{cases},$$

in which we call ρ the "effect size." Under the null, there is no change in the covariance structure, that is, γ_2 is the identity matrix, as is Σ_{y^*} for all y^*. As ρ increases, the relationship between the covariance and y^* becomes stronger. Figure 2 shows that the power for the proposed statistics is near the intended $\alpha = .05$ when $\rho = 0$. Figure 2 also shows that the directional statistic S is the most powerful approach overall.

4.2 | Simulation Model 2 with a continuous y

This simulation model is a bit more complex, following a similar approach used in Cai et al. (2013). The approach generates covariance matrices that are nondirectional in relationship to y and with no overall variation in magnitude, while respecting the need for positive definiteness. To an initial $p \times p$ identity matrix I, $\Sigma^{*(1)}$ was formed by drawing the first $p/2 \times p/2$ off-diagonal elements from a uniform density $U[-\rho, \rho]$, followed by $\Sigma^{*(2)} = \Sigma^{*(1)} + \Sigma^{*(1)T}$ and $\Sigma_1 = \Sigma^{*(2)} + (\lambda_{min}(\Sigma^{*(2)}) + 0.05)I$. Σ_2 is formed by reversing the rows and columns of Σ_1, and finally, $\Sigma_{y^*} = \Sigma_1(1 - y^*) + \Sigma_2 y^*$, where y^* is the result of linear rescaling of y to the $[0, 1]$ interval as in the previous subsection. Here, $\rho \in [0, 1)$ serves as an effect size, and Σ_1 and Σ_2 differ in the groups of genes that show correlation structure but otherwise are the same in the average magnitude of elements and show no directionality. Figure 3 provides the power comparison among the four proposed methods. As expected, S and C have little or no power, whereas M has low power. The statistic Q benefits from aggregation of covariance-squared differences and thus has much more power than the other methods. All methods control type I error properly (dashed line at 0.05 in Figure 3).

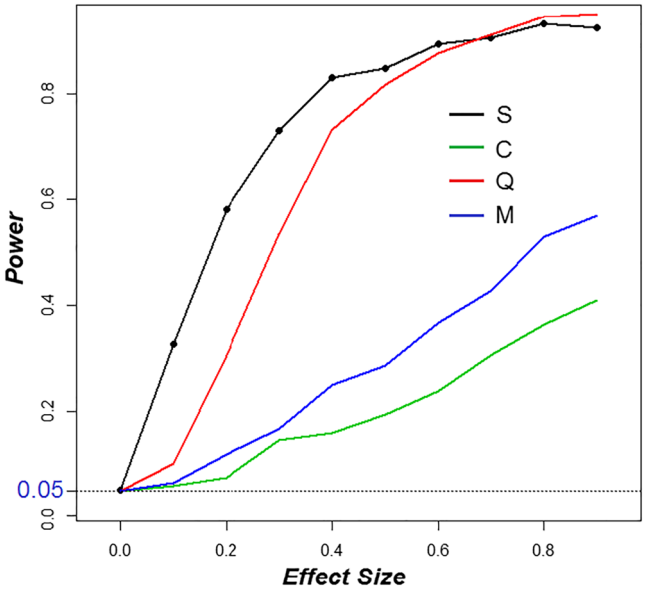

FIGURE 2 Power comparison among *S*, *Q*, *C*, and *M* for Simulation Model 4. The dashed line at $\alpha = .05$ indicates that all the proposed methods control type I error well under the null ($\rho = 0$). The effect size ρ ranges from 0 to .8

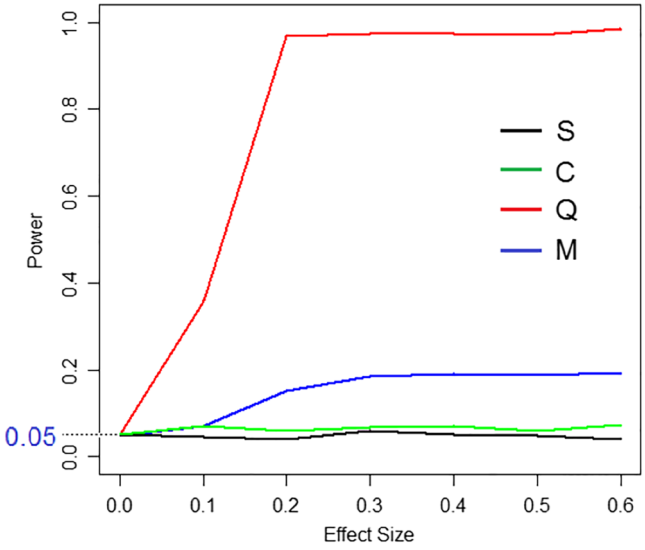

FIGURE 3 Power comparison among *S*, *Q*, *C*, and *M* for Simulation Model 5. The dashed line at $\alpha = .05$ indicates that all the proposed methods control type I error well under the null ($\rho = 0$). The effect size ρ ranges from 0 to .6. *Q* is the most powerful method among these statistics for this simulation model

