Published in final edited form as:
Stat Med. 2008 August 30; 27(19): 3847–3867. doi:10.1002/sim.3274.

Testing whether Genetic Variation Explains Correlation of Quantitative Measures of Gene Expression, and Application to Genetic Network Analysis

Zhaoxia Yu[1,3], **Leiwei Wang**[2], **Michelle A.T. Hildebrandt**[2], and **Daniel J. Schaid**[1,*]

[1]Division of Biostatistics, Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN

[2]Division of Clinical Pharmacology, Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic College of Medicine, Rochester, MN

[3]Department of Statistics, University of California, Irvine, CA (current address)

SUMMARY

Genetic networks for gene expression data are often built by graphical models, which in turn are built from pairwise correlations of gene expression levels. A key feature of building graphical models is evaluation of conditional independence of two traits, given other traits. When conditional independence can be assumed, the traits that are conditioned on are considered to "explain" the correlation of a pair of traits, allowing efficient building and interpretation of a network. Overlaying genetic polymorphisms, such as single nucleotide polymorphisms (SNPs), on quantitative measures of gene expression provides a much richer set of data to build a genetic network, because it is possible to evaluate whether sets of SNPs "explain" the correlation of gene expression levels. However, there is strong evidence that gene expression levels are controlled by multiple interacting genes, suggesting that it will be difficult to reduce the partial correlation completely to zero. Ignoring the fact that some set of SNPs can explain at least part of the correlation between gene expression levels, if not all, might miss important clues on the genetic control of gene expression. To enrich the assessment of the causes of correlation between gene expression levels, we develop methods to evaluate whether a set of covariates (e.g., SNPs, or even a set of quantitative expression transcripts), explains at least some of the correlation of gene expression levels. These methods can be used to assist the interpretation of regulation of gene expression and the construction of gene regulation networks.

Keywords

Fisher's z-transformation; Taylor expansion; model selection; pathway; association; multiple regression; optimal linear composites

1. INTRODUCTION

To understand the complexity of the regulation of gene expression, gene pathways and networks have been used [1-3], which provide graphical representations while reducing complexity by modeling local "neighborhoods" of association of measures of gene expression. Genetic networks are often constructed using probabilistic models [4], such as

*Corresponding Author: Daniel J. Schaid, Ph.D. Harwick 775, Section of Biostatistics Mayo Clinic, 200 First Street, SW Rochester, MN 55905 Tel: 507-284-0639 Fax: 507-284-9542 Email: schaid@mayo.edu.

Gaussian graphical models [5-8]. Assuming that variables have a multivariate normal distribution, a Gaussian graphical model can be built from the inverse of the variance-covariance matrix, also called a concentration matrix, because this defines the conditional independence of pairs of variables after taking other variables into consideration. A Gaussian graphical model contains two parts: the vertices, corresponding to the variables in a graph, and the edges that connect pairs of variables that are not independent, either conditionally or unconditionally. Lack of an edge between two vertices represents conditional independence, i.e., the partial correlation between two variables is zero when conditioning on other variables. In contrast, the fact that some variables can reduce the correlation between a pair of variables is often ignored if the partial correlation is not reduced to zero. This could miss important information, because greater insights to gene regulation can be gained by examining whether a set of covariates, either quantitative measures of other expressed genes or genetic markers such as SNPs, can explain at least some of the correlation between the levels of two expressed genes.

Study of correlated gene expression traits is important because genes that are strongly correlated might have similar functions, a feature emphasized in both cluster analysis [9] and gene network reconstruction [10], or might be regulated by a common mechanism. As reviewed by Gibson and Weir [11], several studies show that an expression quantitative trait locus (eQTL) can be correlated with multiple gene expression levels. Therefore, the correlation between a pair of gene expression levels can be reduced when conditioning on a set of genetic variants that regulates the expression levels. On the other hand, a number of studies have illustrated that the regulation of gene expression is genetically complex, influenced by multiple genes and their interactions [12]. For this reason, it may be difficult to reduce the partial correlation between the levels of two expressed genes completely to zero, especially when not all genetic factors are measured.

Elucidating the shared genetic components of correlated gene expression levels has been accessed in different ways. Using a linkage-based design, Zhu et al. [13] defined a weighted correlation measure to assess the degree of association of LOD scores. They reasoned that associated LOD scores may indicate shared genetic components of correlated expression levels. In a more elaborate model, [14] evaluated the relationship of two correlated expression levels while allowing for complex interactions. Using likelihood ratio tests, they evaluated whether the level of an expressed gene depends on the level of expression of a second gene, as well as interaction of the level of expression of the second gene with its own underlying eQTL. To evaluate whether an eQTL influences several traits (i.e., pleiotropy), Li et al. [15] considered an eQTL to have pleiotropic effects on a pair of gene expression levels if the LOD score for the first gene expression level was substantially different from that when conditioning on the second gene expression level. Although these approaches seem reasonable, their statistical properties are not clear and none of them addressed the question whether the correlation between a pair of gene expression levels can be directly explained by genetic markers. We address this question in a direct way by testing whether the marginal correlation is statistically different from the partial correlation, conditional on other variables.

Note that the marginal and partial correlation coefficients are themselves correlated, which must be addressed in our statistical test. Tests of correlated correlation coefficients have been used in both psychology and economics. For example, they have been used to test if a correlation matrix changes over time, to discover equivalent but less expensive measures, and to interpret outcomes from psychological studies. Although these types of statistical comparisons of correlations have been performed in specialized areas of research, they have not been used to evaluate gene expression levels, particularly for interpreting partial correlations for building gene networks.

This paper is organized as follows. We first introduce methods to test the equality of a marginal and a partial correlation, which can be generalized to test two partial correlations. We then propose a two-stage procedure that first selects a set of covariates to condition on and then tests whether the set "explains" at least some of the correlation between two gene expression levels. Simulations that evaluate Type I error rates and power of our methods are presented. Application of the two-stage procedure to a real data set is presented to illustrate our methods. Finally we discuss how our methods can be used to aid the interpretation of gene expressions networks as well as possible extensions of our work to general studies of multivariate traits.

2. STATISTICAL METHODS

Using the delta method, Olkin and Finn [16] provided the asymptotic variance for the difference of the sample marginal and partial correlations, when the partial correlation conditioned on a single covariate. Unfortunately, when the number of covariates increases, it is difficult to obtain the analytical derivates that are required by the delta method. Alternatively, Steiger and Browne [17] proposed a novel indirect approach for testing equality of interdependent statistics, including the partial correlations. Their method has the advantage that it does not require the complicated analytical derivations required by the delta method, and they showed that their approach is asymptotically equivalent to the delta method. In the following, we illustrate both the delta method and the composite method of Steiger and Browne [17].

2.1 Methods for Testing Equality of Marginal and Partial Correlations

Throughout this paper, we use Y_1 and Y_2 to denote the two gene expression levels we wish to correlate, and $X_1,\ldots,X_p$ to denote the covariates we condition on. These covariates could be quantitative measures of other expressed genes, or measures of genetic markers, such as coding each SNP genotype as 0, 1, or 2 according to the number of copies of the rare allele. The partial correlation coefficient $\rho_{Y_1Y_2\bullet X_1,\ldots,X_p}$ is the correlation between Y_1 and Y_2 after conditioning on $X_1,\ldots,X_p$. This partial correlation can be estimated either by inverting the variance-covariance matrix of the Y's and the X's, or by calculating the correlation of the residuals after regressing each of Y_1 and Y_2 on $X_1,\ldots,X_p$. When the null hypothesis of equality of the marginal correlation, $\rho_{Y_1Y_2}$, and the partial correlation, $\rho_{Y_1Y_2\bullet X_1,\ldots,X_p}$, is rejected, the covariates can be considered to "explain" at least part of the correlation of the two gene expression levels. A difficulty of testing the equality of the marginal and the partial correlations comes from the correlation between them.

2.1.1 The Delta Method—The asymptotic variance of the difference between a sample marginal correlation and a sample partial correlation can be partitioned into three parts:

$$\operatorname{var}\left(r_{Y_1Y_2}\right)+\operatorname{var}\left(r_{Y_1Y_2\bullet X_1,\ldots,X_p}\right)-2\operatorname{cov}\left(r_{Y_1Y_2},r_{Y_1Y_2\bullet X_1,\ldots,X_p}\right). \tag{1}$$

For large sample sizes, it is well known [18] that the sample marginal and partial correlations have variances

$$\begin{aligned}\operatorname{var}\left(r_{Y_1Y_2}\right)&=\left(1-\rho_{Y_1Y_2}^2\right)^2/(N-1)\\ \operatorname{var}\left(r_{Y_1Y_2\bullet X_1,\ldots,X_p}\right)&=\left(1-\rho_{Y_1Y_2\bullet X_1,\ldots,X_p}^2\right)^2/(N-1-p),\end{aligned} \tag{2}$$

where N is the sample size. Because both $\rho_{Y_1Y_2}$ and $\rho_{Y_1Y_2 \bullet X_1,\ldots,X_p}$ are functions of the marginal correlation coefficients, ρ_{ij} 's, it is easy to use the delta method [19] to show that the covariance between $r_{Y_1Y_2}$ and $r_{Y_1Y_2 \bullet X_1,\ldots,X_p}$ can be calculated by

$$\operatorname{cov}\left(r_{Y_1Y_2}, r_{Y_1Y_2 \bullet X_1,\ldots,X_p}\right) = \sum_{(i,j)} \operatorname{cov}\left(r_{Y_1Y_2}, r_{ij}\right) \frac{\partial \rho_{Y_1Y_2 \bullet X_1,\ldots,X_p}}{\partial \rho_{ij}}, \tag{3}$$

where the sum is over the $(p + 2)(p + 1)/2$ ordered pairs of $\{i, j\}$ such that all marginal correlations are considered. The covariance of two marginal correlations was first given by Pearson and Filon [20] and later formalized by Olkin and Siotani [21]. For simplicity, we use $\{i, j, k, l\}$ to denote the indices of four variables. The asymptotic covariance of r_{ij} and r_{kl} is

$$\begin{aligned} \operatorname{cov}\left(r_{ij}, r_{kl}\right) &= \left[\rho_{ij}\rho_{kl}\left(\rho_{ik}^2+\rho_{il}^2+\rho_{jk}^2+\rho_{jl}^2\right)/2+\rho_{ik}\rho_{jl}+\rho_{il}\rho_{jk}\right. \\ &- \left.\left(\rho_{ij}\rho_{ik}\rho_{il}+\rho_{ji}\rho_{jk}\rho_{jl}+\rho_{ki}\rho_{kj}\rho_{kl}+\rho_{li}\rho_{lj}\rho_{lk}\right)\right] / (N-1). \end{aligned} \tag{4}$$

In practice, when calculating formulas (1)-(4), sample correlations are substituted for population correlations. The analytic derivatives were provided by Olkin and Finn [16] when p=1. As mentioned by Steiger and Browne [17] and Lord [22], obtaining the analytic formulas for p>1 is complicated. However, using matrix derivatives (Dwyer [23]), we illustrate that the analytic formulas can be expressed in terms of the derivatives of the inverse of the correlation matrix (see the Appendix). Although the formulas can be implemented in standard software, the number of derivatives to evaluate increases quadratically with p, leading to computational inefficiency.

2.1.2 The Indirect Method Based on Optimal Linear Composites—According to a novel approach by Steiger and Browne [17], tests of marginal correlations can be applied directly to correlations of "optimal linear composites". By their Proposition 4, optimal linear composites can also be used for partial correlations. Let s be the vector of the estimated sample variances and covariances, let the corresponding vector for population parameters be σ, let b be a function of s, and let $c(s,b)$ be a differentiable vector-valued function of both s and b. Suppose $\overset{\bullet}{b}(s)$ satisfies

$$\left.\frac{\partial c'[s,b]}{\partial b}\right|_{b=\overset{\bullet}{b}(s)} = 0, \forall s \in N(\sigma),$$

where N(σ) indicates a neighborhood of σ. Clearly, $\overset{\bullet}{b}(s)$ is a saddle point of $c(s,b)$. It also provides the weights for a linear composite of variables to optimize $c(s,b)$ with respect to b. If $\overset{\bullet}{b}(s)$ is also differentiable, then Proposition 4 of Steiger and Browne (1984) states that:

1. $$\left.\frac{\partial c\left[s, \overset{\bullet}{b}(s)\right]}{\partial s}\right|_{s=\sigma} = \left.\frac{\partial c\left[s, \overset{\bullet}{b}(\sigma)\right]}{\partial s}\right|_{s=\sigma},$$

2. $\sqrt{N-1}\left\{c\left[s, \overset{\bullet}{b}(s)\right] - c\left[\sigma, \overset{\bullet}{b}(\sigma)\right]\right\}$ and $\sqrt{N-1}\left\{c\left[s, \overset{\bullet}{b}(\sigma)\right] - c\left[\sigma, \overset{\bullet}{b}(\sigma)\right]\right\}$ have the same asymptotic distribution.

As shown by Steiger and Browne (1984), several optimal correlation coefficients, including the partial correlation coefficient, meet all the requirements of their Proposition 4. Therefore, we can test the hypothesis that the marginal and partial correlations are equivalent based on optimal linear composites. Let $Y_3 = Y_{1 \bullet X_1,\ldots,X_p}$ and $Y_4 = Y_{2 \bullet X_1,\ldots,X_p}$ denote the residuals after regressing each of Y_1 and Y_2 on $X_1,\ldots,X_p$, respectively. Denote $corr(Y_i,Y_j)$ as $\tilde{\rho}_{ij}$. Noting that $\tilde{\rho}_{34}=\rho_{Y_1Y_2 \bullet X_1,\ldots,X_p}$, we treat the Y_i's as observations and test H_0 : $\tilde{\rho}_{12} - \tilde{\rho}_{34}=0$. By Proposition 4 of Steiger and Browne (1984), we merely need to calculate the sample correlation coefficients, the $\tilde{r}_{ij}$'s ($1 \le i \le j \le 4$), based on the Y_i 's, and then plug them into formula (4). Compared with the delta method, this indirect approach is much more computationally efficient because it does not require the derivatives of the partial correlation with respect to the $(p + 2)(p + 1)/2$ marginal correlations.

Notice that covariance formula (3) is a weighted sum of the covariances between $r_{Y_1Y_2}$ and r_{ij}, where the weights are the derivatives. As displayed in the Appendix, the derivatives are functions of only the following partial correlations (i.e., elements of the inverse of the correlation matrix): partials involving Y_1 with each of the X_i 's, partials involving Y_2 with each of the X_i 's, and the partial of Y_1 and Y_2. Note that we can ignore any partial correlations between X_i and X_j for $1 \le i < j \le p$. Therefore, besides the theoretical support by Proposition 4 of Steiger and Browne [17], the approach based on the delta method and that based on the optimal linear composites use similar information from the correlation matrix. When conditioning on one variable, we can prove that the statistic based on the optimal linear composites is exactly the same as that based on the delta method (not shown). When conditioning on two to five variables, we calculated both statistics using Splus (Insightful Corp., Seattle, WA), with different population parameters, and found that the two statistics agreed up to four significant digits. Presumably, the slight differences were caused by rounding errors. In this article, all the results are based on this optimal linear composites method.

A subtle point to consider for calculating the variance of the difference between the marginal and partial correlations for small sample sizes is that we need to replace N - 1 with N - 1 - p for the variance of the partial correlation and its covariance with the marginal correlation, because of the p degrees of freedom required when regressing each of Y_1 and Y_2 on $X_1,\ldots,X_p$. Also notice that the variance of $r_{Y_1Y_2}$ has a factor $1/(N - 1)$, while the variance of $r_{Y_1Y_2 \bullet X_1,\ldots,X_p}$ has a factor $1/(N - 1 - p)$. Thus, the covariance of $r_{Y_1Y_2}$ and $r_{Y_1Y_2 \bullet X_1,\ldots,X_p}$ requires the factor $1/\sqrt{(N-1)\,(N-1-p)}$ instead of $1/(N - 1)$, as displayed in equation (4). Therefore, when the number of covariates, p, to be conditional on is not small, relative to the sample size, instead of using

$$\operatorname{var}\left(r_{Y_1Y_2} - r_{Y_1Y_2 \bullet X_1,\cdots,X_p}\right) = \operatorname{var}\left(\tilde{r}_{12}\right) + \operatorname{var}\left(\tilde{r}_{34}\right) - 2\operatorname{cov}\left(\tilde{r}_{12}, \tilde{r}_{34}\right), \tag{5}$$

one should use the adjusted formula,

$$\operatorname{var}\left(r_{Y_1Y_2} - r_{Y_1Y_2 \bullet X_1,\cdots,X_p}\right) = \operatorname{var}\left(\tilde{r}_{12}\right) + \operatorname{var}\left(\tilde{r}_{34}\right) - 2\sqrt{\frac{N-1}{N-1-p}}\operatorname{cov}\left(\tilde{r}_{12}, \tilde{r}_{34}\right). \tag{6}$$

Although the above derivations are for comparing the marginal and partial correlations, the methods of Steiger and Browne [17] can be extended to test equivalence of two partial correlations, $\rho_{Y_1Y_2 \bullet \underset{\sim}{X}_1}$ and $\rho_{Y_1Y_2 \bullet \underset{\sim}{X}_2}$, where $\underset{\sim}{X}_1$ is a vector of p_1 covariates and $\underset{\sim}{X}_2$ is a vector of p_2

covariates. Again, this is achieved by using optimal linear composites $Y_{1\bullet \underset{\sim}{x}_1}$, $Y_{2\bullet \underset{\sim}{x}_1}$, $Y_{1\bullet \underset{\sim}{x}_2}$, and $Y_{2\bullet \underset{\sim}{x}_2}$.

2.2 Fisher's z-transformation

The Fisher's z-transformation for a correlation r is defined as

$$z(r) = \frac{1}{2}\log\frac{1+r}{1-r},$$

which has mean 1/2log[(1 + ρ)/(1 - ρ)] and variance 1/(N - 3 - p). If r represents a marginal correlation, then p equals 0. The corresponding test statistic,

$$z = \sqrt{N-3-p}\,[z(r) - z(\rho)],$$

has an asymptotic normal distribution with mean 0 and variance 1. Because it converges to the normal distribution much more rapidly than the correlation coefficient, it has been favored as the test statistic rather than the correlation coefficient [24,25]. Using Fisher's z-transformation for the marginal and partial correlations and then the delta method, we have

$$\operatorname{var}\left[z\left(r_{Y_1Y_2}\right) - z\left(r_{Y_1Y_2\bullet X_1,\cdots,X_p}\right)\right] = \frac{1}{N-3} + \frac{1}{N-3-p} - 2\frac{\operatorname{cov}\left(\tilde{r}_{12},\tilde{r}_{34}\right)}{\left(1-\tilde{r}_{12}\right)\left(1-\tilde{r}_{34}\right)}. \quad (7)$$

Applying the small sample adjustment for the covariance in formula (7), we have

$$\operatorname{var}\left[z\left(r_{Y_1Y_2}\right) - z\left(r_{Y_1Y_2\bullet X_1,\cdots,X_p}\right)\right] = \frac{1}{N-3} + \frac{1}{N-3-p} - 2\sqrt{\frac{N-3}{N-3-p}}\frac{\operatorname{cov}\left(\tilde{r}_{12},\tilde{r}_{34}\right)}{\left(1-\tilde{r}_{12}\right)\left(1-\tilde{r}_{34}\right)}. \quad (8)$$

2.3 The Two-stage Procedure

A difficulty in reconstructing gene expression networks is the computational complexities brought by the large number of variables. When the number of the variables is greater than the sample size, the covariance matrix is not of full rank so its inverse cannot be calculated. Even when the sample size is greater than the number of variables, the inverse of a matrix might not be robust due to small eigenvalues [26]. One way to handle the large number of variables is to use a step-wise procedure, as used by [27,28] for developing a dependency network. Similar to these authors, we initially evaluated a step-wise forward procedure to search for a set of covariates that can explain the maximum portion of the correlation between a pair of gene expression levels by testing equality of two partial correlations sequentially. However, when a covariate explains only a small to moderate portion of the correlation between two gene expression levels, the power to detect its effect is weak. To see this, assume the sample size is 50 and the two gene expression levels Y_1 and Y_2 are sums of independent variables:

$$\begin{aligned} Y_1 &= X_1 + X_2 + X_3 + X_u + \varepsilon_1 \\ Y_2 &= X_1 + X_2 + X_3 + X_u + \varepsilon_2, \end{aligned} \quad (9)$$

where X_i has a Normal distribution with mean 0 and variance 20 and ε_1 and ε_2 are independent random errors with variances 20. Further, assume X_1, X_2, and X_3 are observed but X_u is not. The marginal correlation between Y_1 and Y_2 is 0.8; conditional on X_1, the correlation between them is reduced to 0.75. Using the formula in the Appendix, the squared test statistic has an asymptotic chi-square distribution with one degree of freedom and noncentrality parameter 3.05. Based on the nominal Type I error rate 0.05, this value leads to power of only 0.42. Therefore, for each step of the step-wise procedure, we have limited power to include a new variable.

Note that if a covariate can reduce the correlation between Y_1 and Y_2, it has to be correlated with at least one of them because

$$\rho_{Y_1Y_2\bullet X_1}=\frac{\rho_{Y_1Y_2}-\rho_{Y_1X_1}\rho_{Y_2X_1}}{\sqrt{\left(1-\rho_{Y_1X_1}^2\right)\left(1-\rho_{Y_2X_1}^2\right)}}. \tag{10}$$

Furthermore, we have good power to reject $\rho_{Y_1X_1}=0$ because the sample correlation between X_1 and Y_1 is $20/\sqrt{20*100}\approx 0.45$. Based on the variance formula (2), the square of the test statistic has an asymptotic chi-square distribution with one degree of freedom and noncentrality parameter 15.6, and with corresponding power greater than 0.99. Now, if two of the three observed covariates are found to be significantly associated with either Y_1 or Y_2, the marginal correlation of 0.8 is reduced to a partial correlation of 0.67, leading to a noncentrality parameter with value 6.99 and power of 0.75 for testing the equality of the partial and marginal correlations. Ideally, all three observed covariates are identified, in which case the partial correlation is reduced to 0.5, giving a noncentrality parameter of 13.6 and power of 0.96. Therefore, to increase the power of detecting covariates that can explain at least part of the correlation of two gene expression levels, we propose a two-stage procedure:

1. Find the set of covariates to be conditioned on. We considered two methods:
    a. A "union" method that selects covariates that are associated with *either* Y_1 or Y_2, i.e., the set $L_{\text{Union}}=\{X_i:\rho_{Y_1X_i}\neq 0 \text{ or } \rho_{Y_2X_i}\neq 0\}$;
    b. An "intersection" method that selects covariates that are associated with *both* Y_1 and Y_2, i.e., the set $L_{\text{Intersect}}=\{X_i:\rho_{Y_1X_i}\neq 0 \text{ and } \rho_{Y_2X_i}\neq 0\}$.
2. Test the equality of the marginal and partial correlations.

Note that the two methods to select the set of covariates to be conditioned on lead to different explanations. Formula (10) indicates that the partial and marginal correlations will differ even when a set of covariates is correlated with only one gene expression level. When L_{Union} is used, covariates correlated with either of the two gene expression levels will be selected. In this situation, a difference of marginal and partial correlations can result from identification of determinants of the expression level of one of the genes, and not necessarily from sharing of genetic components between the two expressed genes. This strategy is related to Gaussian graphical models, where a partial correlation between a pair of variables is calculated conditional on all other covariates. On the other hand, $L_{\text{Intersect}}$ considers shared covariates of two gene expression levels, which focuses on selecting covariates that "explain" the correlation between a pair of gene expression levels. In our simulation studies, we considered both the union and intersection strategies.

When selecting covariates, a Type I error rate for each test, α, should be chosen such that the family-wise error rate can be controlled at a desired level, such as 0.05. With the Bonferroni correction, we can choose α to be $0.05/p$. It is well known that the Bonferroni correction is

conservative when variables are correlated, which worsens when markers of high density are used because they tend to be in strong linkage disequilibrium (LD). Alternatively, we may consider dividing 0.05 by the effective number of independent covariates. Using λ to denote the vector of the eigenvalues of the correlation matrix for covariates to be tested, Nyholt [29] and Cheverud [30] defined the effective number of independent covariates as

$$p_{\text{eff}} = 1 + (p-1)\left(1 - \frac{Var(\lambda)}{p}\right). \tag{11}$$

This is based on the fact that at the two extremes, i.e.,λ = (p,0,0,…,0) and λ = (1,1,1,…,1), the variance of λ reaches its maximum value of p and its minimum value of 0, respectively. The first extreme is for when all covariates are perfectly correlated so no correction is necessary. In contrast, the second extreme is for when all covariates are mutually independent and the Bonferroni correction is appropriate to use. By estimating $Var(\lambda)$, the effective number of independent variables, p_{eff}, can be estimated by formula (11).

3. SIMULATION METHODS AND RESULTS

We used simulations to investigate the performance of the indirect test of equality of the marginal and partial correlations based on optimal linear composites and the two-stage procedure. For the indirect test, we focused on evaluating its Type I error rate and simulated data based on predefined marginal correlations. For the two-stage procedure, we studied both Type I error rates and power.

3.1 Test Equality of Marginal and Partial Correlations

Variables were assumed to be jointly Gaussian distributed, each with a marginal mean of 0 and variance of 1. We considered sample sizes 50, 100, and 500 and used a variety of correlation structures. Similar to Steiger and Browne [17], we set

$$\rho_{Y_1Y_2} = \frac{\rho_{Y_1X_1}\rho_{Y_2X_1}}{1 - \sqrt{\left(1-\rho^2_{Y_1X_1}\right)\left(1-\rho^2_{Y_2X_1}\right)}}$$

such that the population parameters satisfy the null hypothesis $\rho_{Y_1Y_2} = \rho_{Y_1Y_2\bullet X_1}$. In addition to the 15 conditions they evaluated (see Table 2 of Steiger and Browne [17]), we added another two conditions which represent situations when the marginal correlation coefficients $\rho_{Y_1X_1}$ and $\rho_{Y_2X_1}$ are very small.

When conditioning on only one variable, for given $\rho_{Y_1X_1}$ and $\rho_{Y_2X_1}$, there is a unique $\rho_{Y_1Y_2}$ that satisfies $\rho_{Y_1Y_2} = \rho_{Y_1Y_2\cdot X_1}$. For general values of p, the situation is much more complicated and there are an infinite number of solutions satisfying $\rho_{Y_1Y_2} = \rho_{Y_1Y_2\cdot X_1,\ldots,X_p}$. To simplify our simulations, for given$\rho_{Y_1X_1}$ and $\rho_{Y_2X_1}$, we made the following restrictions:

1. $\rho_{Y_1X_j} = \rho_{Y_1X_1}$, $j = 2,\ldots,p$,
2. $\rho_{Y_2X_j} = \rho_{Y_2X_1}$, $j = 2,\ldots,p$,
3. $\rho_{X_iX_j} = 0$, $1 \le i < j \le p$.

Under these conditions, the $p \times p$ variance-covariance matrix for $X_1, X_2,\ldots,X_p$ is diagonal. Therefore, using the standard variance formula for the joint distribution of Y_1 and Y_2 conditional on $X_1, X_2,\ldots,X_p$, it can be shown that $\rho_{Y_1Y_2} = \rho_{Y_1Y_2\cdot X_1,\ldots,X_p}$ has the unique solution

$$\rho_{Y_1Y_2} = \frac{p\rho_{Y_1X_1}\rho_{Y_2X_1}}{1 - \sqrt{\left(1 - p\rho^2_{Y_1X_1}\right)\left(1 - p\rho^2_{Y_2X_1}\right)}}.$$

The empirical Type I error rate is the chance of obtaining a p-value that is less than the nominal value of 0.05, among 10000 simulations. To compare the effect of Fisher's z-transformation and the small sample adjustment for the covariance, we report results using four different statistics based on: correlation coefficients (formula (5)), correlation coefficients with adjusted covariance (formula (6)), Fisher's z-transformation (formula (7)), and Fisher's z-transformation with adjusted covariance (formula (8)).

The results for N=50 and N=500 are presented in Tables I-IV. Because the results for N=100 were similar to those for N=50, these results are not shown. In each table, the fourth to the eighth columns are the empirical Type I error rates for tests based on correlation coefficients, Fisher's z-transformation, correlation coefficients with small sample covariance adjustment, and Fisher's z-transformation with small sample covariance adjustment. Similar to the findings of Steiger and Browne [17], the empirical Type I error rates based on optimal linear composites were generally close to their nominal values, except when the correlation between $r_{Y_1Y_2}$ and $r_{Y_1Y_2 \cdot X_1}$ is very high. When $\rho_{Y_1X_1}$ and $\rho_{Y_2X_1}$ are close to zero, it can be shown that the correlation between $r_{Y_1Y_2}$ and $r_{Y_1Y_2 \cdot X_1}$ is very high and the variance of ($r_{Y_1Y_2}$ -$r_{Y_1Y_2 \cdot X_1}$) is near zero (see the variance and covariance formulas (2) and (3)). This can cause conservative Type I error rates, which tend to be extremely conservative when the sample size is small. Conditions 1-3 in Tables I-IV were in this situation. As a result, the corresponding Type I error rates based on all four test statistics were much less than the nominal level. It has been shown elsewhere that Fisher's z-transformation converges to a Normal distribution more rapidly than the correlation coefficient. We found that tests based on Fisher's z-transformation gave Type I error rates slightly closer to 0.05 than tests based on correlation coefficients. This advantage was more obvious for smaller sample sizes (N = 50 and N = 100). Tables I and III also show that using small sample covariance adjustment is slightly more beneficial than the unadjusted statistic.

3.2 Identify “Explanatory” Covariates using the Two-Stage Procedure

All simulations in this section were based on a sample size of 50. We assumed that Y_1 and Y_2 each had variance 100, and that each were sums of subsets of mutually independent covariates X_i (i = 1,…,10,u). Each covariate had a Normal distribution with mean 0 and variance σ^2, and ε_1 and ε_2 were independent random errors with variances such that the total variance of each of Y_i was 100. In real data, it is possible that not all covariates shared by Y_1 and Y_2 are measured. Therefore, we assumed X_i (i = 1,…,10) were measured but X_u was not. To cover different situations, we varied the variance $\sigma^2$to be 10, 15, and 20. In each simulation, we applied our two-stage procedure, with the first step of selecting X's based on the Bonferroni correction and the second step of testing equality of the marginal and partial correlations based on Fisher's z-transformation with the small sample covariance adjustment (formula (8)). For each set of chosen parameters, we repeated the simulation 1000 times and estimated power or Type I error rates by the frequency of observing a p-value less than 0.05 in the second stage.

3.2.1 Type I Error Rate of Two-Stage Procedure—We first considered the situation when X_u was the only variable that contributed to gene expression levels other than random error, i.e.,

$$\begin{aligned} Y_1 &= X_u + \varepsilon_1 \\ Y_2 &= X_u + \varepsilon_2. \end{aligned} \quad \text{(Model 1)}$$

Under this assumption, conditioning on a set of observed covariates $X_1,\ldots,X_{10}$, the theoretical partial correlation is the same to the theoretical marginal correlation. The Type I error rates using both L_{Union} and $L_{\text{Intersect}}$ were conservative (Table V). These small Type I error rates agreed with the results shown in Tables I-IV; the Type I error rates for testing the equality of a marginal and a partial correlation are generally conservative when the $\rho_{X_i Y_j}$'s are very small.

We then considered a situation when Y_1 and Y_2 were determined by a shared unobserved covariate and different sets of observed covariates,

$$\begin{aligned} Y_1 &= X_1 + X_2 + X_3 + X_u + \varepsilon_1 \\ Y_2 &= X_4 + X_5 + X_6 + X_u + \varepsilon_2, \end{aligned} \quad \text{(Model 2)}$$

Note that although this model reflects the hypothesis that no observed shared covariate explains the correlation between Y_1 and Y_2, the partial correlation can still differ from the marginal correlation. For example, conditional on covariates X_1 and X_4, the marginal correlation between Y_1 and Y_2 is changed from 0.2 to a partial correlation of 0.25 when σ^2 is 20. Therefore, when L_{Union} is used, the test in the second step of our two-step procedure can detect a difference between the marginal and partial correlations even though there is no overlap of the *X*'s that are associate with the *Y*'s. Our results (Table V) showed that when $L_{\text{Intersect}}$ was used, Type I error rates were close to or smaller than 0.05. However, this was not true for L_{Union} because the partial and marginal correlations differed under Model 2.

3.2.2 Power of Two-Stage Procedure—To assess power of the two-stage procedure, we simulated data using two models:

$$\begin{aligned} Y_1 &= X_1 + X_2 + X_3 + X_u + \varepsilon_1 \\ Y_2 &= X_1 + X_2 + X_3 + X_u + \varepsilon_2, \end{aligned} \quad \text{(Model 3)}$$

$$\begin{aligned} Y_1 &= \sqrt{1.5}X_1 + \sqrt{0.5}X_2 + X_3 + X_u + \varepsilon_1 \\ Y_2 &= \sqrt{0.5}X_1 + \sqrt{1.5}X_2 + X_3 + X_u + \varepsilon_2. \end{aligned} \quad \text{(Model 4)}$$

In Model 3, each covariate had the same effect on Y_1 and Y_2. In contrast, for Model 4, the effects of covariates X_1 and X_2 on Y_1 and Y_2 differed in complementary ways. Besides power, we also report the frequency of detecting true shared observed covariates (X_i, i=1,2,3) identified in the first stage. The results are illustrated in Table VI. Due to the small sample size we used (N = 50), we had limited power to identify all three covariates that are correlated with each Y_i, even when using L_{Union}. For example, for the union method and the data simulated by Model 3, we found that 19.3%, 44.6%, and 60.8% of the time all three covariates were detected when σ^2 had values of 10, 15, and 20, respectively. However, when at least two of the covariates were selected in the first step of our two-step procedure, we had large power to reject the equality of the marginal and partial correlations.

As expected, when $L_{\text{Intersect}}$ was used, power was smaller than that for L_{Union}. This is because $L_{\text{Intersect}}$ uses a more rigorous method to select covariates to be conditioned on,

which reduces the difference between the marginal and partial correlations. When covariates contributed more to one gene expression level than the other, power was also reduced because of the difficulty to detect significantly associated covariates to condition on (Table IVb).

4. APPLICATION TO REAL DATA

4.1 *Background*

The expression of mRNA was measured on the genes that make up the proteasome. The proteasome is a protein complex that degrades unwanted proteins into short polypeptides and amino acids. Three of its beta subunits, β_1, β_2, and β_5 are essential to break down proteins. Because these subunits are functionally related, the amount of expression of these genes tends to be highly correlated, as illustrated in Table VII. Our goal was to identify any SNPs that tend to "explain" at least part of the correlations, since such SNPs might play an important role in co-regulation of the expression of the beta subunits.

4.2 *Methods*

Gene expression levels of the three beta subunit genes (PSMB1, PSMB2, and PSMB5) were measured, along with 90 single nucleotide polymorphisms (SNPs) within these three genes. All measurements were made on 263 Coriell Cell Repository samples, representing unrelated subjects from four ethnic groups (African American, Caucasian American, Han Chinese American, and Mexican American). Excluding five subjects that had either missing expression levels or missing SNP genotypes, we analyzed the remaining 258 individuals. Because most of the measured SNPs had small minor allele frequencies (MAFs), only nine SNPs with MAF greater than 0.05 were used in our analyses. Because the original gene expression levels were skewed, we transformed them using logarithm base two. To avoid possible confounding effects of ethnicity and gender, we regressed the gene expression levels on indicators of ethnic group and gender and used the residuals in the two-stage procedure. We first used only the nine SNPs as covariates in our two-step procedure, and then repeated our tests for each pair of gene expression levels, using the SNPs and the remaining gene expression level as covariates.

4.3 Results

4.3.1 Using only SNPs as Covariates—When using only SNPs as covariates with p=9 for the Bonferroni correction, no SNP was selected in the first stage, even when L_{Union} was used. We then evaluated whether the effective number of independent covariates would be more useful than the conservative Bonferroni correction. The variance of the eigenvalues of the correlation matrix for the nine SNPs was 1.35. By formula (11), the effective number of independent covariates was p_{eff}=7.8. Using this for the Bonferroni correction, SNP rs941718 in gene PSMB5 was correlated with gene expression level PSMB5. However, the marginal and partial correlations were not statistically significantly different. Details about the correlations are reported in Tables VIII-XI. Notice that the correlations did not change very much after conditioning on individual SNPs. As a result, the partial and marginal correlations were highly correlated.

4.3.2 Using both SNPs and Expression Phenotypes as Covariates—When we used the Bonferroni correction that assumed that all covariates were independent and the $L_{\text{Intersect}}$ set, the first step of our two-step procedure always selected the remaining gene expression level when studying the correlation of a pair of gene expression levels. As shown in Table VII, when conditioning on each of the three gene expression levels, namely PSMB1, PSMB2, and PSMB5, the correlations of the other two expression levels were greatly reduced: from 0.519 to 0.390 (p-value = 3.71×10^{-6}) for PSMB1 and PSMB2; from

0.408 to 0.182 (p-value = 2.28×10^{-10}) for PSMB1 and PSMB5; and from 0.553 to 0.411 (p-value = 4.84×10^{-6}) for PSMB2 and PSMB5. Although we did not find a set of SNPs that explained the pairwise correlations among the gene expression levels of PSMB1, PSMB2, and PSMB5, our results suggest that these subunits might nonetheless be interacting with each other. Conditional on one of them, the correlation between the other two can be substantially reduced, although not to a zero level. Further studies are needed to understand the genetic variants that might further explain these correlations.

5. DISCUSSION

We proposed a statistical framework to first select a set of covariates and then to test if the set explains the correlation of two gene expression traits. This is based on the fact that if a set of covariates causes the correlation between a pair of gene expression levels, then the partial correlation between the two gene expression levels can be reduced relative to the marginal correlation when conditioning on the set of covariates. An implicit assumption of our methods is that relationships among variables are linear. If this is not true, but relationships are monotonic, then our methods could be applied to Spearman correlations and partial correlations, which can be achieved by replacing observations with their ranks.

We first introduced two asymptotic tests to test the difference between a marginal and a partial correlation, which Steiger and Browne [17] showed to be asymptotically equivalent. The results from our simulations suggest that the tests generally achieve the nominal Type I error rate when the correlation between response Y and covariates (X's) is not small, especially for the test based on Fisher's z-transformation with adjustment for small samples. We also developed a two-stage procedure to first search for a set of covariates to condition on, and then to test whether the partial correlation differs from the marginal correlation. Currently we only focus on the relationship of a pair of gene expression levels. We may also want to test whether the correlation matrix of a group of gene expression levels changes after conditioning on some covariates. The work by Steiger and Browne [17] provides guidelines for comparing whether two correlated correlation matrices are statistically different or not. Notice that the null hypothesis, i.e., the partial and marginal correlations are equal, indicates a special structure about the correlation matrix. Therefore, it can be tested using methods derived for testing different structural equation models [31]. However, to do that, we need to write the partial correlation as a function of all pairwise marginal correlations. As mentioned by Lord [22] and Steiger and Browne [17], this again leads to computational complexities.

In the first step of our two-step procedure, we proposed a “union” and an “intersection” method to select covariates for conditioning at the second stage. The union method selects covariates that are associated with at least one of the two gene expression levels, while the intersection method selects covariates that are associated with both gene expression levels. While the union method tends to follow the procedures of a Gaussian graphical model, the intersection method is more closely related with testing whether a set of covariates “explains” the correlation of two gene expression levels. It is possible, however, for the union method to detect more subtle associations. For example, a covariate could be truly associated with each of two expression levels, but achieve statistical significance for only one of the traits due to random fluctuation, and perhaps limited power.

Our goal was to test whether a set of SNPs can explain the correlation between two gene expression levels. If the linear correlation between two gene expression traits is completely caused by a set of SNPs, and there are no other casual factors, then the expected partial correlation will be reduced to zero. Although we expect the partial correlation to be closer to zero than the marginal correlation, it is possible that this may not occur. The sign of the

difference between marginal and partial correlations depends on several factors, such as whether the effect of SNPs on a gene expression level is positive or negative, the sign of the correlation between a pair of gene expression levels, as well as whether there are unidentified or unmeasured factors. Similar to the decomposition of the marginal correlation provided by Jones and West [32], a partial correlation can also be decomposed into sums of "path" weights using marginal correlations in a gene relevance network [33] (i.e. a covariance graph that is based on marginal correlations). Consider a situation where a set of SNPs increases the expression level for one gene but decreases the expression level of another gene, i.e.,

$$\begin{aligned} Y_1 &= X_1 + X_2 + X_3 + X_u + \varepsilon_1 \\ Y_2 &= -X_1 - X_2 - X_3 + X_u + \varepsilon_2, \end{aligned}$$

and all covariates are independent and have a normal distribution with mean 0 and variance 20. The two random errors are assumed independent with variances 20. With these parameters, the marginal correlation between Y_1 and Y_2 is -0.4. In contrast, after conditioning on the three covariates X_1, X_2, and X_3, the partial correlation is 0.5. Therefore, conditional on a subset of the shared covariates, the partial correlation can be on the opposite side of zero from the marginal correlation. For this reason, we prefer two-sided tests than one-sided tests in the second step of our two-step approach.

Zhu et al. [13] found that LD among genetic markers can influence the construction of genetic networks. Suppose there are two SNPs in LD and that each SNP is associated with a different gene expression level. Also assume that only one of the two SNPs was measured. Then we might falsely draw the conclusion that the measured SNP explains the correlation between the two gene expression levels, at least partially. Therefore, although we have used the word "explain" throughout this paper, the results based on our methods can only be interpreted in a statistical way, not in a biological way. In other words, our statistical tests can only provide suggestive results whose biological meaning has to be further identified and confirmed by carefully designed experiments.

As shown by Brem and Krugylak [12], many expression traits are regulated by non-additive genetic mechanisms. Therefore, studying gene-gene interactions (epistatsis) and gene-environment interactions can be crucial to construct a realistic network. Because our current procedure only considers covariates that are at least marginally correlated with gene expression levels, interactions among covariates that represent gene-gene and gene-environmental interactions with weak marginal effects may not be detected. In the future, we plan to generalize our covariate selection method in the first stage such that interactions among genetic markers, quantitative gene expression traits, and environmental factors can be modeled in an effective way.

In the covariate selection step of our union and intersection methods, we selected covariates that showed marginal effects. It is possible to enhance this step by penalization techniques that improve regression model prediction and interpretation (e.g. parsimony), particularly when the number of covariates is much larger than the sample size. Some recent developments in this area are lasso regression, a penalized least squares method that imposes an L_1 penalty on the regression coefficients [34,35], and the elastic net penalty that uses a convex combination of lasso and ridge regression penalties, and hence capitalizes on the strengths of each [36]. Schäfer and Strimmer [37] found that the maximum likelihood estimate or the unbiased empirical covariance matrix is not an accurate estimate of the true covariance matrix when the number of variables is similar to or greater than the number of observations. To obtain an accurate and reliable estimate of the covariance matrix, they

proposed a shrinkage approach. Note, however, that substituting alternative covariate selection methods or estimators for covariance matrices still allows use of our general procedures.

Our methods might be useful while attempting to construct genetic networks by integrating gene expression levels and SNPs. This could be achieved by treating all the gene expression levels other than the pair of interest, as well as the SNPs, as candidate covariates that might explain the correlation. Evaluating whether the partial correlation differs from the marginal correlation can generate more insights to underlying biological processes than simply testing conditional independence, i.e., whether the partial correlation differs from zero, the approach taken by traditional Gaussian graphical models. Note that Kulp and Jagalur [14] also developed regression models to integrate gene expression levels with SNPs. Their approach considered one expressed gene to be a target (dependent variable in regression) and another to be a regulator (independent variable), and allowed the SNPs of the regulator gene to interact with the level of expression of the regulator, building a model for trans-acting regulators where the expression of the target gene depends on both the genotype and expression of the regulator. Our approach is more general as a screening tool, because the SNPs and expression levels we condition on can be anywhere in the genome.

Finally, although our work was motivated to evaluate whether some genetic components can explain at least part of the correlation between a pair of gene expression levels, our proposed methods to test the equality of a marginal and a partial correlation can be used in many other multivariate settings. For example, in an epidemiological study, one could test if the correlation between measures of calcium and low density lipid is due to variation in a set of risk factors, such as body mass index, hypertension, smoking, and diabetes. Our proposed procedure can be used to test if individual risk factors or a set of risk factors can explain at least part of the correlation between traits of interest.

Acknowledgments

This work was supported by the U.S. