*Submitted to the Annals of Applied Statistics*
arXiv: arXiv:0000.0000

# DYNAMIC GENE CO-EXPRESSION ANALYSIS WITH CORRELATION MODELING

By Tae Kim and Dan Nicolae

*University of Chicago*

To better understand the difference in biological processes among populations, we aim to study the effect of genetic ancestry on gene co-expression. We propose a method that can model the covariance between two variables to vary against a continuous covariate. We apply variable transformation to utilize past works on heteroskedasticity, and we offer a score test statistic that is robust to model misspecification, computationally simple, and inferentially unaffected by low sample size. Subsequently, we expand the method to test relationships between one hub-gene and many other genes to obtain a more global view of the co-expression network. Simulations show that the proposed method, compared to alternatives, has higher statistical power under more diverse scenarios with much less computational cost. We apply this method to GTEx data to analyze the dynamic behavior of the African Americans' gene co-expression against their genetic ancestry, and we identify transcription factors whose co-expressions 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 is widely studied to understand how genes are functionally connected and to often provide insights into analyzing the transcriptional regulatory system. This biological system is extremely complicated to fully and accurately comprehend when available data is limited, but we can still get vital pieces of information by focusing on a few key genes and study how they are connected to their other genes. For example, we can isolate one transcription factor gene and study how it is connected to its target genes. We define this quality of a transcription factor as its "local connectivity." This biological problem leads us to the next problem: how does local connectivity vary across various phenotypic conditions? Past works have studied how subjects in different disease statuses show distinct coexpression patterns, contributing to a better understanding of the disease

*Keywords and phrases:* co-expression; network; heteroskedasticity; score test; GTEx; admixed population;

at a molecular level [7]. Another phenotype of interest is the genetic ancestry of admixed population. Genetic ancestry is already known to play a critical role in other molecular phenotypes including DNA methylation and gene expressions [20, 8], and so we believe it has an important role in the inter-gene relationships as well. Therefore, 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 help us better comprehend how genes are differentially regulated in distinct populations.

The above problem can be translated into a statistical problem of modeling the covariance matrix of the expression levels of multiple genes. We can start from its simplest form by studying the expression levels of two genes. We build a statistical model that can explain how their correlation varies against genetic ancestry. This problem can easily be generalized into any covariance modeling problem for bivariate data outside the field of genetics. Variance modeling has been widely studied in the context of heteroskedasticity [2**? ?** ], and correlation modeling under discrete conditions has been studied in the context of differential network [**?** ], but dynamic correlation modeling has been less explored.

Li (2002) addresses the most similar scientific problem to ours [15**?** ]. The paper uses the term "liquid association" (LA) to conceptualize the internal evolution of the coexpression pattern for a pair of genes. He analyzes the coexpression that changes across the different cellular states that cannot be directly observed by using the expression level of another gene to represent the cellular state. Other studies built on the liquid association to better identify cell states that affect coexpression [**? ? ?** ], 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 test local ancestry. Second, it treats the covariate as a random variable that follows a normal distribution, which genetic ancestry does not. Third, it only tests the linear relationship between the covariate and the coexpression. Lastly, the proposed test statistic does not have a closed-form null distribution and requires a permutation test, leading to computational inefficiency.

We propose a different methodology to solve the continuously-varying

covariance problem. We first assume that the two genes' expression levels follow a bivariate normal distribution. Then, we apply a simple variable transformation to induce independence, effectively changing the multivariate covariance modeling problem to a univariate variance modeling problem. Next, we apply a traditional score test for heteroskedasticity [2] 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. When the number of genes 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 tests 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. Next 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 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. *Framework.* We assume the data $\boldsymbol{y_i} \in \mathbb{R}^K$, $i = 1, 2, \cdots, N$ are independent, and we define the data matrix $Y = \{y_{ik}\}_{i=1,k=1}^{N,K} \in \mathbb{R}^{N\times K}$. The covariate matrix $X \in \mathbb{R}^{N\times P} = \{x_{ip}\}_{i=1,p=1}^{N,P}$ is assumed to be full-rank with $P < N$. We assume $\boldsymbol{y_i}$ independently follow $K$-variate normal distribution as follows,

$$\boldsymbol{y_i} = \boldsymbol{b}_0 + \boldsymbol{x}_i^T B + \boldsymbol{u_i}, \boldsymbol{u_i} \sim \mathcal{N}_K(\boldsymbol{0}, \Sigma(\boldsymbol{x}_i^T \boldsymbol{\alpha})) \tag{2.1}$$

$$\Sigma(\boldsymbol{x}_i) = \{\rho_{kl}(\boldsymbol{x}_i^T \boldsymbol{\alpha})\}_{k,l=1}^K$$

where $\rho_{kl}$ is constant when $k = l$ while varies across $\boldsymbol{x}_i$ only when $k \neq l$.

The above model is close to a multivariate regression model of $\boldsymbol{y_i}$ with respect to the covariate $\boldsymbol{x}_i$ with intercept $\boldsymbol{b}_0$, slope $B$, and error term $\boldsymbol{u_i}$. However, it differs from the classical regression model because its error variance $\Sigma(\boldsymbol{x}_i, \boldsymbol{\alpha})$ depends on the covariate $\boldsymbol{x}_i$ through some parameters $\boldsymbol{\alpha}_{k\ell}$. The function $\rho_{k\ell}$ represents the specific form of heteroskedasticity between the two variables $k$ and $\ell$, and possible choices for this function are further discussed in section 2.2.

In the context of gene-coexpression network of African Americans, the data matrix $Y$ is the gene expression level matrix for $N$ individuals at $K$ genes, and therefore, we expect this to be a high dimensional problem with $K > N$. The covariate $\boldsymbol{x}_i$ holds genetic ancestry information of individual $i$. 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 section 4, we focus on scalar $\boldsymbol{x}_i$ for straightforward interpretability.

We aim to model $\Sigma$ to see how the relationship among $K$ variables changes across the different $\boldsymbol{x}_i$. and therefore we assume that the diagonal entries of $\Sigma$ are fixed and do not depend on $\boldsymbol{x}_i$. This may or may not be appropriate depending on the contexts, but it is empirically justified for our example data set of GTEx gene expression level matrix and genetic matrix.

The co-expression network is often considered sparse, meaning most of the non-diagonal entries of $\Sigma$ are zeros. However, when we dynamically model the network to vary across the continuous genetic ancestry, the sparsity assumption becomes highly restrictive. Such assumption would require the co-expression between two genes is 0 across all possible genetic ancestry for most genes. In contrast, we believe it is plausible that genes are independent in some ancestry while highly correlated in some other ancestry. More generally, we lack prior information how the co-expression matrix would change across ancestry, and we are open to find any possible patterns. Therefore, we do not assume any specific structure on $\Sigma$. It is difficult to model a continuously varying $K \times K$ covariance matrix within the space of symmetric, positive definite space. Therefore, we first focus on a pair of two variables 1 and 2 and model their correlation $\rho(\boldsymbol{x}_i)$ in section 2.2. Then in section 2.4, we combine information across multiple pairs of genes to make inference on gene's connectivity across the entire network, discussed in detail.

2.2. *Inference for $K = 2$.* Here, we focus on a simple case of $K = 2$. For the next two sections, we assume that the variance matrix has 1 on all the diagonals for mathematical simplicity. This can be a proper choice for certain applications. For example, the gene expression level has already been quantile normalized as a pre-processing step. In section 2.5, we lift the unit variance assumption and present the method for the general case.

We can re-write (2.1) for $K = 2$ as below,

$$\begin{aligned} \begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} &= \begin{bmatrix} b_{01} \\ b_{02} \end{bmatrix} + \begin{bmatrix} \boldsymbol{x_i^T \beta_1} \\ \boldsymbol{x_i^T \beta_2} \end{bmatrix} + \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \\ \begin{bmatrix} u_{i2} \\ u_{i2} \end{bmatrix} &\sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{12}(\boldsymbol{x}_i, \boldsymbol{\alpha_{12}}) \\ \rho_{12}(\boldsymbol{x}_i, \boldsymbol{\alpha_{12}}) & 1 \end{bmatrix} \right) \end{aligned} \tag{2.2}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the first and second columns of $B$ respectively. The null hypothesis is

$$\rho_{12}(\boldsymbol{x}_i, \boldsymbol{\alpha_{12}}) = \bar{\rho}_{12}. \tag{2.3}$$

In other words, the covariance between variables 1 and 2 do not depend on $\boldsymbol{x}_i$ and is a constant value $\bar{\rho}_{12}$. Both the function $\rho$ and the parameters $\boldsymbol{\alpha}$ depend on the variable pair 1 and 2, and for the rest of the section, we drop the subscripts 12 for notational convenience.

The likelihood of (2.2) is

$$\begin{aligned} \ell(\boldsymbol{\alpha}, \boldsymbol{b}_0, \boldsymbol{\beta}) = \sum_{i=1}^{N} \Bigg( &- log(2\pi) - \frac{\sqrt{1 - \rho^2(\boldsymbol{x}_i, \boldsymbol{\alpha})}}{2} \\ &- \frac{(y_{i1} - \mu_{y_1})^2 + (y_{i2} - \mu_{y_2})^2 - 2\rho(\boldsymbol{x}_i, \boldsymbol{\alpha})(y_{i1} - \mu_{y_1})(y_{i2} - \mu_{y_2})}{2(1 - \rho^2(\boldsymbol{x}_i, \boldsymbol{\alpha}))} \Bigg), \end{aligned} \tag{2.4}$$

where $\mu_{y_1} = b_{01} + \boldsymbol{x}_i^T \boldsymbol{\beta}_1$, $\mu_{y_2} = b_{02} + \boldsymbol{x}_i^T \boldsymbol{\beta}_2$ and $\boldsymbol{b}_0 = [b_{01} \;\; b_{02}]^T$. Since the the model is bivariate, the full likelihood seems tractable, but we propose below a variable transformation to make the likelihood much simpler without losing any information.

$$\begin{aligned} \begin{bmatrix} w_i \\ v_i \end{bmatrix} = \begin{bmatrix} y_{i1} + y_{i2} \\ y_{i1} - y_{i2} \end{bmatrix} &= \begin{bmatrix} b_{01} + b_{02} \\ b_{01} - b_{02} \end{bmatrix} + \boldsymbol{x}_i^T \begin{bmatrix} \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} u_w^2 \\ u_v^2 \end{bmatrix} \\ \begin{bmatrix} u_w^2 \\ u_v^2 \end{bmatrix} &\sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 + 2\rho(\boldsymbol{x}_i, \boldsymbol{\alpha}) & 0 \\ 0 & 2 - 2\rho(\boldsymbol{x}_i, \boldsymbol{\alpha}) \end{bmatrix} \right) \end{aligned} \tag{2.5}$$

The two variables $w_i$ and $v_i$ are independent and the likelihood can be written as a sum of two univariate likelihoods.

(2.6)
$$\ell(\boldsymbol{\alpha}, \boldsymbol{b_0}, \boldsymbol{\beta}) = -\frac{N}{2}log(2\pi) - \frac{1}{2}\sum_i log(2 + 2\rho(\boldsymbol{x}_i, \boldsymbol{\alpha})) - \sum_{i=1}^{N} \frac{(w_i - \mu_w)^2}{2 + 2\rho(\boldsymbol{x}_i, \boldsymbol{\alpha})}$$
$$- \frac{N}{2}log(2\pi) - \frac{1}{2}\sum_i log(2 - 2\rho(\boldsymbol{x}_i, \boldsymbol{\alpha})) - \sum_{i=1}^{N} \frac{(v_i - \mu_v)^2}{2 - 2\rho(\boldsymbol{x}_i, \boldsymbol{\alpha})}$$

where $\mu_w$ and $\mu_v$ are $b_{01}+b_{02}+\boldsymbol{x}_i^T(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)$ and $b_{01}-b_{02}+\boldsymbol{x}_i^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)$ respectively. This transformation effectively changes the problem from modeling covariance to modeling variance, allowing us to apply results from literature on univariate heteroskedasticity.

Given the likelihood in (2.6), we have two well-known tools to test the null hypothesis (2.3) — likelihood ratio test and Rao's score test [1]. Both methods require us to impose certain assumptions on $\rho$, but Rao's score test allows much more flexibility.

First, the likelihood ratio test requires the full specification of the function $\rho$ to estimate the maximum likelihood estimate (MLE) of $\boldsymbol{\alpha}$ both under the null hypothesis and under the alternative hypothesis. One straightforward function for $\rho$ is any kind of sigmoid function bounded to (-1, 1) such as logistic function, hyperbolic tangent function, or any cumulative distribution supported on the whole real line. For the input of $\rho$, we can use simple linear form of $\boldsymbol{x}_i^T\boldsymbol{\alpha}$, or we can also allow non-linearity by using higher order polynomial or even generalized additive models. There are two problems with the likelihood ratio test. One, as mentioned in the previous section, we would like to impose as little assumption on the specific form of heteroskedasticity as possible. If $\rho$ is highly mis-specified, we sacrifice a lot of statistical power. Two, most of the reasonable assumptions of $\rho$, 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

[1] In fact, due to the nuisance parameters, the appropriate terminology is Rao's efficient score test or equivalently Neyman's $C(\alpha)$ test [14, 18]. Breusch and Pagan (1979) refers to the same procedure as Lagrange Multiplier Test [2]. Bera and Bilias (2001) discusses the historical development of three methods - Rao's score test, Neyman's $C(\alpha)$ test, and Lagrange Multiplier test, and here we choose the term "score test" as it is most familiar to the statisticians.[1]

as in our application of gene co-expression network.

Meanwhile, Rao's score test, unlike the likelihood ratio test, only requires the MLE of $\boldsymbol{\alpha}$ under the null hypothesis [23]. Moreover, under a mild assumption of linearity and additivity ($\rho(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \rho(\boldsymbol{x}_i^T \boldsymbol{\alpha})$), the test statistic does not depend on the form of $\rho$ while maintaining its asymptotic property as long as $\rho$ is twice differentiable. Therefore, we define $\boldsymbol{\alpha} \in \mathbb{R}^p$ as the linear coefficients and finalize our model below

$$\rho(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \rho(\boldsymbol{x}_i^T \boldsymbol{\alpha}) \tag{2.7}$$

with the null hypothesis

$$\boldsymbol{\alpha} = \mathbf{0}. \tag{2.8}$$

In some cases, the underlying heteroskedasticity could be more complex than what is formulated in (2.7), for example when there are interactions among the covariates, but we believe the linearity and additivity assumptions are standard for the incipient stage of the analysis. Also, since $\rho$ can take any non-linear form, (2.7) still is a flexible framework especially compared to the likelihood ratio test. In order to test (2.8), we expand the result from Breusch and Pagan (1979) to derive the test statistic [2]. The performance of score test and the likelihood ratio test are further analyzed in the simulation section 3

The score test allows us to replace all the nuisance parameters with the MLEs under the null hypothesis. We therefore replace $\boldsymbol{b}_0$ and $\boldsymbol{\beta}$ with the OLS estimators $(\hat{\boldsymbol{b}}_{\mathbf{0},\text{MLE}}, \hat{\boldsymbol{\beta}}_{\text{MLE}})$ of the intercept and the slope, and define $\hat{u}_w^2$ and $\hat{u}_v^2$ as the OLS residuals in the model (2.5). We also define $\hat{\sigma}_w^2 = \sum_{i=1}^N \hat{u}_w^2/N$ and $\hat{\sigma}_v^2 = \sum_{i=1}^N \hat{u}_v^2/N$ as the MLEs for the error variance under the null hypothesis. The first derivative of the likelihood evaluated at $\boldsymbol{\alpha} = \mathbf{0}, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\text{MLE}}$, and $\boldsymbol{b}_0 = \hat{\boldsymbol{b}}_{\mathbf{0},\text{MLE}}$ is

$$\begin{aligned} \boldsymbol{d}_{\boldsymbol{\alpha}} = \frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{b_0}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} \Big|_{\boldsymbol{\alpha}=\mathbf{0}, \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\text{MLE}}, \boldsymbol{b}_0=\hat{\boldsymbol{b}}_{\mathbf{0},\text{MLE}}} = \\ -\rho'(0) \sum_{i=1}^N \left( \frac{\boldsymbol{x}_i}{\hat{\sigma}_w^2} \left(1 - \frac{\hat{u}_{wi}^2}{\hat{\sigma}_w^2}\right) - \frac{\boldsymbol{x}_i}{\hat{\sigma}_v^2} \left(1 - \frac{\hat{u}_{vi}^2}{\hat{\sigma}_v^2}\right) \right). \end{aligned} \tag{2.9}$$

The second derivative, also evaluated at $\boldsymbol{\alpha} = \mathbf{0}, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\text{MLE}}$, and $\boldsymbol{b}_0 = \hat{\boldsymbol{b}}_{\mathbf{0},\text{MLE}}$

is

$$
\begin{aligned}
I_{\boldsymbol{\alpha}\boldsymbol{\alpha}^{\boldsymbol{T}}} &= \frac{\partial^2 \ell(\boldsymbol{\alpha}, \boldsymbol{b_0}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}\boldsymbol{\alpha}^{\boldsymbol{T}}} \mid_{\boldsymbol{\alpha}=\mathbf{0}, \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}, \boldsymbol{b}_0} \\
&= \hat{\boldsymbol{b}}_{\mathbf{0},\mathrm{MLE}} = 2\rho'(0)^2 \left( \frac{1}{\hat{\sigma}_w^4} + \frac{1}{\hat{\sigma}_v^4} \right) \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T .
\end{aligned} \tag{2.10}
$$

The test statistic $q$ for the variable pair (1,2) is, as defined in Breusch and Pagan (1979),

$$
\begin{aligned}
q &= \boldsymbol{d}_{\boldsymbol{\alpha}}^T I_{\boldsymbol{\alpha}\boldsymbol{\alpha}^T}^{-1} \boldsymbol{d}_{\boldsymbol{\alpha}} \\
&= \frac{1}{2\left(\frac{1}{\hat{\sigma}_w^4} + \frac{1}{\hat{\sigma}_v^4}\right)} \left( \left( \sum_{i=1}^{N} \boldsymbol{x_i} \left( \frac{1}{\hat{\sigma}_w^2} - \frac{\hat{u}_{wi}^2}{\hat{\sigma}_w^4} \right) - \left( \frac{1}{\hat{\sigma}_v^2} - \frac{\hat{u}_{vi}^2}{\hat{\sigma}_v^4} \right) \right)^T \right. \\
&\quad \left. \left( \sum_{i=1}^{N} \boldsymbol{x_i^T x_i} \right)^{-1} \left( \sum_{i=1}^{N} \boldsymbol{x_i} \left( \frac{1}{\hat{\sigma}_w^2} - \frac{\hat{u}_{wi}^2}{\hat{\sigma}_w^4} \right) - \left( \frac{1}{\hat{\sigma}_v^2} - \frac{\hat{u}_{vi}^2}{\hat{\sigma}_v^4} \right) \right) \right)
\end{aligned} \tag{2.11}
$$

where the unknown function $\rho$ has been canceled out. Every component of the test statistic is easily acquired from the data, and the computational burden is low. Most importantly, it is flexible as it allows any form of heteroskedasticity $\rho$. Under this setting, Breusch and Pagan (1979) proves that $q$ asymptotically follows $\chi_P^2$ [2].

$$
q \to \chi_P^2 \tag{2.12}
$$

However, even though the introduced test statistic has convenient asymptotic properties, the sample size is not large enough in many applications. The statistic has its error in the order of $N^{-1}$ [11], and many Monte Carlo experiments show that the test rejects the null hypothesis less frequently than indicated by its nominal size [9, 10, 12] . In response, corrections have been suggested [11, 6, 6, 12], and we apply Honda's method to ensure the asymptotic properties even under the samll sample size. The details of the small-sample correction as well as the derivation of the test statistic, are shared in Appendix.

2.3. *Connection to Liquid Association.* Li (2002) addresses a similar scientific problem to ours when $K = 2$ [15]. The paper uses the term "liquid association" (LA) to conceptualize the internal evolution of co-expression pattern for a pair of genes [15]. The paper analyzes the co-expression pattern that changes across different cellular state that cannot be directly observed by using the expression level of another gene to represent the cellular state.

This expression level can be considered analogous to the sample-specific covariate $X$ in our model, and then LA becomes similar to our test statistic: quantity to measure the change in $cor(Y_1, Y_2)$ across a continuous variable $X$.

It defines the function $g$ to denote the mean of correlation between two genes $Y_1$ and $Y_2$ conditional on $X$, $g(X) = E(Y_1Y_2|X)$, and assumes that $X$ follows standard normal variable, he uses Stein's lemma to show $Eg'(X) = Eg(X)X = E(Y_1Y_2X)$ to finally use $\sum_{i=1}^{N} y_{i1}y_{i2}x_i$ as the test statistic. We can write the test statistic in the context of our model with function $\rho$, if we assume $X$ to be a 1-dimensional random vector.

$$\begin{aligned} Y_1 &= \beta_1 X + \epsilon_1, \quad \epsilon_1 \sim N(0,1) \\ Y_2 &= \beta_2 X + \epsilon_2, \quad \epsilon_2 \sim N(0,1) \\ E(\epsilon_1\epsilon_2|X) &= \rho(X\alpha) \end{aligned}$$

Then, using its derivation process, we can write $g(X)$ and $Eg'(X)$ as follows.

$$\begin{aligned} g(X) &= E(Y_1Y_2|X = x) \\ &= E(\beta_1\beta_2X^2 + \beta_1\epsilon_2X + \beta_2\epsilon_1X + \epsilon_1\epsilon_2|X) \\ &= \beta_1\beta_2X^2 + \rho(X\alpha) \\ Eg'(X) &= Eg(X)X \\ &= \beta_1\beta_2E(X^3) + E\left(X \cdot \rho(X\alpha)\right) \\ &= E(X \cdot \rho(X\alpha)) \end{aligned} \tag{2.13}$$

Under our null hypothesis $\alpha = 0$, $\rho(X\alpha) = \rho(0)$ is constant and independent of $X$, so we eventually get $Eg'(X) = E(X) = 0$.

However, the liquid association model is different from ours in a few critical ways. First, it restricts the covariate to be a 1-dimensional vector while our method can be expanded to multidimensional covariate $X$. Second, it treats $X$ as a random variable that follows standard normal distribution, while we treat it as fixed, and our analysis does not depend on any distributional assumption. Therefore, our method can be generalized and applied to a wide range of problems. Third, the liquid association only measures the strength of linear correlation between the co-expression and the cellular state, while our method can detect any type of cases that deviate from null hypothesis with various shapes of $\rho$. Lastly, the LA test statistic does not have a closed-form null distribution and requires permutation test for inference, leading to computational inefficiency. We compare the performance in detail in section 3.

2.4. *Inference for $K > 2$.* In section 2.2, we proposed the test statistic $q$ that tests a pair of variables 1 and 2 to measure the evidence that their correlation changes with respect to the covariate $X$. Then we made a small sample correction to obtain $\tilde{q}_{12}$ that closely follows $\chi^2_P$ distribution even when sample size is small.

As a natural extension to the pair-wise test statistic, we can repeat the procedure for all variable pairs $k$ and $\ell$ to obtain $\tilde{q}_{k\ell}$. In this section, we propose a way to combine the test statistics to test a new global null hypothesis with improved statistical power. The global null hypothesis for variable 1 extends (2.8) as follows,

$$\boldsymbol{H}_0^{(1)} : \boldsymbol{\alpha}_{12} = \boldsymbol{\alpha}_{13} = \cdots = \boldsymbol{\alpha}_{1K} = \mathbf{0}, \tag{2.14}$$

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 $X$.

Combining the test statistics can have either positive or negative impact; the procedure can accrue relevant evidence to improve the statistical power, or it can accumulate noise to do the exact opposite. Therefore, we must carefully decide how to combine the test statistics based on the alternative hypothesis we would like to leverage against, and the alternative hypothesis must be constructed to reflect our prior knowledge about the network structure. Chen (2012) discusses two ways to construct the alternative hypothesis [4]. 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. Chen (2012) focuses on the sparse alternative and proposes the exponential-combination framework. Here, we do not assume that our alternative is sparse and propose a simpler linear combination of the test statistics

$$d_1 = \tilde{q}_{12} + \tilde{q}_{13} + \cdots + \tilde{q}_{1K} = \sum_{k=2}^{K} \tilde{q}_{1k}. \tag{2.15}$$

We believe combining the test statistics like (2.15) improves the statistical power of tests for any network whose structure is similar to scale-free topology [13], i.e. where the "hot spot" variables or "hub" variables are connected to a lot of other nodes forming cliques or modules. In the context of gene co-expression 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 those genes regulated by that transcription factor are likely to be correlated with the covariate as well. The effect sizes for each gene pair may be too small to be detected, but combining them by simple addition like in (2.15) can form a stronger signal.

In order to test the significance of $d_1$ against the null hypothesis (2.14), we need the null distribution of $d_1$. Although $\tilde{q}_{1k}$ separately follow $\chi^2_P$, they are correlated to one another, so their null distribution is not trivial. Incidentally, 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 inference. 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 (2.11) can be alternatively written as follows, where we define $r_p$.