4.3 | Simulation Model 3 with discrete *y* (to assess type I error)

This simulation procedure was originally from Li and Chen (2012). We assume the first population $X_1 \sim N(0, \Sigma_1)$; whereas the second population $X_2 \sim N(0, \Sigma_2)$, where

$$\Sigma_{1ij} = \begin{cases} 1+\theta_1^2, & \text{if} \quad i=j \\ \theta_1, & \text{if} \quad i=j+1, \quad j-1 \\ 0, & \text{if} \quad i \neq j-1, j, j+1 \end{cases}, \Sigma_{2ij} = \begin{cases} 1+\theta_1^2+\theta_2^2, & \text{if} \quad i=j \\ \theta_1(1+\theta_2), & \text{if} \quad i=j+1, \quad j-1 \\ 0, & \text{if} \quad i \neq j-1, j, j+1 \end{cases}.$$

The difference between the two covariance matrices is

$$\Sigma_{2ij} - \Sigma_{1ij} = \begin{cases} \theta_2^2, & \text{if} \quad i=j \\ \theta_1\theta_2, & \text{if} \quad i=j+1, \quad j-1 \\ 0, & \text{if} \quad i \neq j-1, j, j+1 \end{cases}.$$

To assess type I error, we set $\theta_2 = 0$, which implies the null $\Sigma_1 - \Sigma_2 = 0$. We show results for $n_1 = n_2 = \{20, 50, 80, 100\}$ and feature dimension $p = \{32, 64, 128, 256, 512, 700\}$. The number of simulations was 1,000 for each setting. The asymptotic results were used to obtain *p*-values for

TABLE 1 Type I error comparison, Simulation Model 1, $X_k \sim N(0, \Sigma_k)$, $\Sigma_1 = \Sigma_2$

$n_1 = n_2$	Method	$p = 32$	$p = 64$	$p = 128$	$p = 256$	$p = 512$	$p = 700$
20	*S*	.042	.043	.051	.035	.033	.044
	Q	.055	.058	.046	.043	.043	.052
	C	.041	.049	.029	.053	.053	.061
	Li–Chen	.044	.054	.051	.048	.051	.038
	Cai	.092	.14	.139	.204	.211	.263
	M	.074	.085	.085	.132	.177	.209
	M_p	.053	.054	.050	.052	.051	.050
50	*S*	.047	.042	.053	.035	.049	.035
	Q	.052	.041	.045	.049	.042	.046
	C	.055	.049	.051	.046	.057	.042
	Li–Chen	.052	.060	.033	.043	.054	.049
	Cai	.042	.068	.058	.065	.055	.059
	M	.049	.042	.058	.034	.043	.033
	M_p	.059	.054	.051	.048	.051	.050
80	*S*	.043	.057	.051	.043	.047	.04
	Q	.065	.051	.040	.046	.044	.048
	C	.056	.045	.047	.051	.063	.05
	Li–Chen	.054	.060	.047	.048	.052	.053
	Cai	.052	.056	.043	.052	.058	.041
	M	.074	.043	.034	.036	.042	.027
	M_p	.046	.046	.051	.047	.050	.049
100	*S*	.056	.05	.042	.051	.047	.05
	Q	.039	.051	.050	.040	.060	.053
	C	.057	.042	.028	.043	.057	.055
	Li–Chen	.056	.049	.052	.046	.049	.048
	Cai	.050	.052	.043	.039	.036	.047
	M	.072	.052	.04	.045	.032	.027
	M_p	.047	.054	.048	.050	.048	.046

S and *C* and 1,000 permutations for *Q*. Although the parametric method works well for *M* for a moderate sample size, to ensure robustness for this statistic for the entire range of simulations, we also report permutation-based *p*-values, labelled M_p.

