Submitted to the Annals of Applied Probability
arXiv: arXiv:0000.0000

DYNAMIC GENE COEXPRESSION ANALYSIS WITH CORRELATION MODELING

By Tae Kim and Dan Nicolae

University of Chicago

In many biological studies on the transcriptome, the correlation of genes might fluctuate with quantitative factors such as genetic ancestry. We propose a method that can model the covariance between two variables to vary against a continuous covariate. For the bivariate case, we propose a score test statistic that is computationally simple and robust to model misspecification of the covariance term ρ. Subsequently, we expand the method to test relationships between one highly connected gene, such as transcription factors, and several other genes to obtain a more global view of the dynamic of the coexpression network. Simulations show that the proposed method has higher statistical power than the alternatives, works under more diverse scenarios, and is computationally much cheaper. We apply this method to African American subjects from GTEx to analyze the dynamic behavior of their gene coexpression against genetic ancestry, and we identify transcription factors whose coexpressions with their target genes change with the genetic ancestry. We believe this method can be applied to a wide array of problems that require covariance modeling.

1. Introduction.

Gene coexpression, the covariance structure of gene expression data, shows how genes are functionally connected and provides insights into the design of the transcriptional regulatory system. Ideally, such a complicated biological system can be fully understood through longitudinal observations in multiple and diverse cell types that capture the dynamics of the system. In reality, however, such comprehensive measurements are often unavailable or too expensive, and the expression dynamics must be captured instead through cross-sectional or tissue-specific data sets. In such cases, investigating the dependence structure can be useful. The dependence structure can be especially valuable for characterizing how few key genes are connected to the rest of the transcriptome. For example, we can focus on one transcription factor — genes that help turn transcription of genes on and off — and study how it is connected to its target genes. To further investigate

Keywords and phrases: coexpression; network; heteroskedasticity; score test; GTEx; admixed population

this problem, we define "local connectivity" of a transcription factor as its overall connectivity to its target genes.

Consider this biological problem: how does local connectivity vary across various phenotypic conditions? Past studies have investigated similar problems, such as how subjects in distinct disease groups show distinct coexpression patterns, contributing to a better understanding of disease at a molecular level (De la Fuente (2010)). Here, we focus on understanding how coexpression changes with quantitative traits, not discrete conditions such as disease status, and hence we study the dynamic nature of coexpression. As an example of a quantitative trait, we use genetic ancestry. Ancestry is known to play a critical role in other molecular phenotypes including DNA methylation and gene expressions (Galanter et al. (2017); Price et al. (2008)), and so we believe it has an important role in gene dynamics and gene networks as well. In this paper, we study how the local connectivity of transcription factor genes changes with ancestry. Specifically, we study the gene coexpression of African American subjects to identify candidate transcription factors whose effects on their targets vary with the proportion of African ancestry in their genome. This analysis will lead to a better comprehension of how genes are differentially regulated in distinct populations.

The above biological problem can be investigated using multivariate statistical models of gene expression with a covariance structure (characterizing connectivity) that depends on one or more features (such as ancestry). This paper focuses on testing the contribution of ancestry on the covariance matrix, and we start from its simplest form by studying the expression levels of two genes. We construct a statistical model that can explain how their correlation varies against genetic ancestry and use that to test if the correlation is constant across conditions. We generalize it to the local connectivity of a transcription factor by meta-analyzing the pairwise statistics. Note that covariance modeling for multivariate data is important in many applications outside the field of genetics. Variance modeling has been widely studied in the context of heteroskedasticity (Breusch and Pagan (1979); Glejser (1969); White et al. (1980)), and correlation modeling under discrete conditions has been studied in the context of the differential network (Ideker and Krogan (2012)), but dynamic correlation modeling has been less explored.

Li (2002) and Li et al. (2004) addresses the most similar scientific problem to ours, using the term "liquid association" (LA) to conceptualize the internal evolution of the coexpression pattern for a pair of genes. They analyze the coexpression that changes across different unobserved cellular states that are represented by the expression level of another gene as a proxy. Other studies have built on the liquid association to better identify cell states that affect coexpression (Yan et al. (2017); Yu (2018)), most focusing on expanding the test to genome-scale. However, methods based on liquid association have some limitations. First, it restricts the covariate to be a 1-dimensional vector, and cannot be generalized to more realistic scenarios. Second, it treats the covariate as a random variable that follows a normal distribution, which genetic ancestry does not, so it cannot be used for our application. Third, it only tests the linear relationship between the covariate and the coexpression. Lastly, the corresponding test statistic does not have a closed-form null distribution and requires a permutation test, leading to computational inefficiency.

We propose a methodology for the continuously-varying covariance problem. We transform the likelihood function of bivariate normal variables to effectively change the multivariate covariance modeling problem to a univariate variance modeling problem. We then apply a traditional score test for heteroskedasticity (Breusch and Pagan (1979)) where the null hypothesis is that the coexpression does not vary with the covariate. This method is generalizable to non-normal, multivariate covariates, and it is also applicable to a non-linear relationship between the variance and the covariate. Moreover, the score test statistic asymptotically follows a chi-squared distribution, and hence it is easily expandable to a large number of tests without excessive computational burden. Subsequently, we tackle the local connectivity problem by expanding the scope of the problem from the relationship of two genes to the relationships between one gene and multiple genes by combining the test statistics. When the number of genes in the local cluster is smaller than the sample size, the desired statistical properties apply to the new combined test statistic as well.

The rest of the paper is organized as follows. First, we lay out the framework for the score test that investigates whether the covariance between bivariate normal variables varies against a continuous covariate X. Then we propose a way to combine the pair-wise test statistics for one gene and test the global null that the local connectivity of one variable does not change with genetic ancestry. In the simulation section, we show that the proposed method has distinct advantages compared to alternatives such as the likelihood ratio test or liquid association. Finally, we share our real data analysis results using Gene-Tissue Expression (GTEx) data for African Americans' transcriptome and genome. We end with a discussion about limitations of the method, possible future directions, and potential applications to fields outside genetics.