(2.16)
$$q = \sum_{p=1}^{P}\left(\frac{1}{\sqrt{N}}\sqrt{\frac{\hat{\sigma}_w^4\hat{\sigma}_v^4}{\hat{\sigma}_w^4+\hat{\sigma}_v^4}}\sum_{i=1}^{N} x_{ip}\left(\frac{\hat{\sigma}_w^2-\hat{u}_{wi}^2}{\hat{\sigma}_w^4}-\frac{\hat{\sigma}_v^2-\hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right)\right)^2 = \sum_{p=1}^{P} r_p^2$$

For each $p$, $r_p$ follows the standard normal distribution by the central limit theorem.

$$E\left(x_{ip}\left(\frac{\hat{\sigma}_w^2-\hat{u}_{wi}^2}{\hat{\sigma}_w^4}-\frac{\hat{\sigma}_v^2-\hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right)\right) = x_{ip}E\left(\frac{\hat{\sigma}_w^2-\hat{u}_{wi}^2}{\hat{\sigma}_w^4}\right)E\left(\frac{\hat{\sigma}_v^2-\hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right) = 0$$

$$Var\left(x_{ip}\left(\frac{\hat{\sigma}_w^2-\hat{u}_{wi}^2}{\hat{\sigma}_w^4}-\frac{\hat{\sigma}_v^2-\hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right)\right) = x_{ip}^2\left(Var\left(\frac{\hat{\sigma}_w^2-\hat{u}_{wi}^2}{\hat{\sigma}_w^4}\right)+Var\left(\frac{\hat{\sigma}_v^2-\hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right)\right)$$
$$= \frac{\hat{\sigma}_w^4+\hat{\sigma}_v^4}{\hat{\sigma}_w^4\hat{\sigma}_v^4}$$

This confirms the previous result

$$\sum_{p=1}^{P} r_p^2 \to \chi^2_P$$

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 $r$

$$\boldsymbol{r}_{1,p} = \begin{bmatrix} r_{12,p} \\ r_{13,p} \\ \cdots \\ r_{1K,p} \end{bmatrix} \to N_{K-1}(\boldsymbol{0}, H_1), \qquad \forall p = 1, \cdots, P \tag{2.17}$$

$H_1$ 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 (2.16), it is easy to see that $H_1$ has 1 at the diagonals. Also, $\eta_{1k,1\ell}$ converges in probability to

(2.18)
$$\frac{\rho_{1k}^3\rho_{1\ell}^3+\rho_{1k}^3\rho_{1\ell}+\rho_{1k}\rho_{1\ell}^3-3\rho_{1k}^2\rho_{1\ell}^2\rho_{k\ell}+2\rho_{1k}\rho_{1\ell}\rho_{k\ell}^2-\rho_{1k}^2\rho_{k\ell}-\rho_{1\ell}^2\rho_{k\ell}-\rho_{1k}\rho_{1\ell}+\rho_{k\ell}}{(1-\rho_{1k})^2(1-\rho_{1\ell})^2\sqrt{(1+\rho_{1k}^2)(1+\rho_{1\ell}^2)}}.$$

where the derivation is in Appendix.

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

$$d_1=\sum_{p=1}^{P}\|\boldsymbol{r}_{1,p}\|_2^2=\sum_{p=1}^{P}\sum_{k=2}^{K}r_{1k,p}^2, \tag{2.19}$$

and we can derive the distribution of $d_1$ as well. Let $H_1=U_1\Lambda_1U_1^T$ be the eigen-decomposition of the covariance matrix $H_1$ in (2.17), where the diagonal matrix $\Lambda$ has eigenvalues $\lambda_{12},\cdots,\lambda_{1K}$ in a decreasing order. Then, we can next consider the transformation $\boldsymbol{r}_{\boldsymbol{1,p}}^{*}=U\boldsymbol{r}_{1,p}$ that follows normal distribution with diagonal covariance matrix $N_{K-1}(\boldsymbol{0},\Lambda_1)$. Note that $\|\boldsymbol{r}_{1,p}\|_2^2=\|U\boldsymbol{r}_{1,p}\|_2^2$ due to the orthogonal invariance of L2 norm. Then,

$$\begin{aligned} d_1=&\sum_{p=1}^{P}{r_{12,p}^{*}}^2+\cdots+\sum_{p=1}^{P}{r_{1K,p}^{*}}^2\\ &\sum_{p=1}^{P}{r_{1k,p}^{*}}^2\sim\Gamma\left(\frac{P}{2},\frac{\lambda_{1k}}{2}\right),\quad k=2,\cdots,K \end{aligned} \tag{2.20}$$

Assuming that we know the true, symmetric, positive definite $H$, we can acquire positive $\lambda_{1k}$ 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 [17].

In (2.18), we define the element-wise mapping $\phi:\Sigma\to H$. It is clear from the construction of $H$ that if we can estimate a well-conditioned, symmetric, positive definite correlation matrix $\hat{\Sigma}$, $\phi(\hat{\Sigma})$ is also symmetric and positive

definite. As $N \to \infty$, $\hat{\Sigma}$ converges to $\Sigma$, so we can easily acquire the asymptotic null distribution of $d_1$. When $N$ is sufficiently larger than $K$, empirical covariance matrix has nice asymptotic properties to guarantee that the test statistics in (2.15) converges in distribution to (2.20).

However, when $K$ is much larger than $n$, an accurate estimation of $\Sigma$ is a difficult problem, especially when there is no structural assumption such as sparsity or low rank. So we instead turn to the permutation test, which is valid under the independence assumption in (2.1). Empirically, we justify the exchangeability of GTEx subjects through some exploratory analysis which shows that the covariance matrix of the gene expression levels has small non-diagonal elements. Also the principal components didn't show any clustering. We conclude that permutation test is well justified, and so we shuffle the covariate vector and test against the network data to preserve the correlation structure of the network.

We use the sequential precision-improvement permutation test, similar to one suggested by Chen (2012) [4]. Permutation test often results in a poor 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 many genes can be computationally wasteful. In order to find balance, as we proceed with incrementally larger number of permutations, we count the number of cases that led to more extreme degree statistics than observed $d_k$. After the minimum number of permutations predefined by the user (1000 in our analysis), if more 2 more extreme cases were found, the permutation was terminated. If there are less than 2 such cases observed, we add 100 more permutations and re-check for early termination. We repeat until it reaches the predefined maximum number of permutation, which is designed to give a good enough resolution of $p$-value, given the number of tests that we are performing.

2.5. *Generalization to Non-unit variance.* So far, we have assumed that all $K$ variables have variance 1 so that the variance matrix $\Sigma$ has unit diagonals. However in most cases, different variables have different variability. For example, some functional genes have expression level more highly variable than other house-keeping genes. In that case, we generalize (**??**) to

$$\boldsymbol{y_i} = \boldsymbol{b}_0 + \boldsymbol{x}_i^T B + \boldsymbol{u_i}, \boldsymbol{u_i} \sim \mathcal{N}_K(\boldsymbol{0}, \Sigma(\boldsymbol{x}_i^T \boldsymbol{\alpha})) \tag{2.21}$$

$$\Sigma(\boldsymbol{x}_i) = \{\rho_{k_1 k_2}(\boldsymbol{x}_i^T \boldsymbol{\alpha})\}_{k_1 k_2 = 1}^K$$

where $\rho_{k_1k_2}$ is constant when $k_1 = k_2$ while varies across $\boldsymbol{x}_i$ only when $k_1 \neq k_2$. Then we can make the variable transformation as follows to create two independent univariate distributions.

$$
\begin{aligned}
&\begin{bmatrix} w_{i,12} \\ v_{i,12} \end{bmatrix} = \begin{bmatrix} \frac{y_{i1}}{\sqrt{\rho_{11}}} + \frac{y_{i2}}{\sqrt{\rho_{22}}} \\ \frac{y_{i1}}{\sqrt{\rho_{11}}} - \frac{y_{i2}}{\sqrt{\rho_{22}}} \end{bmatrix} = \boldsymbol{x}_i^T \begin{bmatrix} \frac{\boldsymbol{\beta}_1}{\sqrt{\rho_{11}}} + \frac{\boldsymbol{\beta}_2}{\sqrt{\rho_{11}}} \\ \frac{\boldsymbol{\beta}_1}{\sqrt{\rho_{11}}} - \frac{\boldsymbol{\beta}_2}{\sqrt{\rho_{22}}} \end{bmatrix} + \begin{bmatrix} u_{wi} \\ u_{vi} \end{bmatrix} \\
&\begin{bmatrix} u_{wi} \\ u_{vi} \end{bmatrix} \sim \mathcal{N}\left( \mathbf{0}, \begin{bmatrix} 2 + \frac{2\rho_{12}(\boldsymbol{x}_i^T\boldsymbol{\alpha})}{\sqrt{\rho_{11}\rho_{22}}} & 0 \\ 0 & 2 - \frac{2\rho_{12}(\boldsymbol{x}_i^T\boldsymbol{\alpha})}{\sqrt{\rho_{11}\rho_{22}}} \end{bmatrix} \right)
\end{aligned} \tag{2.22}
$$

The score test allows us to replace all the nuisance parameters with their MLEs, so we can replace $\rho_{11}$ with $\hat{\rho}_{11,\text{MLE}} = \frac{\sum_{i=1}^N {y_{i1}}^2}{N}$ and $\rho_{22}$ with $\hat{\rho}_{22,\text{MLE}} = \frac{\sum_{i=1}^N {y_{i2}}^2}{N}$. We re-define $\hat{u}_{wi}$ and $\hat{u}_{vi}$ as follows.