Table 1 shows that for this multivariate normal model, most methods perform well and control type I error. The Cai et al. (2013) method and our similar *M* statistics are noticeably anticonservative for $\alpha = .05$ for the smaller sample size ($n_1 = n_2 = 20$) and more so as *p* increases. As stated above, for comparison, we have also included in the table a statistic M_p, which is the permutation-based version of our *M*. For larger sample sizes, the asymptotic *p*-values for *M* are close to nominal, even for large *p*. For the setting with $n_1 = n_2 = 20$, $p = 50$, 100,000 simulations were performed to provide greater insight into tail behaviour (Figure 4); *p*-values for *Q* perform well, which is sensible, as the permutation null holds. In addition, we also show good results for a ''residualized'' *Q* (lower right panel), in which each row of *X* is residualized using simple linear regression for the effect of *y*. The rationale for such an approach might be to ensure that the test statistic is sensitive to the changes in covariance only, not to any linear association with *y*. Here, the residualization is also performed inside the permutation loop.

4.4 | Simulation Model 4 with discrete *y* (to assess type I error)

Here, we follow the previous Simulation Model but with skewed data elements. Specifically, let $G(w; 4, 0.5)$ denote the Gamma distribution function with shape parameter 4 and scale 0.5 evaluated at *w*. Then, if $W \sim G$, $X = W - 2$ has mean zero and variance 1, that is, follows a centred Gamma. The elements of X_1 and X_2 are drawn as shown above, following the same null covariance structure that was used in Simulation Model 3.

Here, the Cai approach in Cai et al. (2013) becomes conservative, with increasing both the sample size and feature size (Table 2). The Li–Chen method is anticonservative, but the type I error becomes closer to nominal as the sample size and feature size increase.

4.5 | Simulation Model 5 with discrete *y* (to assess power)

For power comparisons, we return to the multivariate normal data elements. We use Simulation Model 3 but with covariance matrices determined by $\theta_1 = 2, \theta_2 = 1$ (one of the simulation models also used by Li and Chen, 2012, and summarized in their Table 3). Although this simulation model

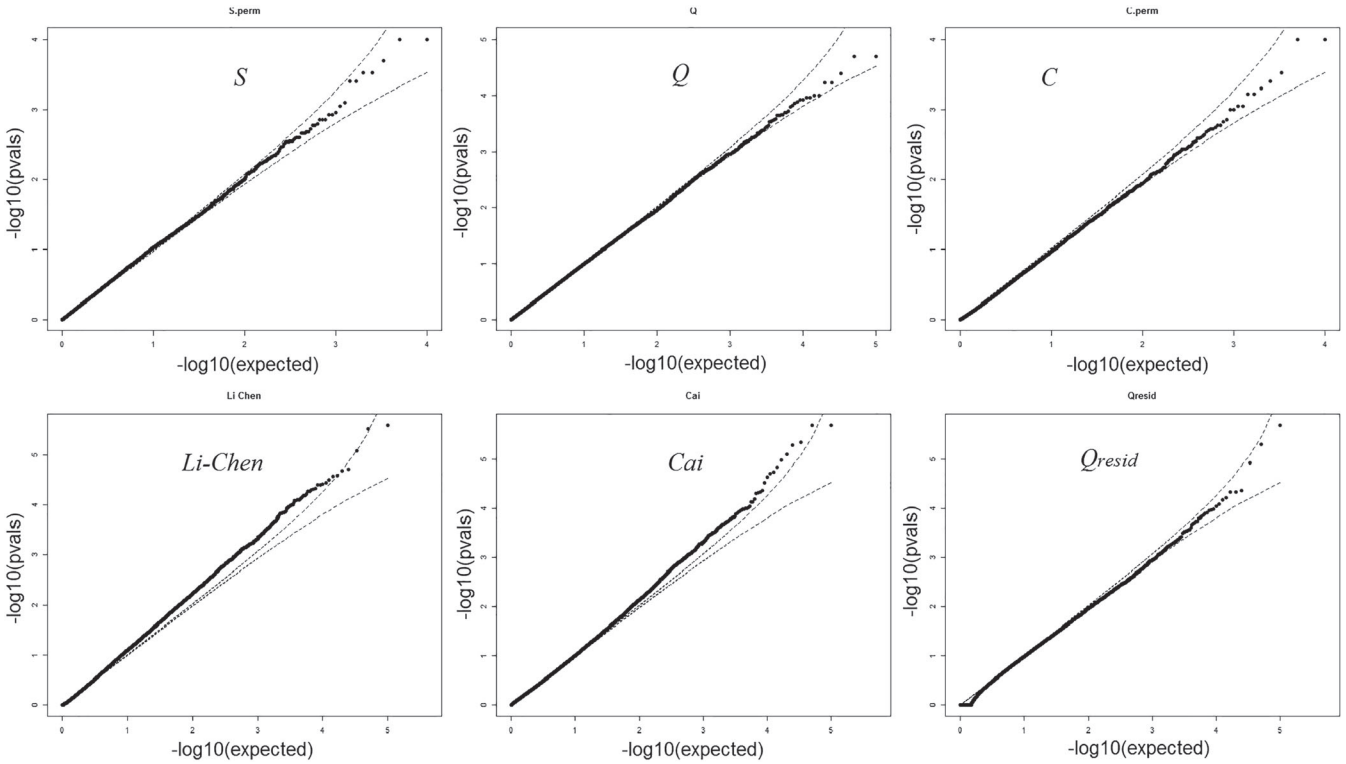

FIGURE 4 QQplots for several of the proposed methods and existing methods for the null two-sample problem of Simulation Model 1, $p = 50$, $n_1 = n_2 = 20$

TABLE 2 Type I error comparison, Simulation Model 2, $\Sigma_1 = \Sigma_2$, elements following centred Gamma

$n_1 = n_2$	Method	$p = 32$	$p = 64$	$p = 128$	$p = 256$	$p = 512$	$p = 700$
20	*S*	.034	.042	.039	.039	.039	.04
	Q	.057	.046	.066	.050	.047	.042
	C	.035	.05	.047	.05	.037	.05
	Li–Chen	.158	.112	.083	.071	.053	.063
	Cai	.048	.048	.058	.055	.083	.085
	M	.083	.102	.108	.145	.152	.171
	M_p	.039	.039	.050	.037	.042	.046
50	*S*	.049	.061	.062	.048	.055	.044
	Q	.055	.049	.055	.043	.051	.042
	C	.045	.053	.051	.048	.044	.05
	Li–Chen	.048	.048	.058	.055	.083	.085
	Cai	.016	.013	.010	.007	.003	.004
	M	.039	.062	.053	.054	.047	.042
	M_p	.051	.05	.055	.042	.050	.048
80	*S*	.051	.046	.05	.049	.045	.056
	Q	.056	.048	.043	.042	.038	.057
	C	.062	.049	.042	.047	.047	.061
	Li–Chen	.165	.141	.090	.059	.051	.056
	Cai	.019	.010	.005	.006	.005	.002
	M	.046	.051	.036	.049	.048	.038
	M_p	.059	.039	.045	.045	.048	.051
100	*S*	.054	.037	.047	.047	.047	.048
	Q	.045	.042	.049	.049	.051	.039
	C	.041	.038	.045	.058	.049	.053
	Li–Chen	.176	.133	.088	.069	.050	.046
	Cai	.013	.009	.007	.003	.003	.003
	M	.048	.044	.039	.034	.037	.048
	M_p	.059	.041	.058	.052	.053	.050

TABLE 3 Power comparison, Simulation Model 3, $X_k \sim N(0, \Sigma_k)$, $\Sigma_1 \neq \Sigma_2$

$n_1 = n_2$	Method	$p = 32$	$p = 64$	$p = 128$	$p = 256$	$p = 512$	$p = 700$
20	*S*	.18	.156	.179	.163	.164	.158
	Q	.211	.231	.235	.234	.221	.213
	C	.63	.826	.972	.999	1	1
	Li–Chen	.273	.273	.252	.285	.269	.272
	Cai	.138	.140	.164	.204	.233	.282
	M	.125	.111	.115	.157	.182	.189
	M_p	.129	.072	.050	.061	.083	.054
50	*S*	.456	.474	.456	.452	.473	.437
	Q	.705	.751	.803	.809	.772	.789
	C	.989	1	1	1.000	1	1
	Li–Chen	.752	.800	.824	.861	.839	.857
	Cai	.234	.163	.146	.136	.104	.084
	M	.217	.14	.134	.126	.097	.077
	M_p	.270	.133	.092	.122	.034	.051
80	*S*	.644	.695	.686	.692	.677	.703
	Q	.955	.972	.991	.995	.992	.992
	C	1	1	1	1.000	1	1
	Li–Chen	.941	.980	.992	.994	.996	.998
	Cai	.496	.420	.377	.316	.246	.189
	M	.468	.417	.367	.305	.229	.175
	M_p	.574	.394	.333	.242	.253	.201
100	*S*	.761	.776	.788	.796	.813	.814
	Q	.991	.997	.999	1.000	1.000	1.000
	C	1	1	1	1.000	1	1
	Li–Chen	.997	1.000	.999	1.000	1.000	1.000
	Cai	.700	.652	.557	.508	.423	.406
	M	.689	.633	.56	.494	.403	.398
	M_p	.700	.649	.601	.487	.375	.374

was used by Li and Chen (2012) to support their proposed statistic, our proposed *C* has consistently the highest power for all the n, p settings. The Li–Chen statistic shows power slightly higher than that of *Q*, even though they both are based on the Frobenius norm. We speculate that the reason is related to the fact that permutation testing is conditional on the observed data, and the power difference nearly disappears at the larger sample sizes. It is perhaps a bit surprising that *S* is less powerful than *Q*, as the covariance differences are directional. However, the squared terms in *Q* also may effectively act to reduce noise, and we have observed situations in which *S* is more powerful. The Cai and *M* statistics show the lowest power, as they use only the most extreme covariance difference element and do not aggregate over the large number of covariance difference elements.

5 | ANALYSIS OF REAL DATASETS

5.1 | A continuous phenotype example

To illustrate the utility of covariance testing in association with a continuous phenotype, we reanalyzed the well-known data of van de Vijver et al. (2002), in which gene expression in breast tumours of 295 patients younger than 55 was examined for association with disease-free survival. To identify biological pathways (gene sets) of interest in the comparisons, groups of genes for 372 KEGG and 8,039 Gene Ontology pathways were identified, so that for each test, *p* represents the number of genes in the pathway. To identify pathways of greatest interest, we aggregate over all gene pairs in each pathway, using statistics as described below. We used martingale residuals, adjusted for age and sex, and 50 surrogate variables, obtained from the expression data, as a quantitative phenotype *y*; *p*-values for *Q* were determined by 100,000 permutations for high accuracy, and the remaining statistics used the asymptotic approximations. Of the four statistics proposed, three achieved false discovery $q < 0.15$ for the most significant gene set using Benjamini–Hochberg adjustment. These were *Q* (GO:0043254 regulation of protein complex assembly, 193 genes, $q = 0.0066$), *C* (GO:0022408 negative regulation of cell–cell adhesion, 75 genes, q=0.124), and *M* (GO:1900221 regulation of beta-amyloid clearance, five genes, q=0.055).

To further examine the finding for statistic *C*, we performed proportional hazards regression using the risk score vector $\{b_k\}$ as a predictor for disease-free survival while including the covariates described. The resulting Wald statistic for the risk scores was $p = 8.0 \times 10^{-15}$. This striking result for the risk scores would be highly significant by any conceivable multiple test correction, even though the *q*-value for *C* was of

borderline significance. To visualize, we divided the risk score vector for C into tertiles, and Figure 5 shows the corresponding Kaplan–Meier curves for disease-free survival. The result shows that high "connectivity" of the genes in the pathway is associated with reduced disease-free survival, consistent with observations that loss of cellular adhesion promotes metastasis (Martin, Ye, Sanders, Lane, & Jiang, 2013). Note that this observation would not be apparent from standard gene-set enrichment approaches using overall expression levels.

5.2 | Discrete phenotype example 1

The second real dataset is gene expression data on kidney transplant tissue (Modena et al., 2016), in which those with acute rejection ($y = 1, n_1 = 54$) were compared with normal outcomes ($y = 2, n_2 = 99$). Pathway analyses proceeded similarly as with the previous example, with 7,266 Gene Ontology BP and 402 KEGG pathways examined. The most significant pathways for each statistic are listed in Figure 6. Each heatmap depicts the matrix corresponding to the statistic (e.g., for S, it depicts $\hat{\Sigma}_1 - \hat{\Sigma}_2$). The most significant pathways for each statistic are as follows, with multiple comparison false discovery q-values: GO:0035754 B cell chemotaxis (five genes, for S, $q = 1.4 \times 10^{-6}$), GO:0070193 synaptonemal complex organization (11 genes, for Q, $q = 0.03$), and GO:0009394 2′-deoxyribonucleotide metabolic process (28 genes, for C, $q = 4.2 \times 10^{-18}$).

As another illustration for this dataset, we show the results for the most significant M statistic for GO:0021889 olfactory bulb interneuron differentiation (Figure 6d, 13 genes, p value = .0005, q n.s.). To best illustrate the changes in correlation rather than the changes in variance, for this statistic, we row-scaled the data to have variance 1 for each gene. The gene pair {*ATF5*, *ERRB4*} shows the most significant change, with a high negative correlation in the AR group and little correlation in the normal group. *ATF5* has been associated with transplant rejection in multiple organ systems (Morgun et al., 2006). There is little literature on *ERBB4* and transplant rejection, but the gene has been associated with kidney nephropathy (Sandholm et al., 2012) and thought to be protective of polycystic kidney disease in a mouse model (Zeng, Miyazawa, Kloepfer, & Harris, 2014).

5.3 | Discrete phenotype example 2

The third real dataset is a targeted reanalysis of a brain expression dataset compiled by Fulcher and Fornito (2016). The authors had used previous mouse brain connectome findings to classify each of 213 brain regions as "hub" and "nonhub" regions using imaging-based connectivity, with expression data obtained for each region. A primary biological finding in their paper was that hub regions involved coexpression of genes involved in energy metabolism. However, to test this hypothesis, the authors needed to use an indirect means, creating a threshold-based connectivity score for each gene and testing for enrichment using methods that do not acknowledge gene–gene correlations. We reasoned that a gene-set approach using S and C might be able to obtain a similar finding directly. Defining $y = 1$ for hub regions and $y = 0$ for nonhub regions, and considering each region to be a sample, we performed testing for each of 5,944 Gene Ontology pathways containing at least five genes. For S, the most significant pathways associated with an increase of covariance for hub regions were GO:0005746 ($q = 1.09 \times 10^{-4}$, five genes, "mitochondrial respiratory chain") and GO:0045039 ($q = 1.37 \times 10^{-4}$, five genes, "protein import into mitochondrial inner membrane"). Similarly, for C, for the same GO category, we obtained $q = 9.85 \times 10^{-4}$. These pathways clearly involve energy metabolism, whereas the top pathways showing increased covariance in nonhub regions were not involved in energy metabolism. For example, in the nonhub regions, for S, we obtained

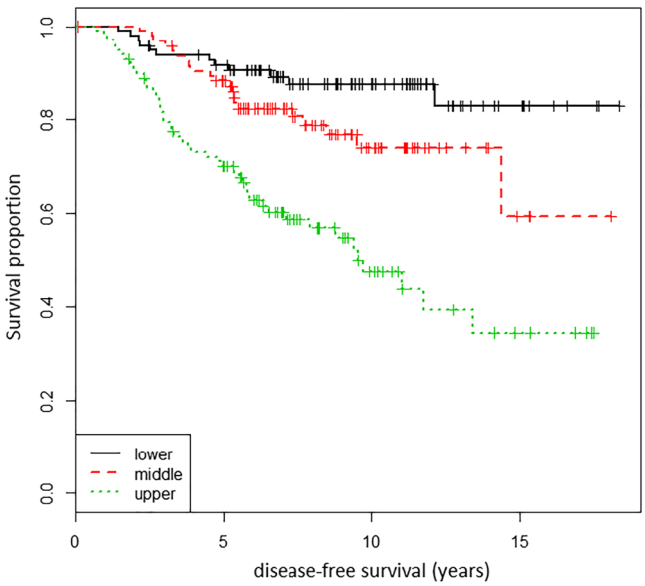

FIGURE 5 Kaplan–Meier curves for disease-free survival and breast cancer data of van de Vijver et al. (2002; $p = 8 \times 10^{-15}$). The curves correspond to tertiles of the risk scores for statistic C, for pathway GO:0022408 "negative regulation of cell–cell adhesion"

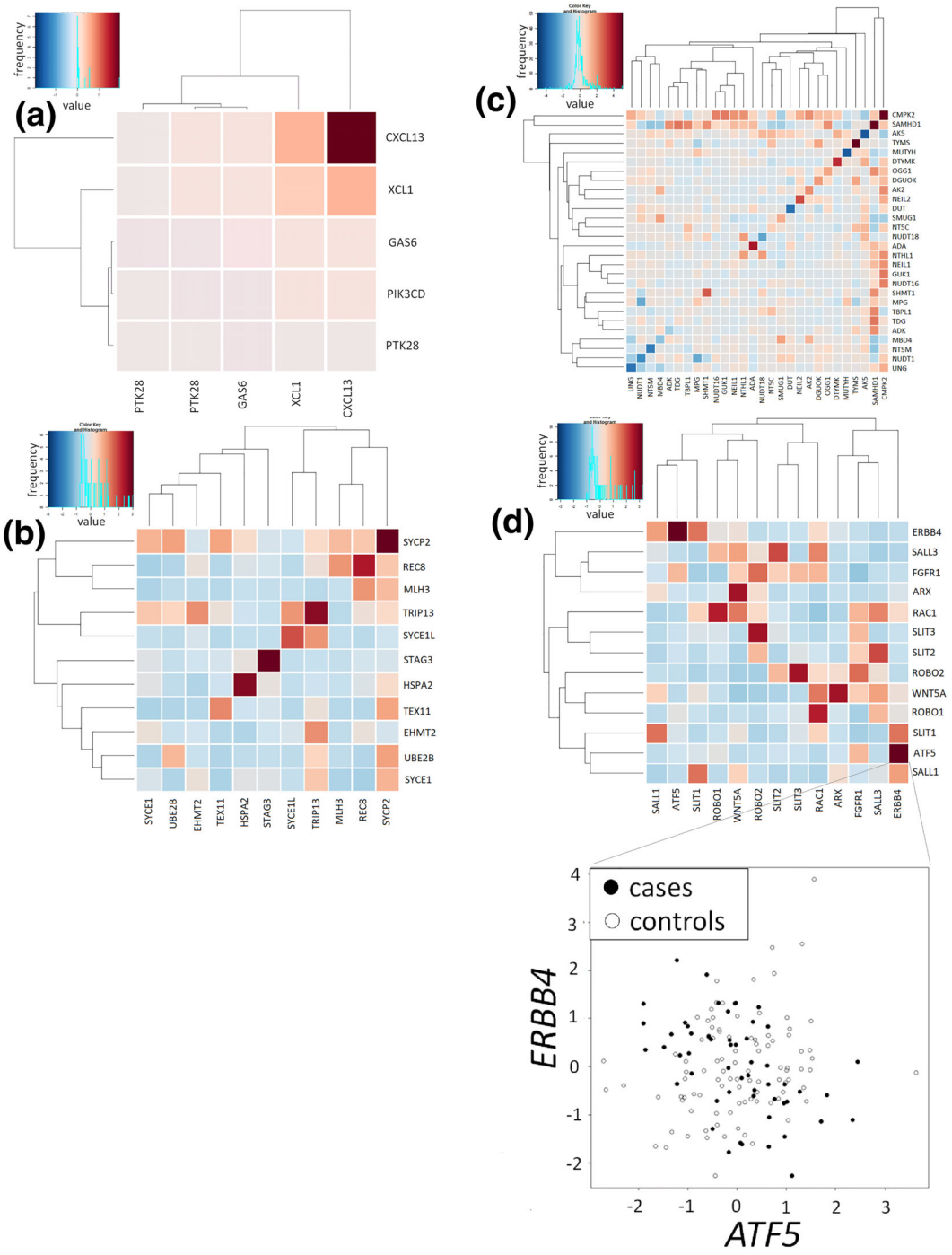

FIGURE 6 Illustration of the four statistics in the kidney transplant data, using gene set analysis of gene expression in those with acute rejection ($n_1 = 54$) versus normal ($n_2 = 99$). All panels except for Panel (a) were zero-centred to better illustrate the covariance changes. (a) Heatmap of $\hat{\Sigma}_1 - \hat{\Sigma}_2$ for GO:0035754, the most significant pathway for *S*. (b) Heatmap of $(\hat{\Sigma}_1 - \hat{\Sigma}_2)^{\circ 2}$ for GO:0070193, the most significant pathway for *Q*. (c) Heatmap of $(\hat{\Sigma}_1^{\circ 2} - \hat{\Sigma}_2^{\circ 2})$ for GO:0009394, the most significant pathway for *C*. (d) Heatmap of M_{ij} values for GO:0021889, the most significant pathway for *M*. The inset shows the covariance in acute rejection versus controls for *ATF5* and *ERBB4*

GO:0015677 ($q = 4.08 \times 10^{-6}$, seven genes, "copper ion import") and GO:0019218 ($q = 4.75 \times 10^{-4}$, five genes, "regulation of steroid metabolic process").

6 | DISCUSSION

We have proposed four covariance test statistics in a straightforward trend-testing framework that applies to general *y*. The approach is not limited by *p*, *n*, or whether *y* is discrete or continuous. For most of the statistics, a natural risk score is an output. The availability of testing for a

continuous y is a distinct advantage over previous methods, making covariance testing a simple approach that can be applied in a huge variety of settings. We propose that the approach can be part of a standard testing toolkit and used to evaluate, for example, pathway associations, in high-throughput data or the statistical significance of network discoveries.

7 | DATA AVAILABILITY STATEMENT

The three datasets described in Section 5 are in https://sites.google.com/ncsu.edu/zhouslab/home/software?authuser=0.

8 | SOFTWARE

The accompanying software *CorDiff* is available on R CRAN (https://cran.r-project.org/web/packages/CorDiff/index.html).

ORCID


Yi-Hui Zhou https://orcid.org/0000-0002-4092-7463

REFERENCES

Anderson, T. W. (1962). An introduction to multivariate statistical analysis. (*No. 519.9 A53*). New York: Wiley.

Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach. *Bioinformatics*, *21*(9), 1943–1949.

Cai, T., Liu, W., & Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association, 108*(501), 265–277.

Cai, T. T., & Sun, W. (2017). Large-scale global and simultaneous inference: Estimation and testing in very high dimensions. *Annual Review of Economics*, *9*, 411–439.