2. Methods.

2.1. *Test for connectivity between two genes .* Consider 2-dimensional data for N subjects $\boldsymbol{y}_i \in \mathbb{R}^2$, $i = 1, 2, \cdots, N$ independently following the bivariate normal distribution. There are two covariate matrices $Z \in \mathbb{R}^{N\times R}$ and $X \in \mathbb{R}^{N\times P}$, each for the mean term and the variance term, respectively. They are both assumed to be full rank. We notate each element of X and Z as $\{x_{ip}\}_{i=1,p=1}^{N,P}$ and $\{z_{ir}\}_{i=1,r=1}^{N,R}$.

$$\begin{gathered}\begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{z}_i^T \boldsymbol{\beta}_1 \\ \boldsymbol{z}_i^T \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} u_{i1} \\ \hat{u}_{i2} \end{bmatrix} \\ \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{12}(\boldsymbol{x}_i) \\ \rho_{12}(\boldsymbol{x}_i) & \sigma_2^2 \end{bmatrix} \right)\end{gathered} \tag{2.1}$$

$\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are coefficients with length R for the mean term Z. σ_1^2 and σ_2^2 are fixed scalars but ρ_{12} varies with the covariate. We define $\boldsymbol{\alpha} \in \mathbb{R}^P$ and a scalar α_0 as the linear coefficients to model ρ_{12} as follows.

$$\rho_{12}(\boldsymbol{x}_i) = \rho_{12}(\boldsymbol{x}_i^T \boldsymbol{\alpha} + \alpha_0) \tag{2.2}$$

Our parameter of interest is $\boldsymbol{\alpha}$ while all others — α_0, $\boldsymbol{b}$, $\boldsymbol{\beta}$, σ_1^2, σ_2^2 — are nuisance parameters. Our goal is to develop a method that tests the following null hypothesis

$$H_0 : \boldsymbol{\alpha} = \boldsymbol{0}. \tag{2.3}$$

Under the null hypothesis, $\rho_{12}(\boldsymbol{x}_i^T \boldsymbol{\alpha} + \alpha_0) = \rho_{12}(\alpha_0)$ is a constant regardless of $\boldsymbol{x}_i$. The linearity and additivity assumptions in (2.2) are standard, and since ρ_{12} can take any non-linear form, (2.2) still is a flexible framework. Before we introduce the detailed methodology, we first present our result below.

PROPOSITION 1. *Consider the model in (2.1). If the non-diagonal term of Σ follows the function ρ_{12} as defined in (2.2), the score statistic does not depend on the unknown function ρ_{12} and follows χ_P^2 under the null hypothesis (2.3).*

The model (2.1) is close to a multivariate regression model of $\boldsymbol{y}_i$ against the mean term $\boldsymbol{z}_i$ with intercept $\boldsymbol{b}_0$, slope $\begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 \end{bmatrix}$, and error term $\boldsymbol{u}_i$. The difference is that its error variance depends on the covariate $\boldsymbol{x}_i$. The function ρ_{12} represents an unknown form of heteroskedasticity between the two variables 1 and 2. In the context of gene coexpression of African Americans, $\boldsymbol{y}_i$ is gene expression level of an African American individual i at two selected genes, and $\boldsymbol{x}_i$ is a P-dimensional covariate matrix for individual i that holds information about genetic ancestry. It can be a scalar that represents the proportion of African ancestry in the genome, a vector of the first few principal components of the genotypes, or a vector of local ancestry at multiple loci. In the application example in section 4, we focus on scalar $\boldsymbol{x}_i$ for straightforward interpretability.

We have two well-known tools to test the null hypothesis (2.3): likelihood ratio test and Rao's score test (Breusch and Pagan (1979)).

The likelihood ratio test has a few disadvantages compared to the score test. It requires the full specification of the function ρ to estimate the maximum likelihood estimate (MLE) of $\boldsymbol{\alpha}$ both under the null hypothesis and under the alternative hypothesis. Possible functions of ρ_{12} are any kind of sigmoid function bound to $(-\sqrt{\sigma_1^2\sigma_2^2}, \sqrt{\sigma_1^2\sigma_2^2})$ such as logistic function, hyperbolic tangent function, or any cumulative distribution supported on the whole real line. This modeling strategy leads to one disadvantage; as mentioned in the previous section, we would like to impose as few assumptions on the specific form of heteroskedasticity as possible. If ρ_{12} is highly misspecified, we sacrifice the statistical power. Another disadvantage is that most of the reasonable assumptions of ρ, such as the sigmoid functions mentioned above, do not lead to a closed form MLE of $\boldsymbol{\alpha}$ under the alternative hypothesis. It would require us to numerically optimize the likelihood, leading to computational inefficiency, especially when the test space is large as in our application of gene coexpression.

On the other hand, Rao's score test, unlike the likelihood ratio test, only requires the MLE of $\boldsymbol{\alpha}$ under the null hypothesis (Rao and Statistiker (1973)). Moreover, under our linear and additive model ($\rho_{12}(\boldsymbol{x}_i) = \rho_{12}(\boldsymbol{x}_i^T\boldsymbol{\alpha})$), the test statistic does not depend on the form of ρ_{12} while maintaining its asymptotic property as long as ρ_{12} is twice differentiable. In order to test (2.3), we expand the result from Breusch and Pagan (1979) to derive the test statistic.

The Fisher Information of the model (2.1) is block diagonal: one block for mean parameters $\boldsymbol{b}$, $\boldsymbol{\beta}$ and another block for variance parameters σ_1^2,

σ_2^2, α_0, and $\boldsymbol{\alpha}$. Therefore, to derive the score statistic, we only need the first and second derivatives of the log likelihood with respect to the variance parameters. Moreover, we can plug in the MLEs of the nuisance parameters to get the score statistic. So, we replace u_{i1} with the OLS residuals $\hat{u}_{i1}$ from the regression of (2.1). The score statistic is as follows, and the detailed derivation is in the Appendix A.

(2.4)

$$q_{12} = \frac{1}{N(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\rho}_{12}^2)} \sum_{i=1}^{N}(\hat{u}_{i1}\hat{u}_{i2} - \hat{\rho}_{12})\tilde{\boldsymbol{x}}_i^T(\sum_{i=1}^{N} \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T)^{-1}\sum_{i=1}^{N}(\hat{u}_{i1}\hat{u}_{i2} - \hat{\rho}_{12})\tilde{\boldsymbol{x}}_i$$

where $\hat{\rho}_{12} = \rho_{12}(\hat{\alpha}_0) = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_{i1}\hat{u}_{i2}$, and $\tilde{\boldsymbol{x}}_i$ is the covariate with the intercept term: $\begin{bmatrix}1 & \boldsymbol{x}_i\end{bmatrix}^T$. The first derivative of ρ'_{12} has been canceled out, and therefore the score statistic does not depend on the function ρ_{12}; it only depends on the MLE of the covariance $\hat{\rho}_{12}$. Every component of the test statistic is easily acquired from the data, and the computational burden is low. Just as importantly, it is flexible as it allows any form of ρ_{12}. Under this setting, q asymptotically follows χ_P^2 (Breusch and Pagan (1979)).

However, even though the introduced test statistic has convenient asymptotic properties, the inference might not be correct under finite sample size. The error is in the order of N^{-1} (Harris (1985)), and many Monte Carlo experiments show that the test rejects the null hypothesis less frequently than indicated by its nominal size (Godfrey (1978); Griffiths and Surekha (1986). In response, corrections have been suggested (Cribari-Neto and Ferrari (2001); Harris (1985)), and we apply the method from Honda (1988) to ensure the validity of the asymptotic properties even under the small sample size. The details of the small-sample correction, as well as the detailed derivation of the test statistic, are shown in the Appendix.

2.2. *Test for Local Connectivity .* In section (2.1), we proposed the statistic q to test a pair of variables 1 and 2 to measure the evidence that their correlation changes with the covariate X. As a natural extension to the pairwise test statistic, we can repeat the procedure for more than 2 variables. In particular, we can study one transcription factor with multiple target genes to test whether the local connectivity of the transcription factor varies with genetic ancestry. In this section, we propose a way to combine the test statistics to test a new global null hypothesis. The global null hypothesis for variable 1 extends (2.3) as follows,

$$\boldsymbol{H}_0^{(1)} : \boldsymbol{\alpha}_{12} = \boldsymbol{\alpha}_{13} = \cdots = \boldsymbol{\alpha}_{1K} = \boldsymbol{0}, \tag{2.5}$$

where the superscript in $\boldsymbol{H}_0^{(1)}$ indicates that the null hypothesis applies to variable 1. Under $\boldsymbol{H}_0^{(1)}$, no other variables' correlation with variable 1 changes across the different values of X. We believe testing the global null (2.5) improves the statistical power when the "hot spot" variables or "hub" variables are connected to a lot of other nodes forming cliques or modules. In the context of gene coexpression network, we know that transcription factors regulate the gene expression of multiple genes, and if one transcription factor varies with respect to the covariate, the transcriptions of its regulated genes are likely to be correlated with the covariate as well.

We propose a way to combine the test statistics to test (2.5). Chen et al. (2012) discusses two ways to construct the alternative hypothesis for testing the global null. One way, called a sparse alternative, is to test whether only a small number among all tests have non-zero effects while all other tests are null. Another way is to test if at least one test has a non-zero effect size. Based on our prior knowledge in biology and coexpression network, we assume that there are many small signals instead of few big ones, so we choose the latter. We propose a simple linear combination of the test statistics

$$d_1 = q_{12} + q_{13} + \cdots + q_{1K} = \sum_{k=2}^{K} q_{1k}. \tag{2.6}$$

Note that q has been constructed from normalized data of w and v. It is, therefore, the most natural to add them without additional weighting. We believe the statistic d_1 in (2.6) improves the statistical power because even if the effect sizes for each gene pair may be too small to be detected, when combined, they can form a stronger signal.

We derive the null distribution of d_1 to test (2.5). Although each q_{1k} follows χ^2_P, they are correlated to one another, so the null distribution is not trivial. We devise a way to simulate the null distribution as a finite sum of Gamma distributions so that we can estimate the empirical quantile of d_1. We therefore propose the following.

PROPOSITION 2. *Under the setting of Proposition 1, and two additional assumptions that (1) the covariates have been orthogonalized so that* $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T$ *is a* P *by* P *identity matrix and that (2) none of the variables* $y_1, \cdots, y_K$ *are perfectly correlated,* d_1 *asymptotically follows the finite sum of Gamma distributions as defined in (2.11) under the global null hypothesis (2.5).*

To derive the empirical null distribution of d_1, We start by re-writing the pairwise score statistic q (omitting gene pair index for now) as a sum of χ^2_1 variables as below. Our null hypothesis tests for all covariates at the same time, so we can orthogonalize X to make $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T$ an identity matrix without affecting the testing procedure. Let $\tilde{X}$ be the orthogonalized covariate matrix, and $\tilde{x}_{ip}$ be the corresponding entries with $\sum_{i=1}^{N} \tilde{x}_{ip} = 0$ and $\sum_{i=1}^{N} \tilde{x}_{ip}^2 = n$. Then (2.4) can be alternatively written as follows, where we define r_p.

$$q = \sum_{p=1}^{P} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{x_{ip}\left(\hat{u}_{i1}\hat{u}_{i2} - \hat{\rho}_{12}\right)}{\sqrt{\hat{\sigma}_1^2 \hat{\sigma}_2^2 + \hat{\rho}_{12}^2}} \right)^2 = \sum_{p=1}^{P} r_p^2 \tag{2.7}$$

For each p, r_p asymptotically follows the standard normal distribution by the Central Limit Theorem.

Now, we acquire a closed-form covariance structure of r. First, we begin with a multivariate central limit theorem to write the following in terms of r.

$$\boldsymbol{r}_{1,p} = \begin{bmatrix} r_{12,p} \\ r_{13,p} \\ \cdots \\ r_{1K,p} \end{bmatrix} \rightarrow N_{K-1}(\mathbf{0}, H_1) \qquad \forall p = 1, \cdots, P \tag{2.8}$$

H_1 is a $(K-1) \times (K-1)$ matrix where $(k-1, \ell-1)$th element is $\eta_{1k,1\ell}$ for $k, \ell = 2, \cdots K$. From (2.7), it is easy to see that H_1 has 1 at the diagonals. Also, $\eta_{1k,1\ell}$ converges in probability to

$$\frac{\hat{\rho}_{23} + \hat{\rho}_{12}\hat{\rho}_{13}}{\sqrt{(\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\rho}_{12}^2)(\hat{\sigma}_1^2\hat{\sigma}_3^2 - \hat{\rho}_{13}^2)}} \tag{2.9}$$

Note that d_1 can be written as the sum of L2 norm of $\boldsymbol{r}_{1,p}$ with a known distribution,

$$d_1 = \sum_{p=1}^{P} \|\boldsymbol{r}_{1,p}\|_2^2 = \sum_{p=1}^{P}\sum_{k=2}^{K} r_{1k,p}^2. \tag{2.10}$$

Let $H_1 = U_1 \Lambda_1 U_1^T$ be the eigen-decomposition of the covariance matrix H_1 in (2.8), where the diagonal matrix Λ has eigenvalues $\lambda_{12}, \cdots, \lambda_{1K}$ in a decreasing order. Then, we can next consider the transformation $\boldsymbol{r}_{1,p}^* = U\boldsymbol{r}_{1,p}$ that follows normal distribution with diagonal covariance matrix Λ_1.

Note that $\|\boldsymbol{r}_{1,p}\|_2^2 = \|U\boldsymbol{r}_{1,p}\|_2^2$ due to the orthogonal invariance of L2 norm. Then,

$$\begin{aligned} &\sum_{p=1}^{P} {r^*_{1k,p}}^2 \sim \Gamma\left(\frac{P}{2}, \frac{\lambda_{1k}}{2}\right), \qquad k = 2, \cdots, K \\ &d_1 = \sum_{p=1}^{P} {r^*_{12,p}}^2 + \cdots + \sum_{p=1}^{P} {r^*_{1K,p}}^2 \sim \sum_{k=2}^{K} \Gamma\left(\frac{P}{2}, \frac{\lambda_{1k}}{2}\right) \end{aligned} \tag{2.11}$$

Assuming that we know the true, symmetric, positive definite H_1, we can acquire positive λ_{1k} for $k = 2, \cdots, K$, and we have expressed the null distribution of d_1 as the sum of distributions of independent gamma variables. We can computationally simulate this null distribution easily. Alternatively, Moschopoulos (1985) provides another interpretation by expressing the cumulative distribution in a form of infinite sum, but the method is inconvenient in practice.

In (2.9), we define the element-wise mapping $\phi : \Sigma \to H$. Although we do not know true Σ in practice, we can use its maximum likelihood estimate instead as we do with other nuisance parameters. We can see from the construction of H that a well-conditioned, symmetric, positive definite estimate of $\hat{\Sigma}$ leads to a symmetric and positive-definite $\phi(\hat{\Sigma})$ as well. When N is sufficiently larger than K, empirical covariance matrix $\hat{\Sigma}$ of $\begin{bmatrix} y_{1i} & y_{2i} & y_{3i} \end{bmatrix}$ is a consistent estimator for Σ, and replacing Σ with $\hat{\Sigma}$ in computing H can guarantee that the test statistic in (2.6) converges in distribution to (2.11).

However, when K is larger than n, an accurate estimation of Σ is a difficult problem. We therefore turn to a permutation test and shuffle the covariate X to test against the true response data that preserves the correlation structure of the network while maintaining the dependence structure of Y. The permutation procedure is valid under the assumptions in (2.1).

We use the sequential precision-improvement permutation test, similar to one suggested by Chen et al. (2012). Permutation test often results in a limited resolution of p-values which can lead to imprecise inference especially when we need to correct for the testing of multiple hypotheses. Meanwhile, performing a large number of permutations for every can be computationally wasteful. To find balance, we terminate the permutation procedure if the signal is not strong enough.

The detailed procedure is as follows. For every permutation $b = 1, \cdots, B$, we permute the rows (samples) of the covariate matrix X to create X_b. Then we compute d_{kb} using X_b and the data matrix Y. Then, d_{kb} should follow the null distribution, and the p-value for d_k should be computed as a quantile of d_k compared to the empirical distribution of d_{kb} for each b. After the minimum number of permutations pre-defined by the user (1000 in our case), we count the number of permutations where d_{kb} is larger than d_k. If there are two or more such cases, we terminate the permutation procedure early. Most genes fall into this category leading to p-value greater than 0.002. If there are less than 5 such cases observed, we iteratively perform 100 more permutations and re-check the number of d_{kb} with values larger than d_k. We repeat until the number of permutations B reaches the predefined maximum number of permutation (10^5 in our case), which is designed to give a good enough resolution of p-value given the number of tests that we are performing.

3. Simulation Studies.

Here, we evaluate the proposed method through simulations. We focus on the pairwise analysis and compare the performance of the proposed score test with two other alternatives - liquid association and the likelihood ratio test.

First, we check the calibration of test statistics under the null hypothesis. We sample X from the univariate standard normal distribution to match the required setting of liquid association. We simulate the data matrix Y from

$$\boldsymbol{y}_i \sim \mathcal{N}_2\left(\boldsymbol{b}_0 + \boldsymbol{z}_i^T\boldsymbol{\beta}, \begin{bmatrix} 1 & \bar{\rho} \\ \bar{\rho} & 1 \end{bmatrix}\right)$$

where $\bar{\rho}$ was randomly selected from uniform distribution ranging from -1 and 1 and each element of $\boldsymbol{b}_0$ and $\boldsymbol{\beta}$ from standard normal distribution. We use $\boldsymbol{z}_i = \boldsymbol{x}_i$. We test different sample sizes of $N = 500, 100, 30$ to check the behavior of each method under the null hypothesis. For each N, we sample X once, and generate Y 1,000 times. The likelihood ratio test was designed to assume hyperbolic tangent model for ρ,

$$\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}) = \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\alpha}} - 1}{e^{\boldsymbol{x}_i^T\boldsymbol{\alpha}} + 1}, \tag{3.1}$$

which is the inverse of Fisher transformation, $\frac{1}{2}\boldsymbol{x}_i^T\boldsymbol{\alpha} = \frac{1}{2}log\left(\frac{1+\rho}{1-\rho}\right)$. Fisher-transformed ρ asymptotically follows normal distribution, so it works well when X is drawn from normal distribution. We use *optim* function in R to

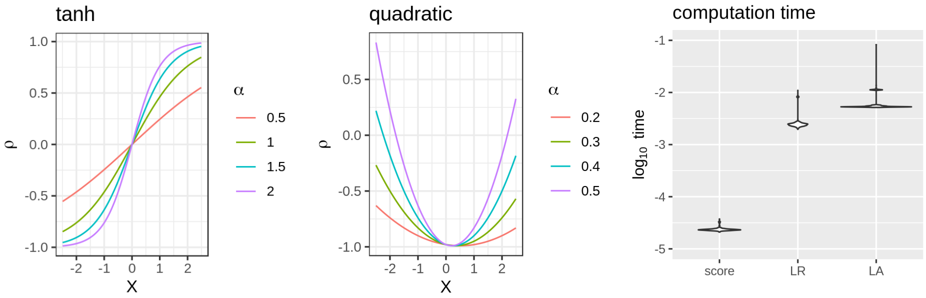

Fig 1. *The first and second plots show the two example functions ρ used to generate data. The third plot shows for each method the distribution of the 1000 separate run-times for computing a single score test. The proposed method performs better than likelihood ratio test in the scale of* 10^2.

find $\hat{\boldsymbol{\alpha}}_{\text{MLE}}$ under the alternative hypothesis.

The results show that all three methods control the type I error at the nominal size well, where score and likelihood ratio test statistics both follow χ^2_1 closely.

Next we generate the data under the alternative hypothesis to compare the statistical power. So we focus on the case of $N = 70$ to reflect the sample size of GTEx data. We again draw X from standard normal distribution. Then, for $i = 1, \cdots, N$, we generate $\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})$ from hyperbolic tangent function in (3.1). Given ρ, we draw Y from (2.1) with varying levels of α, 1000 times each. The hyperbolic tangent model places the likelihood ratio test at an advantage because the model is correctly specified, so as a contrasting case, we use a quadratic model to generate ρ as follows,

$$\rho(\alpha_0 + \boldsymbol{x}_i^T\boldsymbol{\alpha}) = (-0.1 + \boldsymbol{x}_i^T\boldsymbol{\alpha})^2 - 0.99, \tag{3.2}$$

where subtracting 0.99 is to ensure numerical stability. Since the likelihood ratio test assumes a wrong model, it is expected to lose power. Also, since quadratic function is highly non-linear, liquid association is expected to have poor performance as well. Figure 1 (a) and (b) show the shape of ρ with respect to X with varying levels of α.

Table 1 summarizes the result. It counts the proportion of simulations which showed p-values less than 0.05 out of 1,000 total simulations. When ρ is generated from hyperbolic tangent function, likelihood ratio test generally

TABLE 1

Proportion of simulations for each method that showed p-value < 0.05 at given data generating model and α level. We use two functions for ρ, hyperbolic tangent and quadratic function. The likelihood ratio test was conducted under the assumption that ρ is hyperbolic tangent function.

ρ			tanh				quadratic	
α	0	0.25	0.5	0.75	1	0.2	0.3	0.4	0.5
score	0.052	0.181	0.467	0.806	0.928	0.764	0.660	0.604	0.553
LA	0.051	0.181	0.472	0.779	0.911	0.048	0.054	0.054	0.052
LR	0.056	0.199	0.544	0.885	0.979	0.643	0.472	0.398	0.344

outperforms the other two methods, as expected, since the model is correctly specified in LR test. Score test, although it does not assume any model on ρ, does not lose as much power as liquid association does. Meanwhile, when ρ is generated from quadratic function, score function clearly outperforms the other two methods. Hence, the simulations show that the proposed score test is robust to the shape of heteroskedasticity. Figure 1 (c) shows the distribution of computation times of each method to compute the test statistic once in the scale of log_{10} for 1000 simulations under quadratic model with $\alpha = 0.5$. The score test is the most efficient, because the likelihood ratio test requires numerical estimation of MLEs both under the null and under the alternative hypothesis while liquid association requires permutation test for inference. The simulations were done sequentially (non-parallel) with LAMBDA QUAD workstation with Intel Xeon W-2175 processor.

4. Applications to GTEx Data. We next apply this method to 71 African American samples from GTEx Consortium et al. (2015) whose expression level data for muscle skeletal tissue is available. We aim to find transcription factors that change their coexpression pattern with their target genes as the genome's proportion of African ancestry changes. We acquire a list of transcription factors from TF checkpoint database in Chawla et al. (2013). We also acquire a list of target genes for each transcription factors from TF2DNA database in Pujato et al. (2014). We only take into consideration target genes with the highest binding scores. We use the proportion of African ancestry for each individual as the covariate, and it is inferred from the genotype data using the software LAMP by Paşaniuc, Kennedy and Măndoiu (2009). The proportion of African ancestry levels from the subjects range from 14% to 96%.

For each of the $k = 1, \cdots, K = 848$ transcription factor encoding gene, we compute the pair-wise statistic q_{kj} for all its target genes $j = 1, \cdots, J_k$, where J_k is the number of target genes for each transcription factor k. Then, we compute $d_k = \sum_{j=1}^{J_k} q_{kj}$ to test the hypothesis that the correlation between the transcription factor k and its targets remain the same across different genetic industry. We first divide d_k with the number of targets J_k to compute the average score of all the target genes for the given TF k, and we make a heuristic comparison against χ_1^2 distribution. Under the null hypothesis, the expectation of d_k/J_k is 1, although the variance is not trivial due to high dependence. Then, we choose 10 genes with the top d_k/J_k values to perform the permutation test.

Table 2 summarizes the top 5 transcription factors with the highest average d_k values and their p-values computed from sequential permutation tests. The adjusted p-values were computed by

$$\text{adjusted } p = p \times (\text{number of transcription factors})/(\text{rank of } p).$$

Gene Name	p-value	Adjusted p-value
MECOM	$< 10^{-5}$	< 0.01
ZNF423	$< 10^{-5}$	< 0.01
SPIB	0.001	0.326
ZNF618	0.002	0.339
ZSCAN	0.002	0.415

TABLE 2
Top 5 transcription factors with the lowest p-value

Two genes, MECOM and ZNF423, maintain their significance level of 0.01 under the Bonferroni criterion for 912 transcription factors. For the top gene MECOM, the two highest contributing target genes were C3orf70 and P2RY1, each having scores of 26.74 and 25.69 (should follow χ_1^2 under the null hypothesis). Although these values look very high, they do not pass the Bonferroni test after taking into account the number of tests. The significance of MECOM (and ZNF423) is achieved through combining the scores across all the target genes.

5. Discussion. We proposed a method to test whether the covariance between bivariate normal variables changes with continuous covariates. We further expanded our scope of analysis by looking at local connectivity — how one variable's connectivity with multiple other variables change with continuous covariates. We provided a real data example by identifying major transcription factor genes that are differentially connected with its targets by African Americans' genetic ancestry.

Our method is more flexible than other alternatives for covariance testing, but it still has some limitations. First, when there are more variables than the available sample size, as often is the case for many modern data sets, we ultimately turn to a permutation test, not being able to take the computational advantage of the theoretical results. There could be other ways, depending on contexts, to estimate Σ to avoid the permutation test. For example, one could impose the sparsity assumption on the covariance matrix, and that can lead to a reliable estimate of Σ and subsequently H in (2.8). Such assumption is too restrictive in our context, but in other applications, one can take the liberty to make structural assumptions on the covariance matrix. Nonetheless, it is challenging, to say the least, to find consistent estimators for the eigenvalues of Σ in high-dimensional data. Second, the score statistic q was derived under the normality assumption of the data set, although simulations show that the result is quite robust to distribution mis-specification. These limitations of the methods propose possible future research topics: (1) better way to estimate H in (2.8) when $K > N$ to preserve asymptotic results, and (2) a non-parametric version of correlation analysis that can generalize to any underlying distributions.

In practice, users of this method should note the method's sensitivity to correct mean modeling. For example, if correct data generating procedure is $\boldsymbol{z}_i = \boldsymbol{x}_i + \boldsymbol{x}_i^2$, missing the quadratic term (mis-specifying $\boldsymbol{z}_i$ as $\boldsymbol{x}_i$) would incorrectly assign the effects of $\boldsymbol{x}_i^2$ to the residuals. Then, the score statistic computed from the residuals will have some remaining effects from the mean term, and hence the type I error will be inflated. Therefore, the users are expected to be careful in regressing out all the relevant terms so that the mean term effects do not spill over to the variance term.

We believe the proposed method can be applied to data problems in diverse domains, especially where certain hub variables are connected to many other variables as in the transcriptional regulatory network. Many network problems, such as protein interactions, metabolic networks, co-authorship network, and semantic network, are known to have, or have something close to, scale-free topology, indicating that the important variables in those networks can be tested against other variables. The proposed correlation analysis will provide insights into the building blocks of diverse network problems by looking at the pairwise and more than pairwise relationships among the variables.

References.

BREUSCH, T. S. and PAGAN, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society* 1287–1294.

CHAWLA, K., TRIPATHI, S., THOMMESEN, L., LÆGREID, A. and KUIPER, M. (2013). TFcheckpoint: a curated compendium of specific DNA-binding RNA polymerase II transcription factors. *Bioinformatics* **29** 2519–2520.

CHEN, L. S., HSU, L., GAMAZON, E. R., COX, N. J. and NICOLAE, D. L. (2012). An exponential combination procedure for set-based association tests in sequencing studies. *The American Journal of Human Genetics* **91** 977–986.

CONSORTIUM, G. et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348** 648–660.

CRIBARI-NETO, F. and FERRARI, S. L. (1995). An improved Lagrange multiplier test for heteroskedasticity. *Communications in Statistics-Simulation and Computation* **24** 31–44.

CRIBARI-NETO, F. and FERRARI, S. L. (2001). Monotonic improved critical values for two $\chi 2$ asymptotic criteria. *Economics Letters* **71** 307–316.

DE LA FUENTE, A. (2010). From differential expressionto differential networking–identification of dysfunctional regulatory networks in diseases. *Trends in genetics* **26** 326–333.

GALANTER, J. M., GIGNOUX, C. R., OH, S. S., TORGERSON, D., PINO-YANES, M., THAKUR, N., ENG, C., HU, D., HUNTSMAN, S., FARBER, H. J. et al. (2017). Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. *Elife* **6** e20532.

GLEJSER, H. (1969). A new test for heteroskedasticity. *Journal of the American Statistical Association* **64** 316–323.

GODFREY, L. G. (1978). Testing for multiplicative heteroskedasticity. *Journal of Econometrics* **8** 227–236.

GRIFFITHS, W. and SUREKHA, K. (1986). A Monte Carlo evaluation of the power of some tests for heteroscedasticity. *Journal of Econometrics* **31** 219–231.

HARRIS, P. (1985). An asymptotic expansion for the null distribution of the efficient score statistic. *Biometrika* **72** 653–659.

HONDA, Y. (1988). A size correction to the Lagrange multiplier test for heteroskedasticity. *Journal of Econometrics* **38** 375–386.

IDEKER, T. and KROGAN, N. J. (2012). Differential network biology. *Molecular systems biology* **8**.

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

LI, K.-C., LIU, C.-T., SUN, W., YUAN, S. and YU, T. (2004). A system for enhancing genome-wide coexpression dynamics study. *Proceedings of the National Academy of Sciences* **101** 15561–15566.

LONSDALE, J., THOMAS, J., SALVATORE, M., PHILLIPS, R., LO, E., SHAD, S., HASZ, R., WALTERS, G., GARCIA, F., YOUNG, N. et al. (2013). The genotype-tissue expression (GTEx) project. *Nature genetics* **45** 580.

PAŞANIUC, B., KENNEDY, J. and MĂNDOIU, I. (2009). Imputation-based local ancestry inference in admixed populations. In *International Symposium on Bioinformatics Research and Applications* 221–233. Springer.

Price, A. L., Patterson, N., Hancks, D. C., Myers, S., Reich, D., Cheung, V. G. and Spielman, R. S. (2008). Effects of cis and trans genetic ancestry on gene expression in African Americans. *PLoS genetics* **4** e1000294.

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

Pujato, M., Kieken, F., Skiles, A. A., Tapinos, N. and Fiser, A. (2014). Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. *Nucleic acids research* **42** 13500–13512.

Rao, C. R. and Statistiker, M. (1973). *Linear statistical inference and its applications* **2**. Wiley New York.

White, H. et al. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica* **48** 817–838.

Yan, Y., Qiu, S., Jin, Z., Gong, S., Bai, Y., Lu, J. and Yu, T. (2017). Detecting subnetwork-level dynamic correlations. *Bioinformatics* **33** 256–265.

Yu, T. (2018). A new dynamic correlation algorithm reveals novel functional aspects in single cell and bulk RNA-seq data. *PLoS computational biology* **14** e1006391.

Appendix A. Derivation of test statistic. We start from the likelihood of (2.1).

$$\begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_i^T \boldsymbol{\beta}_1 \\ \boldsymbol{z}_i^T \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix}$$

$$\begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}(\tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}}) \\ \rho_{12}(\tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\alpha}}) & \sigma_2^2 \end{bmatrix} \right)$$

where $\tilde{\boldsymbol{\alpha}} = \begin{bmatrix} \alpha_0 & \boldsymbol{\alpha}^T \end{bmatrix}$, and $\tilde{\boldsymbol{x}}_i$ is $\begin{bmatrix} 1 & \boldsymbol{x}_i \end{bmatrix}^T$. We compute the first and second derivative of the log likelihood with respect to $\tilde{\boldsymbol{\alpha}}$ when all other nuisance parameters are replaced with their respective maximum likelihood estimates. $\tilde{\boldsymbol{\alpha}}$ is similarly replaced with $\begin{bmatrix} \hat{\alpha}_0 & \boldsymbol{0} \end{bmatrix}^T$ so that $\rho_{12}(\hat{\alpha}_0)$ is equal to $\frac{1}{N}\sum_{i=1}^N \hat{u}_{i1}\hat{u}_{i2}$. $\hat{u}_{i1}$ and $\hat{u}_{i2}$ is the OLS residuals of the above regression under the null hypothesis.

Score function. We first derive the first derivative of the log likelihood with respect to $\boldsymbol{\alpha}$. We start from differentiating with respect to the variance matrix Σ.

$$\frac{\partial \ell}{\partial \Sigma} = -\frac{1}{2} \sum_{i=1}^{N} \left(\frac{\partial log|\Sigma|}{\partial \Sigma} + \frac{\partial \boldsymbol{u}_i^T \Sigma^{-1} \boldsymbol{u}_i}{\partial \Sigma} \right)$$

Here, we use ρ_{12} instead of $\rho_{12}(\boldsymbol{x}_i^T \boldsymbol{\alpha})$ for notational convenience. Similarly, $\hat{\rho}_{12}$ is used instead of $\rho_{12}(\hat{\alpha}_0)$. The first component is as follows:

$$\frac{\partial log(|\Sigma|)}{\partial \sigma_1^2} = \frac{\sigma_2^2}{\sigma_1^2\sigma_2^2 - \rho_{12}^2}, \quad \frac{\partial log(|\Sigma|)}{\partial \sigma_2^2} = \frac{\sigma_1^2}{\sigma_1^2\sigma_2^2 - \rho_{12}^2}, \quad \frac{\partial log(|\Sigma|)}{\partial \alpha} = \frac{-2\rho_{12}\rho_{12}'}{\sigma_1^2\sigma_2^2 - \rho_{12}^2} x_i$$

For the second component, we first compute the following:

$$\begin{aligned} \boldsymbol{u}_i^T \Sigma^{-1} \boldsymbol{u}_i &= \frac{1}{\sigma_1^2\sigma_2^2 - \rho_{12}^2} \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} \sigma_2^2 & -\rho_{12} \\ -\rho_{12} & \sigma_1^2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ &= \frac{1}{\sigma_1^2\sigma_2^2 - \rho_{12}^2} (u_1^2\sigma_2^2 + u_2^2\sigma_1^2 - 2u_1u_2\rho_{12}) \end{aligned}$$

Then, we can get the second component as follows,

$$\begin{aligned}
\frac{\partial u^T\Sigma^{-1}u}{\partial\sigma_1^2} &= \sum_{i=1}^{N}\frac{u_{i2}^2(\sigma_1^2\sigma_2^2-\rho_{12}^2)-(u_{i2}^2\sigma_1^2+u_{i1}^2\sigma_2^2-2u_{i1}u_{i2}\rho_{12})\sigma_2^2}{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^2}\\
\frac{\partial u^T\Sigma^{-1}u}{\partial\sigma_2^2} &= \sum_{i=1}^{N}\frac{u_{i1}^2(\sigma_1^2\sigma_2^2-\rho_{12}^2)-(u_{i2}^2\sigma_1^2+u_{i1}^2\sigma_2^2-2u_{i1}u_{i2}\rho_{12})\sigma_1^2}{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^2}\\
\frac{\partial u^T\Sigma^{-1}u}{\partial\alpha} &= \sum_{i=1}^{N}\frac{-2u_{i1}u_{i2}\rho_{12}'(\sigma_1^2\sigma_2^2-\rho_{12}^2)-(u_{i2}^2\sigma_1^2+u_{i1}^2\sigma_2^2-2u_{i1}u_{i2}\rho_{12})(-2\rho_{12})(\rho_{12}')}{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^2}\boldsymbol{x}_i\\
&= \sum_{i=1}^{N}\rho_{12}'\frac{4\sigma_1^2\sigma_2^2\rho_{12}-2u_{i1}u_{i2}(\sigma_1^2\sigma_2^2-\rho_{12}^2)}{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^2}
\end{aligned}$$

Plugging in the MLEs for each nuisance parameter, we can replace σ_1^2 with $\hat{\rho}_{11}=\sum_{i=1}^N u_{2i}^2$ and σ_2^2 with $\hat{\rho}_{22}=\sum_{i=1}^N \hat{u}_{i2}^2$. We also replace $\boldsymbol{\alpha}$ with $\mathbf{0}$ as specified in H_0, and $\rho_{12}(\hat{\alpha}_0)$ is equal to $\sum_{i=1}^N \hat{u}_{i1}\hat{u}_{i2}$.

$$\tilde{d}_{\sigma_1^2} = \sum_{i=1}^{N}\hat{\rho}_{12}'\frac{(\hat{\sigma}_1^2\hat{\sigma}_2^2+\hat{\rho}_{12}^2)(\hat{u}_{i1}\hat{u}_{i2}-\hat{\rho}_{12})}{(\hat{\sigma}_1^2\hat{\sigma}_2^2-\hat{\rho}_{12}^2)^2}\tilde{\boldsymbol{x}}_i$$

where $\hat{\rho}_{12}'=\rho_{12}'(\hat{\alpha}_0)$. Ultimately, the first two entries of the score vector that corresponds to the variance parameters σ_1^2 and σ_2^2 become zero, and only the entries related to the parameter of interest ($\tilde{\boldsymbol{\alpha}}$) are relevant.

Fisher Information. Again, we start from the second derivative of the entire variance matrix Σ with three parameters σ_1^2, σ_2^2, and $\tilde{\boldsymbol{\alpha}}$. We use the property

$$\mathcal{I}(\Sigma)_{m,n} = \frac{1}{2}tr\left(\Sigma^{-1}\frac{\partial\ell}{\partial\Sigma}_m\Sigma^{-1}\frac{\partial\ell}{\partial\Sigma}_n\right)$$

Then, we get the following for each element of $\mathcal{I}(\Sigma)$:

$$\begin{aligned}
\mathcal{I}(\Sigma)_{\sigma_1^2\sigma_1^2} &= \frac{1}{2}\frac{\sigma_2^4}{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^2} \\
\mathcal{I}(\Sigma)_{\sigma_2^2\sigma_2^2} &= \frac{1}{2}\frac{\sigma_1^4}{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^2} \\
\mathcal{I}(\Sigma)_{\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}} &= 2\sum_{i=1}^{N}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T\frac{\rho_{12}^2\rho_{12}'^2+\sigma_1^2\sigma_2^2\rho_{12}'^2}{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^2} \\
\mathcal{I}(\theta)_{\sigma_1^2\sigma_2^2} &= -\frac{\rho_{12}\rho_{12}'\sigma_2^2}{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^2}\tilde{\boldsymbol{x}}_i \\
\mathcal{I}(\theta)_{\sigma_1^2\tilde{\boldsymbol{\alpha}}} &= -\frac{\rho_{12}\rho_{12}'\sigma_1^2}{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^2}\tilde{\boldsymbol{x}}_i
\end{aligned}$$

Finally, putting them altogether ,

$$\mathcal{I}(\sigma_1^2,\sigma_2^2,\tilde{\boldsymbol{\alpha}}) = -\frac{1}{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^2}\begin{bmatrix} \frac{1}{2}\sigma_2^4 & \frac{1}{2}\rho_{12}^2 & -\rho_{12}\rho_{12}'\sigma_2^2\tilde{\boldsymbol{x}}_i^T \\ \frac{1}{2}\rho_{12}^2 & \frac{1}{2}\sigma_1^4 & -\rho_{12}\rho_{12}'\sigma_1^2\tilde{\boldsymbol{x}}_i^T \\ -\rho_{12}\rho_{12}'\sigma_2^2\tilde{\boldsymbol{x}}_i & -\rho_{12}\rho_{12}'\sigma_1^2\tilde{\boldsymbol{x}}_i & (\rho_{12}^2\rho_{12}'^2+\sigma_1^2\sigma_2^2\rho_{12}'^2)\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T \end{bmatrix}$$

Consider the above Fisher Information matrix as a block matrix. Since the first two entries of the score vector are zeros, we only need the bottom right block of the inverted $\mathcal{I}$. After plugging in the MLEs, it turns out to be

$$\tilde{\mathcal{I}}_{\tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}}^{-1} = \frac{(\sigma_1^2\sigma_2^2-\rho_{12}^2)^4}{\rho_{12}'^2(\sigma_1^2\sigma_2^2+\rho_{12}^2)}(\boldsymbol{x}_i\boldsymbol{x}_i^T)^{-1}$$

Finally, we get our score statistic

$$q_{12} = \frac{1}{N(\hat{\sigma}_1^2\hat{\sigma}_2^2+\hat{\rho}_{12}^2)}\sum_{i=1}^{N}(\hat{u}_{i1}\hat{u}_{i2}-\hat{\rho}_{12})\tilde{\boldsymbol{x}}_i^T(\sum_{i=1}^{N}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^T)^{-1}\sum_{i=1}^{N}(\hat{u}_{i1}\hat{u}_{i2}-\hat{\rho}_{12})\tilde{\boldsymbol{x}}_i$$

Appendix B. Small Sample Correction. Although the introduced test statistic q asymptotically follows χ_1^2, the statistic has its error in the order of N^{-1} (Harris (1985)) with finite sample size N, and many Monte Carlo experiments show that the test rejects the null hypothesis less frequently than indicated by its nominal size (Godfrey (1978); Griffiths and Surekha (1986); Honda (1988)). In response, Harris (1985) used Edgeworth expansion to obtain the distribution and moment generating function to order N^{-1} of the test statistic (Harris (1985)). Building on this expansion, Honda (1986) and Cribari-Neto and Ferrari (2001) proposed corrections to the critical value or to the test statistic that allows better inference even when the sample size is small while preserving the asymptotic properties.

Honda (1988) provided a closed-form formula to adjust the critical value in the order of $O(N^{-1})$ to correct the type I error of the test. This adjustment, only depending on the covariate, sample size, and the degrees of freedom, but not on the data, is a cubic function with respect to C_γ, the critical value at the level of type I error γ, i.e. $P(\chi^2_P \geq C_\gamma) = \gamma$, and we refer to this cubic function as g defined as follows.

$$\begin{aligned} g(C_\gamma) = C_\gamma + C_\gamma\left(\frac{A_3 - A_2 + A_1}{12NP}\right) + C_\gamma^2\left(\frac{A_2 - 2A_3}{(12NP(P+2)}\right) + \\ C_\gamma^3\left(\frac{A_3}{12NP(P+2)(P+4)}\right) = C_\gamma + \tilde{g}(C_\gamma) \end{aligned} \tag{5.1}$$

where the scalars A_1, A_2, and A_3 follow the notation of Honda (1988) directly.

One of the desirable properties of g would be monotonicity, because regardless of sample size, we would like to maintain the same ordering of the strength of evidence against the null. This turns out to be almost always true in practice. The derivative of $g(C)$ is

$$g'(C_\gamma) = \frac{A_3}{12NP}\left(\frac{A_3 - A_2 + A_1 + 12NP}{A_3} + \frac{2(A_2 - 2A_3)}{(P+2)A_3}C_\gamma + \frac{3}{(P+2)(P+4)}C_\gamma^2\right)$$

A_3 is strictly positive by definition, and we can solve the above quadratic equation to see in which case the derivative is positive (Cribari-Neto and Ferrari (1995)). In other words, we can study when the following discriminant is complex.

$$\sqrt{\left(\frac{2(A_2 - 2A_3)^2}{(P+2)A_3}\right)^2 - 4 \cdot \frac{3A_3(A_3 - A_2 + A_1)}{(P+2)(P+4)A_3} - 4 \cdot \frac{3 \cdot 12NP}{(P+2)(P+4)}}$$

The first two terms inside the square root are $O(1)$ and the last term is $O(N)$, so we can see that the discriminant becomes complex quickly as N increases. Also when the covariates are simulated from normal distribution, $g'(C)$ was always positive unless $n < 3$.

Similar argument is offered in Cribari-Neto and Ferrari (1995). Based on the correction of the critical value in (5.1), Cribari (1995) proposes to subtract the correction $\tilde{g}(C_\gamma)$ so that

$$P(q \geq g(C_\gamma)) = P(q \geq C_\gamma + \tilde{g}(C_\gamma)) = P(q - \tilde{g}(C_\gamma) \geq C_\gamma).$$

This treats the correction as de-biasing instead of changing the overall shape of the distribution. Although this adjustment of the test statistic corrects the size of the test at a given threshold, it prevents further analysis when we combine the test statistics in (2.6).

Instead, we aim to adjust the test statistic so that the overall shape of null distribution is closer to χ^2_P. We assume a large enough sample size for monotonicity of g and define the inverse function of g to propose the new adjusted test statistic $\tilde{q}_{12} = g^{-1}(q)$

$$\gamma = P(\chi^2_P \geq C_\gamma) = P(q \geq g(C_\gamma)) = P(g^{-1}(q) \geq C_\gamma).$$

Our final test statistic $\tilde{q}_{12}$ is the real solution to the following equation

$$q - g(C_\gamma) = 0$$

which is guaranteed to be unique by the monotonicity of g. The cubic equation can be solved both analytically and numerically efficiently given the covariate X.

Appendix D. Data. We use the genotype data and normalized gene expression level data from GTEx V6p release (Lonsdale et al. (2013)) to apply the method to the African American samples and their gene coexpression network. The data has been pre-processed by GTEx as explained in the GTEx portal (https://gtexportal.org). In order to select African Americans from the available samples, we first inferred the local ancestry of the samples who identified themselves as European Americans or African Americans and verified that their genetic ancestry is consistent. For local ancestry inference, we use the software LAMP that reaches as high as 98% accuracy level for distinguishing YRI and CEU ancestry (Paşaniuc, Kennedy and Măndoiu (2009)). We also need the reference minor allele frequency from the pure population, so we used data from 1000 Genome Project. For the initial setting of hyperparameters in LAMP, we use 7 for the number of generations of admixture, 0.2 and 0.8 for the initial proportion of CEU and YRI population, and 10^{-8} for recombination rate, but the results are robust to these settings. LAMP returns local ancestry at each SNP as the count of African chromosomes (0, 1, or 2) at each locus, and we use the SNP closest to the center of the gene to represent the local ancestry of the entire gene. Around

92% of the genes show no recombination event in all of the subjects, and less than 3% of the genes have more than one individual with ancestry switch within the gene, so we believe this is a valid approximation.

We compute the covariate of proportion of ancestry by averaging the inferred local ancestry, and this estimate is cross-checked with principal component analysis which can effectively cluster the subjects into subpopulations (Pritchard, Stephens and Donnelly (2000)). We also include pure YRI and CEU population for PCA, and most African Americans lie strictly between the YRI and CEU population showing a two-way admixture between pure Europeans and pure Africans. We observed some outliers that were not placed between pure populations, and so we removed them. We also observed some self-identified Europeans whose genetic ancestry is more than 10% African, and we include them in our analysis as African Americans.

The expression levels provided by GTEx were measured using RNA-seq for 38,498 genes in the autosomal chromosomes. For each tissue, only genes with RPKM higher than 0.1 were included. Then the expression levels are normalized, log-transformed, and corrected for technical artifacts by GTEx Consortium.

5747 South Ellis Avenue
Chicago, IL 60637