$$
\hat{u}_{wi} = \frac{\hat{u}_{1i}}{\sqrt{\hat{\rho}_{11,\text{MLE}}}} + \frac{\hat{u}_{2i}}{\sqrt{\hat{\rho}_{22,\text{MLE}}}}, \qquad \hat{u}_{vi} = \frac{\hat{u}_{1i}}{\sqrt{\hat{\rho}_{11,\text{MLE}}}} - \frac{\hat{u}_{2i}}{\sqrt{\hat{\rho}_{22,\text{MLE}}}} \tag{2.23}
$$

Defining $\hat{\sigma}_w^2$ and $\hat{\sigma}_v^2$ as $\sum_{i=1}^N \frac{u_{wi}^2}{N}$ and $\frac{u_{vi}^2}{N}$, respectively, we can re-write the test statistic $q$ in the same way.

$$
\begin{aligned}
\boldsymbol{d}_{\boldsymbol{\alpha}} &= \frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{b_0}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} \Big|_{\boldsymbol{\alpha}=\mathbf{0}, \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\text{MLE}}, \boldsymbol{b}_0=\hat{\boldsymbol{b}}_{\mathbf{0},\text{MLE}}, \rho_{11}=\hat{\rho}_{11,\text{MLE}}, \rho_{22}=\hat{\rho}_{22,\text{MLE}}} \\
&= -\rho_{12}'(0) \sum_{i=1}^N \left( \frac{\boldsymbol{x}_i}{\hat{\sigma}_w^2} \left( 1 - \frac{\hat{u}_{wi}^2}{\hat{\sigma}_w^2} \right) - \frac{\boldsymbol{x}_i}{\hat{\sigma}_v^2} \left( 1 - \frac{\hat{u}_{vi}^2}{\hat{\sigma}_v^2} \right) \right)
\end{aligned} \tag{2.24}
$$

The second derivative is similarly

$$
\begin{aligned}
I_{\boldsymbol{\alpha}\boldsymbol{\alpha}^T} &= \frac{\partial^2 \ell(\boldsymbol{\alpha}, \boldsymbol{b_0}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha} \boldsymbol{\alpha}^T} \Big|_{\boldsymbol{\alpha}=\mathbf{0}, \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\text{MLE}}, \boldsymbol{b}_0=\hat{\boldsymbol{b}}_{\mathbf{0},\text{MLE}}, \rho_{11}=\hat{\rho}_{11,\text{MLE}}, \rho_{22}=\hat{\rho}_{22,\text{MLE}}} \\
&= 2\rho_{12}'(0)^2 \left( \frac{1}{\hat{\sigma}_w^4} + \frac{1}{\hat{\sigma}_v^4} \right) \sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{x}_i^T.
\end{aligned} \tag{2.25}
$$

Then the test statistic is the same as (2.11) with the same asymptotic distribution.

$$
\begin{aligned}
q &= \boldsymbol{d}_{\boldsymbol{\alpha}}^T I_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1} \boldsymbol{d}_{\boldsymbol{\alpha}} \\
q &\to \chi_1^2
\end{aligned}
$$

Inference for $K > 2$ also remains the same.

In practice, we can scale each column of the data matrix $Y$ by its maximum likelihood estimate of variance as the first step of the analysis.

$$\tilde{Y} = \left[\tilde{\boldsymbol{y}}_{\cdot 1}, \cdots, \tilde{\boldsymbol{y}}_{\cdot K}\right]$$
$$\tilde{\boldsymbol{y}}_{\cdot k} = \frac{\boldsymbol{y}_{\cdot k}}{\sqrt{\sum_{i=1}^{N} y_{ik}^2/N}}$$

Replacing the data matrix $Y$ with its scaled version $\tilde{Y}$, all the testing and inference procedures introduced in the previous sections remain the same.

**3. Simulation Studies.** In this section, 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 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

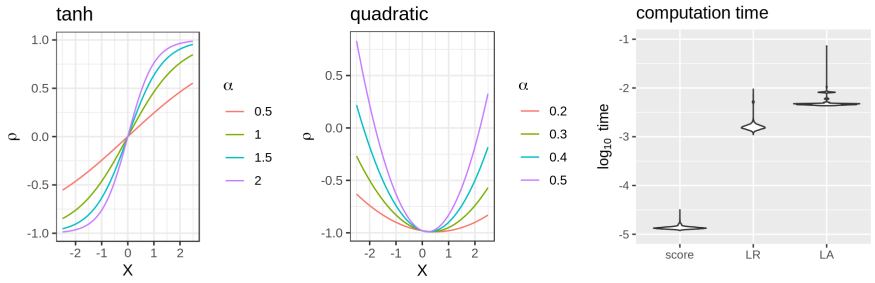
$$y_i \sim \mathcal{N}_2\left(\boldsymbol{b}_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}, \begin{bmatrix} 1 & \bar{\rho} \\ \bar{\rho} & 1 \end{bmatrix}\right)$$

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 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 $\rho$,

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

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 $\rho$ asymptotically follows normal distribution, so it works well when $X$ is drawn from normal distribution. We use *optim* function in R to 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 $\chi_1^2$




(a) (b) (c)

Fig 1: (a)-(b) The two heteroskedastic functions $\rho$ used to generate data. Simulation results in Table (1) show that score test outperforms LA in both cases, and that LA particularly suffers in capturing the non-linear relationship in quadratic model. Likelihood ratio test performs better than score test when the model is correctly specified, but when the model is mis-specified, score test performs better. (c) The computational burden is much lower in the proposed method compared to the other two. LR test requires numerical optimization for finding MLEs, and LA requires permutation test.

closely. So we focus on the small sample case $N = 30$ as that is close to the size of the real data of African Americans' gene expression level.

Next we generate the data under the alternative hypothesis to compare the statistical power. For fixed $N = 30$, 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 $\rho$, we draw $Y$ from (**??**) with varying levels of $\alpha$, 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 $\rho$ as follows. Subtracting 0.99 is to ensure numerical stability.

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

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 $\rho$ with respect to $X$ with varying levels of $\alpha$.

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 $\rho$ is generated from hyperbolic tangent function, likelihood ratio test generally outperforms the other two methods, as expected since the model is correctly specified in LR test. Score test, although does not assume any model on $\rho$, does not lose as much power as liquid association does. Meanwhile, when $\rho$ is generated from quadratic function, score function clearly outperforms the other two methods. The proposed score test is robust to the shape of heteroskedasticity. Figure 1 (c) shows the distribution of computation times of each method 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.

| $\rho$ | tanh | | | | | quadratic | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 0 | 0.5 | 1 | 1.5 | 2 | 0.2 | 0.3 | 0.4 | 0.5 |
| score | 0.052 | 0.206 | 0.542 | 0.795 | 0.910 | 0.627 | 0.587 | 0.539 | 0.531 |
| LA | 0.054 | 0.180 | 0.511 | 0.722 | 0.828 | 0.042 | 0047 | 0.058 | 0.079 |
| LR | 0.046 | 0.247 | 0.693 | 0.965 | 0.992 | 0.533 | 0.438 | 0.371 | 0.338 |

TABLE 1

*Proportion of simulations for each method that showed p-value* $< 0.05$ *at given data generating model and* $\alpha$ *level. We use two functions for* $\rho$*, hyperbolic tangent and quadratic function. The likelihood ratio test was conducted under the assumption that* $\rho$ *is hyperbolic tangent (tanh) function. Proposed score test performs better than liquid association in all cases.*

**4. Applications to GTEx Data.** We next apply this method to African American samples from GTEx. We aim to find a group of genes that change its co-expression structure as the genome's proportion of African ancestry changes. The proportion of African ancestry for each individual is defined as global ancestry, and it is referred from software LAMP [19]. The data sets are explained in more detail in the Appendix.

We first conduct the data analysis on the supra-pubic skin tissue (not sun exposed), where 31 African American samples are available. Due to low sample size, we restrict our search space to only transcription factors, which are known to have high correlation with many other genes. Therefore, if their impact sizes on other genes are different based on genetic ancestry, such relationship could have important biological implications.

4.1. *Data.* We use the genotype data and normalized gene expression level data from GTEx V6p release [16] to apply the method to the African American samples and their gene co-expression 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 [19]. 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 define global ancestry as a value between 0 and 1 that quantifies the proportion of African chromosome in each subject. We first estimate it by averaging the inferred local ancestry, and this estimate is cross-checked with principal component analysis which can effectively cluster the subjects into subpopulations [21]. 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.

We limit our analysis of real data to transcription factors. We acquired a

list of transcription factors from TFcheckpoint database [3]. We also acquire a list of target gense for each transcription factors from TF2DNA database [22]. We only took into consideration target genes with the highest binding scores.

4.2. *Analysis.* For each transcription factor encoding gene $k$, we compute the pair-wise test-statistic $q_{kj}$ for all its target genes $j = 1, \cdots, J_k$. Then, we compute $d_k = \sum_{j=1}^{J_k} q_{kj}$ to test the hypothesis that correlation between the transcription factor $k$ and its targets remain the same across different genetic industry.

We first compare two discrete networks between European Americans and African Americans using binary indicator vector as the covariate $X$, i.e. $x_i = 1$ if subject $i$ is African American, and $x_i = 0$ if subject $i$ is European American. We first compute scores $q_{kj}$ for all the transcription factor encoding gene $k$ and its target genes $j = 1, \cdots, J_k$. Then we sum over all the targets for each transcription factor to compute $d_k$. Then we divide the sum with the number of targets to make a heuristic comparison against $\chi^2_1$ distribution. This procedure essentially observes the average score of all the target genes for a given transcription factor encoding gene, and under the null, the expectation is 1, although the variance is not trivial due to high dependence. Then we choose the top 10 genes with the highest score to further analyze them using the permutation test.

We repeat the same procedure for the African American samples, where $X$ is the vector of global ancestry ranging from 0.3 to 0.95.

4.3. *Results.* Table (2) summarizes the top 20 transcription factors with the highest $d_k$ values with their $p$-values computed from sequential permutation tests for two cases: (1) comparing two discrete networks, one of African Americans and the other of European Americans, (2) comparing continuously varying coexpression network among African Americans with respect to their global ancestry.

| AA vs EA | | | AA only | | |
|---|---|---|---|---|---|
| genes | $p$-value | tissue | genes | $p$-value | tissue |
| ZNF474 | $< 10^{-6}$ | adrenal gland | ZBTB20 | $1.622{\cdot}10^{-5}$ | adipose |
| HOXA4 | $< 10^{-6}$ | artery coronary | KLF7 | $3.461{\cdot}10^{-5}$ | adipose |
| SP5 | $< 10^{-6}$ | artery coronary | ISL2 | $4.00{\cdot}10^{-5}$ | whole blood |
| ZNF638 | $< 10^{-6}$ | breast tissue | ZNF285 | $6.579{\cdot}10^{-5}$ | heart |
| MEF2C | $< 10^{-6}$ | fibroblasts | E2F5 | $8.00{\cdot}10^{-5}$ | adrenal gland |
| HMBOX1 | $< 10^{-6}$ | fibroblasts | SREBF1 | $8.55{\cdot}10^{-5}$ | nerve tibial |
| TFCP2L1 | $< 10^{-6}$ | esophagus | ZKSCAN3 | $8.62{\cdot}10^{-5}$ | heart |
| SMARCC2 | $< 10^{-6}$ | esophagus | NR4A2 | $9.00{\cdot}10^{-5}$ | whole blood |
| TFDP2 | $< 10^{-6}$ | esophagus | ZNF682 | $1.29{\cdot}10^{-4}$ | adipose |
| DZIP1L | $< 10^{-6}$ | lung | TFAP2E | $1.44{\cdot}10^{-4}$ | adrenal gland |
| TAX1BP1 | $< 10^{-6}$ | stomach | SMAD7 | $1.48{\cdot}10^{-4}$ | artery tibial |
| ZFHX4 | $< 10^{-6}$ | testis | ZSCAN4 | $1.65{\cdot}10^{-4}$ | muscle skeletal |
| ZBTB20 | $2.28{\cdot}10^{-5}$ | adipose | ZNF528 | $1.90{\cdot}10^{-4}$ | skin |
| HOXB7 | $2.53{\cdot}10^{-5}$ | artery coronary | HOXB6 | $2.30{\cdot}10^{-4}$ | esophagus |
| ZSCAN4 | $3.58\cdot10^{-5}$ | muscle skeletal | FOXK1 | $2.70{\cdot}10^{-4}$ | lymphocytes |
| ZSCAN9 | $6.19\cdot10^{-5}$ | muscle skeletal | ZNF440 | $2.86{\cdot}10^{-4}$ | skin |
| ZNF799 | $6.43\cdot10^{-5}$ | testis | HOXB6 | $3.03{\cdot}10^{-4}$ | artery coronary |
| ZKSCAN3 | $8.37\cdot10^{-5}$ | heart | PLAGL2 | $3.17{\cdot}10^{-4}$ | artery tibial |
| SMAD7 | $1.19\cdot10^{-4}$ | artery tibial | MYB | $3.57{\cdot}10^{-4}$ | nerve tibial |
| POU6F2 | $1.14{\cdot}10^{-4}$ | testis | POU6F2 | $3.92{\cdot}10^{-4}$ | testis |

TABLE 2

## 5. Appendix.

5.1. *Derivation of $q$.* The likelihood of the model in (2.5) is

$$\ell(\boldsymbol{\alpha},\boldsymbol{\beta}) = -\frac{N}{2}log(2\pi) - \frac{1}{2}\sum_i log(2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})) - \sum_{i=1}^N \frac{(w_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}$$
$$- \frac{N}{2}log(2\pi) - \frac{1}{2}\sum_i log(2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})) - \sum_{i=1}^N \frac{(v_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}.$$

The first derivative is

$$\boldsymbol{d}_{\boldsymbol{\alpha}} = \frac{\partial \ell}{\partial \boldsymbol{\alpha}}$$
$$= -\frac{1}{2}\sum_{i=1}^N \frac{2\rho'(\boldsymbol{x}_i^T\boldsymbol{\alpha})\boldsymbol{x}_i}{2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}\left(1-\frac{u_w^2}{2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}\right) + \frac{1}{2}\sum_{i=1}^N \frac{2\rho'(\boldsymbol{x}_i^T\boldsymbol{\alpha})\boldsymbol{x}_i}{2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}\left(1-\frac{u_v^2}{2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}\right).$$

Now, we replace $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ with the maximum likelihood estimators under the null hypothesis. $\boldsymbol{\beta}$ is therefore replaced by the OLS estimators, and $\boldsymbol{\alpha}$

is 0. Moreover, the MLE for $2+2\rho(\mathbf{0})=\hat{\sigma}_w^2$ and MLE for $2-2\rho(\mathbf{0})=\hat{\sigma}_v^2$.

$$\tilde{\boldsymbol{d}}_{\boldsymbol{\alpha}} = \boldsymbol{d}_{\boldsymbol{\alpha}} \mid_{\boldsymbol{\alpha}=\mathbf{0},\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}} \tag{5.1}$$

$$= -\frac{1}{2}\sum_i \left( \frac{2\rho'(0)\boldsymbol{x}_i}{\hat{\sigma}_w^2}\left(1-\frac{\hat{u}_{wi}^2}{\hat{\sigma}_w^2}\right) - \frac{2\rho'(0)\boldsymbol{x}_i}{\hat{\sigma}_v^2}\left(1-\frac{\hat{u}_{vi}^2}{\hat{\sigma}_v^2}\right)\right) \tag{5.2}$$

$$= -\rho'(0)\sum_i \left( \frac{\boldsymbol{x}_i}{\hat{\sigma}_w^2}\left(1-\frac{\hat{u}_{wi}^2}{\hat{\sigma}_w^2}\right) - \frac{\boldsymbol{x}_i}{\hat{\sigma}_v^2}\left(1-\frac{\hat{u}_{vi}^2}{\hat{\sigma}_v^2}\right)\right) \tag{5.3}$$

For the second derivative,

$$\begin{aligned}
I_{\boldsymbol{\alpha}\boldsymbol{\alpha}^{\boldsymbol{T}}} &= \frac{\partial^2 \ell}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^{\boldsymbol{T}}} \\
&= -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\alpha}}\sum_{i=1}^N \boldsymbol{x}_i \frac{2\rho'(\boldsymbol{x}_i^T\boldsymbol{\alpha})}{2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}\left(1-\frac{u_{vi}^2}{2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}\right) \\
&+ \frac{1}{2}\sum_{i=1}^N \boldsymbol{x}_i \frac{-2\rho'(\boldsymbol{x}_i^T\boldsymbol{\alpha})}{2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}\left(1-\frac{u_{vi}^2}{2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}\right) \\
&= -\frac{1}{2}\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^T\Bigg( \frac{(2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))(2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})-u_{wi}^2)2\rho''(\boldsymbol{x}_i^T\boldsymbol{\alpha})+(2u_{wi}^2-(2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})))(2\rho'(\boldsymbol{x}_i^T\boldsymbol{\alpha})^2}{(2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))^3} \\
&\quad - \frac{(2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))(2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})-u_{vi}^2)(-2\rho''(\boldsymbol{x}_i^T\boldsymbol{\alpha}))+(2u_{vi}^2-(2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})))(-2\rho'(\boldsymbol{x}_i^T\boldsymbol{\alpha}))^2}{(2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))^3}\Bigg) \\
&= -\frac{1}{2}\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^T\Bigg( 2\rho''(\boldsymbol{x}_i^T\boldsymbol{\alpha})\left(\frac{(2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))(2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})-u_{wi}^2)}{(2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))^3} + \frac{(2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))(2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha})}{(2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))^3}\right. \\
&\quad \left. + 2\rho'(\boldsymbol{x}_i^T\boldsymbol{\alpha})^2\left(\frac{2u_{wi}^2-(2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))}{(2+2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))^3} + \frac{2u_{vi}^2-(2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))}{(2-2\rho(\boldsymbol{x}_i^T\boldsymbol{\alpha}))^3}\right)\Bigg)
\end{aligned}$$

Plugging in $\boldsymbol{\alpha}=\mathbf{0}$ and $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$, above becomes

$$\begin{aligned}
\tilde{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}^{\boldsymbol{T}}} &= I_{\boldsymbol{\alpha}\boldsymbol{\alpha}} \mid_{\boldsymbol{\alpha}=\mathbf{0},\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}} \\
&= -\frac{1}{2}\sum_{i=1}^N \boldsymbol{x_i}\boldsymbol{x_i}^T\Bigg( \frac{\hat{\sigma}_w^2(\hat{\sigma}_w^2-\hat{u}_{wi}^2)2\rho''(0)+(2\hat{u}_{wi}^2-\hat{\sigma}_w^2)(2\rho'(0))^2}{\hat{\sigma}_w^6} - \frac{\hat{\sigma}_v^2(\hat{\sigma}_v^2-\hat{u}_{vi}^2)2\rho''(0)+(2\hat{u}_{vi}^2-\hat{\sigma}_v^2)(2\rho'(0)}{\hat{\sigma}_v^6} \\
&= -\rho''(0)\sum_{i=1}^N \boldsymbol{x_i}\boldsymbol{x_i}^T\left(\frac{\hat{\sigma}_w^2(\hat{\sigma}_w^2-\hat{u}_{wi}^2)}{\hat{\sigma}_w^6} - \frac{\hat{\sigma}_v^2(\hat{\sigma}_v^2-\hat{u}_{vi}^2)}{\hat{\sigma}_v^6}\right) - 2\rho'(0)^2\sum_{i=1}^N \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{x_i}\left(\frac{2\hat{u}_{wi}^2-\hat{\sigma}_w^2}{\hat{\sigma}_w^6} - \frac{2\hat{u}_{vi}^2-\hat{\sigma}_v^2}{\hat{\sigma}_v^6}\right)
\end{aligned}$$

Since $\hat{\sigma}_w^2 = \frac{1}{N}\sum_{i=1}^N \hat{u}_{wi}^2$,

$$\sum_{i=1}^N \boldsymbol{x_i}\boldsymbol{x_i}^T\frac{\hat{\sigma}_w^2(\hat{\sigma}_w^2-\hat{u}_{wi}^2)}{\hat{\sigma}_w^6} = \frac{1}{\hat{\sigma}_w^4}\sum_{i=1}^N\left(\frac{1}{N}\boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{x_i}\sum_{j=1}^N \hat{u}_{wj}^2 - \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{x_i}\hat{u}_{wi}^2\right) = 0$$

and similarly,

$$\sum_{i=1}^{N} \boldsymbol{x_i^T x_i} \frac{\hat{\sigma}_v^2(\hat{\sigma}_v^2 - \hat{u}_{vi}^2)}{\hat{\sigma}_v^6} = 0.$$

Therefore,

$$\begin{aligned}\tilde{I}_{\boldsymbol{\alpha\alpha}} &= I_{\boldsymbol{\alpha\alpha}} \mid_{\boldsymbol{\alpha}=\mathbf{0},\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}} \\ &= 2\rho'(0)^2 \left(\frac{1}{\hat{\sigma}_w^4} + \frac{1}{\hat{\sigma}_v^4}\right) \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T\end{aligned}$$

Therefore, the test statistic $q$ is

$$q = \tilde{d}_{\boldsymbol{\alpha}} \tilde{I}_{\boldsymbol{\alpha\alpha}}^{-1} \tilde{d}_{\boldsymbol{\alpha}} \tag{5.4}$$

$$= \frac{1}{2} \cdot \frac{1}{\frac{1}{\hat{\sigma}_w^4} + \frac{1}{\hat{\sigma}_v^4}} \left(\sum_{i=1}^{N} \boldsymbol{x_i} \left(\frac{1}{\hat{\sigma}_w^2} - \frac{\hat{u}_{wi}^2}{\hat{\sigma}_w^4}\right) - \left(\frac{1}{\hat{\sigma}_v^2} - \frac{\hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right)\right)^T \left(\sum_{i=1}^{N} \boldsymbol{x_i^T x_i}\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{x_i} \left(\frac{1}{\hat{\sigma}_w^2} - \frac{\hat{u}_{wi}^2}{\hat{\sigma}_w^4}\right) - \left(\frac{1}{\hat{\sigma}_v^2} -\right.\right.$$

5.2. *Derivation of* $\eta_{12,13}$. We start from the test statistic $q$ computed when the covariates have been orthogonalized.

$$q = \sum_{p=1}^{P} \left(\frac{1}{\sqrt{N}} \sqrt{\frac{\hat{\sigma}_w^4 \hat{\sigma}_v^4}{\hat{\sigma}_w^4 + \hat{\sigma}_v^4}} \sum_{i=1}^{N} x_{ip} \left(\frac{\hat{\sigma}_w^2 - \hat{u}_{wi}^2}{\hat{\sigma}_w^4} - \frac{\hat{\sigma}_v^2 - \hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right)\right)^2 = \sum_{p=1}^{P} r_p^2$$

$r_p$ for each $p$ follows standard normal distribution by the central limit theorem.

$$E\left(x_{ip}\left(\frac{\hat{\sigma}_w^2 - \hat{u}_{wi}^2}{\hat{\sigma}_w^4} - \frac{\hat{\sigma}_v^2 - \hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right)\right) = x_{ip} E\left(\frac{\hat{\sigma}_w^2 - \hat{u}_{wi}^2}{\hat{\sigma}_w^4}\right) E\left(\frac{\hat{\sigma}_v^2 - \hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right) = 0$$

$$\begin{aligned}Var\left(x_{ip}\left(\frac{\hat{\sigma}_w^2 - \hat{u}_{wi}^2}{\hat{\sigma}_w^4} - \frac{\hat{\sigma}_v^2 - \hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right)\right) &= x_{ip}^2 \left(Var\left(\frac{\hat{\sigma}_w^2 - \hat{u}_{wi}^2}{\hat{\sigma}_w^4}\right) + Var\left(\frac{\hat{\sigma}_v^2 - \hat{u}_{vi}^2}{\hat{\sigma}_v^4}\right)\right) \\ &= \frac{\hat{\sigma}_w^4 + \hat{\sigma}_v^4}{\hat{\sigma}_w^4 \hat{\sigma}_v^4}\end{aligned}$$

We can re-write $r_{12,p}$ as following as a pre-processing to compute $cov(r_{12,p}, r_{12,p})$,

$$\hat{\sigma}_w^2 = 2 + 2\hat{\rho}_{12}$$

$$\hat{\sigma}_v^2 = 2 - 2\hat{\rho}_{12}$$

where $\rho_{12}$ is the constant correlation between variables 1 and 2 under the null hypothesis. Note that asymptotically $\hat{\sigma}_w^2$ converges to $\sigma_w^2$ in probability, and so does $\hat{\rho}_{12}$ to $\rho_{12}$.

$$
\begin{aligned}
&r_{p,12} \\
&= \frac{1}{\sqrt{2N}} \frac{1}{\sqrt{8(1+\hat{\rho}_{12}^2)}} \sum_{i=1}^{N} x_{ip} \left( (2-2\hat{\rho}_{12}) - \frac{2-2\hat{\rho}_{12}}{2+2\hat{\rho}_{12}} (\hat{u}_{1i}+\hat{u}_{2i})^2 \right) - \left( (2+2\hat{\rho}_{12}) - \frac{2+2\hat{\rho}_{12}}{2-2\hat{\rho}_{12}} (\hat{u}_{1i}-\hat{u}_{2i}) \right) \\
&= \frac{1}{\sqrt{16N(1+\hat{\rho}_{12}^2)}} \sum_{i=1}^{N} x_{ip} \left( \frac{4((\hat{\rho}_{12}^3 - \hat{\rho}_{12}) - \hat{u}_{1i}\hat{u}_{2i}(\hat{\rho}_{12}^2+1) + \hat{\rho}_{12}(\hat{u}_{1i}^2+\hat{u}_{2i}^2))}{(1-\hat{\rho}_{12})(1+\hat{\rho}_{12})} \right) \\
&= \frac{1}{\sqrt{N(1+\hat{\rho}_{12}^2)}} \sum_{i=1}^{N} x_{ip} \left( \frac{(\hat{\rho}_{12}^3 - \hat{\rho}_{12}) - \hat{u}_{1i}\hat{u}_{2i}(\hat{\rho}_{12}^2+1) + \hat{\rho}_{12}(\hat{u}_{1i}^2+\hat{u}_{2i}^2))}{(1-\hat{\rho}_{12})(1+\hat{\rho}_{12})} \right)
\end{aligned}
$$

Similarly,

$$
r_{p,13} = \frac{1}{\sqrt{N(1+\hat{\rho}_{13}^2)}} \sum_{i=1}^{N} x_{ip} \left( \frac{(\hat{\rho}_{13}^3 - \hat{\rho}_{13}) - \hat{u}_{1i}\hat{u}_{3i}(\hat{\rho}_{13}^2+1) + \hat{\rho}_{13}(\hat{u}_{1i}^2+\hat{u}_{3i}^2))}{(1-\hat{\rho}_{13})(1+\hat{\rho}_{13})} \right)
$$

$$
cov(r_{12,p}, r_{13,p}) = E(r_{12,p}r_{13,p}) - E(r_{12,p})E(r_{13,p}) = E(r_{12,p}r_{13,p})
$$

Then, after some algebra,

(5.5)

$$
\begin{aligned}
\eta_{12,13} = E \sum_{i=1}^{N} x_{ip}^2 &\frac{\hat{u}_{i1}^2\hat{u}_{2i}\hat{u}_{3i}(\hat{\rho}_{12}^2+1)(\hat{\rho}_{13}^2+1) + \hat{\rho}_{12}\hat{\rho}_{13}(\hat{u}_{1i}^2+\hat{u}_{2i}^2)(\hat{u}_{1i}^2+\hat{u}_{3i}^2)}{N(1-\hat{\rho}_{12})^2(1-\hat{\rho}_{13})^2\sqrt{(1+\hat{\rho}_{12}^2)(1+\hat{\rho}_{13}^2)}} \\
&- \frac{\hat{\rho}_{12}(\hat{\rho}_{13}^2+1)(\hat{u}_{1i}\hat{u}_{3i})(\hat{u}_{1i}^2+\hat{u}_{2i}^2) + \hat{\rho}_{13}(\hat{\rho}_{12}^2+1)(\hat{u}_{1i}\hat{u}_{2i})(\hat{u}_{1i}^2+\hat{u}_{3i}^2)}{N(1-\hat{\rho}_{12})^2(1-\hat{\rho}_{13})^2\sqrt{(1+\hat{\rho}_{12}^2)(1+\hat{\rho}_{13}^2)}}
\end{aligned}
$$

(5.6)

$$
\begin{aligned}
&= \frac{(\rho_{23}+2\rho_{12}\rho_{23})(\rho_{12}^2+1)(\rho_{13}^2+1) + \rho_{12}\rho_{13}(6+2(\rho_{12}+\rho_{13}+\rho_{23}))}{(1-\rho_{12})^2(1-\rho_{13})^2\sqrt{(1+\rho_{12}^2)(1+\rho_{13}^2)}} \\
&- \frac{\rho_{12}(\rho_{13}^2+1)(3\rho_{13}+\rho_{13}+2\rho_{12}\rho_{23}) - \rho_{13}(\rho_{12}^2+1)(3\rho_{12}+\rho_{12}+2\rho_{13}\rho_{23})}{(1-\rho_{12})^2(1-\rho_{13})^2\sqrt{(1+\rho_{12}^2)(1+\rho_{13}^2)}}
\end{aligned}
$$

from

$$
E(\hat{u}_{i1}^4) = 3, \quad E(\hat{u}_{i1}^3\hat{u}_{i2}) = 3\rho_{12}, \quad E(\hat{u}_{i1}^2\hat{u}_{i2}^2) = 1+2\rho_{12}^2,
$$
$$
E(\hat{u}_{i1}^2\hat{u}_{i2}\hat{u}_{i3}) = \rho_{23} + 2\rho_{12}\rho_{13}.
$$

5.3. *Small Sample Correction.* Although the introduced test statistic has convenient asymptotic properties, the sample size is not large enough in many applications. The statistic has its error in the order of $N^{-1}$ [11], and many Monte Carlo experiments show that the test rejects the null hypothesis less frequently than indicated by its nominal size [9, 10, 12] . In response, Harris (1985) used Edgeworth expansion to obtain the distribution and moment generating function to order $n^{-1}$ of the test statistic [11]. Building on this expansion, Honda (1986) and Cribari-Neto and Ferrari 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. [5, 6, 12]

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 $C_\gamma$, the critical value at the level of type I error $\gamma$, i.e. $P(\chi^2_P \geq C_\gamma) = \gamma$, and we refer to this cubic function as $g$ defined as follows.

$$g(C_\gamma) = C_\gamma + C_\gamma\left(\frac{A_3 - A_2 + A_1}{12NP}\right) + C_\gamma^2\left(\frac{A_2 - 2A_3}{(12NP(P+2)}\right) + C_\gamma^3\left(\frac{A_3}{12NP(P+2)(P+4)}\right) = C_\gamma + \tilde{g}(C_\gamma) \tag{5.7}$$

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 [5]. 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 (1995) [5]. Based on the correction of the critical value in (5.7), 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 section 2.4.

Instead, we aim to adjust the test statistic so that the overall shape of null distribution is closer to $\chi^2_P$. 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 $X$.

**References.**

[1] Anil K Bera and Yannis Bilias. Rao's score, neyman's c ($\alpha$) and silvey's lm tests: an essay on historical developments and some new results. *Journal of Statistical Planning and Inference*, 97(1):9–44, 2001.

[2] Trevor S Breusch and Adrian R Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294, 1979.

[3] Konika Chawla, Sushil Tripathi, Liv Thommesen, Astrid Lægreid, and Martin Kuiper. Tfcheckpoint: a curated compendium of specific dna-binding rna polymerase ii transcription factors. *Bioinformatics*, 29(19):2519–2520, 2013.

[4] Lin S Chen, Li Hsu, Eric R Gamazon, Nancy J Cox, and Dan L Nicolae. An exponential combination procedure for set-based association tests in sequencing studies. *The American Journal of Human Genetics*, 91(6):977–986, 2012.

[5] Francisco Cribari-Neto and Silvia LP Ferrari. An improved lagrange multiplier test for heteroskedasticity. *Communications in Statistics-Simulation and Computation*, 24(1):31–44, 1995.

[6] Francisco Cribari-Neto and Silvia LP Ferrari. Monotonic improved critical values for two $\chi 2$ asymptotic criteria. *Economics Letters*, 71(3):307–316, 2001.

[7] Alberto De la Fuente. From 'differential expression'to 'differential networking'–identification of dysfunctional regulatory networks in diseases. *Trends in genetics*, 26(7):326–333, 2010.

[8] Joshua M Galanter, Christopher R Gignoux, Sam S Oh, Dara Torgerson, Maria Pino-Yanes, Neeta Thakur, Celeste Eng, Donglei Hu, Scott Huntsman, Harold J Farber, et al. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. *Elife*, 6:e20532, 2017.

[9] Leslie G Godfrey. Testing for multiplicative heteroskedasticity. *Journal of Econometrics*, 8(2):227–236, 1978.

[10] WE Griffiths and K Surekha. A monte carlo evaluation of the power of some tests for heteroscedasticity. *Journal of Econometrics*, 31(2):219–231, 1986.

[11] P Harris. An asymptotic expansion for the null distribution of the efficient score statistic. *Biometrika*, 72(3):653–659, 1985.

[12] Yuzo Honda. A size correction to the lagrange multiplier test for heteroskedasticity. *Journal of Econometrics*, 38(3):375–386, 1988.

[13] Steve Horvath and Jun Dong. Geometric interpretation of gene coexpression network analysis. *PLoS computational biology*, 4(8):e1000117, 2008.

[14] S Kocherlakota and K Kocherlakota. Neyman's c ($\alpha$) test and rao's efficient score test for composite hypotheses. *Statistics & probability letters*, 11(6):491–493, 1991.

[15] Ker-Chau Li. Genome-wide coexpression dynamics: theory and application. *Proceedings of the National Academy of Sciences*, 99(26):16875–16880, 2002.

[16] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580, 2013.

[17] Peter G Moschopoulos. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathssematics*, 37(1):541–544, 1985.

[18] Jerzy Neyman. C($\alpha$) tests and their use. 1979.

[19] Bogdan Paşaniuc, Justin Kennedy, and Ion Măndoiu. Imputation-based local ancestry inference in admixed populations. In *International Symposium on Bioinformatics Research and Applications*, pages 221–233. Springer, 2009.

[20] Alkes L Price, Nick Patterson, Dustin C Hancks, Simon Myers, David Reich, Vivian G Cheung, and Richard S Spielman. Effects of cis and trans genetic ancestry on gene expression in african americans. *PLoS genetics*, 4(12):e1000294, 2008.

[21] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[22] Mario Pujato, Fabien Kieken, Amanda A Skiles, Nikos Tapinos, and Andras Fiser. Prediction of dna binding motifs from 3d models of transcription factors; identifying tlx3 regulated genes. *Nucleic acids research*, 42(22):13500–13512, 2014.

[23] Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

5747 South Ellis Avenue
Chicago, IL 60637