Choi, Y., & Kendziorski, C. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics*, *25*(21), 2780–2786.

Fulcher, B. D., & Fornito, A. (2016). A transcriptional signature of hub connectivity in the mouse connectome. *Proceedings of the National Academy of Sciences*, *113*(5), 1435–1440.

Goeman, J. J., Van De Geer, S. A., De Kort, F., & Van Houwelingen, H. C. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*, *20*(1), 93–99.

Good, P. I. (2002). Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods*, *1*(2), 34.

Hoff, P. D., & Niu, X. (2012). A covariance regression model. *Statistica Sinica*, *22*(2)729–753.

Hu, R., Qiu, X., & Glazko, G. (2009). A new gene selection procedure based on the covariance distance. *Bioinformatics*, *26*(3), 348–354.

Hu, R., Qiu, X., Glazko, G., Klebanov, L., & Yakovlev, A. (2009). Detecting intergene correlation changes in microarray analysis: A new approach to gene selection. *BMC Bioinformatics*, *10*(1), 20.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., & Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, *92*(6), 841–853.

Isogai, T. (2016). Building a dynamic correlation network for fat-tailed financial asset returns. *Applied Network Science*, *1*(1), 7.

John, S. (1971). Some optimal multivariate tests. *Biometrika*, *58*(1), 123–127.

Li, K.-C. (2002). Genome-wide coexpression dynamics: Theory and application. *Proceedings of the National Academy of Sciences*, *99*(26), 16875–16880.

Li, J., & Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, *40*(2), 908–940.

Martin, T. A., Ye, L., Sanders, A. J., Lane, J., & Jiang, W. G. (2013). Cancer invasion and metastasis: Molecular and cellular perspective. *Metastatic Cancer: Clinical and Biological Perspectives*.

McKenzie, A. T., Katsyv, I., Song, W.-M., Wang, M., & Zhang, B. (2016). DGCA: A comprehensive r package for differential gene correlation analysis. *BMC Systems Biology*, *10*(1), 106.

Modena, B. D., Kurian, S. M., Gaber, L. W., Waalen, J., Su, A. I., Gelbart, T., ..., & Friedewald, J. J. (2016). Gene expression in biopsies of acute rejection and interstitial fibrosis/tubular atrophy reveals highly shared mechanisms that correlate with worse long-term outcomes. *American Journal of Transplantation*, *16*(7), 1982–1998.

Morgun, A., Shulzhenko, N., Perez-Diez, A., Diniz, R. V., Sanson, G. F., Almeida, D. R., Matzinger, P., & Gerbase-DeLima, M. (2006). Molecular profiling improves diagnoses of rejection and infection in transplanted organs. *Circulation Research*, *98*(12), e74–e83.

Sandholm, N., Salem, R. M., McKnight, A. J., Brennan, E. P., Forsblom, C., Isakova, T., ..., & Swan, E. J. (2012). New susceptibility loci associated with kidney disease in type 1 diabetes. *PLoS Genetics*, *8*(9), e1002921.

Sun, L., Craiu, R. V., Paterson, A. D., & Bull, S. B. (2006). Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, *30*(6), 519–530.

Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., ..., & Parrish, M. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, *347*(25), 1999–2009.

Yuan, H., Deng, M., Tang, N. L., & Li, Z. (2016). A network based covariance test for detecting multivariate eqtl in Saccharomyces cerevisiae. *BMC Systems Biology*, *10*(1), S8.

Zeng, F., Miyazawa, T., Kloepfer, L. A., & Harris, R. C. (2014). Deletion of erbb4 accelerates polycystic kidney disease progression in cpk mice. *Kidney International*, *86*(3), 538.

Zhou, Y.-H., Mayhew, G., Sun, Z., Xu, X., Zou, F., & Wright, F. A. (2013). Space–time clustering and the permutation moments of quadratic forms. *Stat*, *2*(1), 292–302.

Zhou, Y.-H., & Wright, F. A. (2015). Hypothesis testing at the extremes: Fast and robust association for high-throughput data. *Biostatistics*, *16*, 611–625.

Zhu, L., Lei, J., Devlin, B., & Roeder, K. (2017). Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes. *The Annals of Applied Statistics*, *11*(3), 1810.

Zou, T., Lan, W., Wang, H., & Tsai, C.-L. (2017). Covariance regression analysis. *Journal of the American Statistical Association*, *112*(517), 266–281.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Zhou Y-H. Set-based differential covariance testing for genomics. *Stat*. 2019;8:e235. https://doi.org/10.1002/sta4.235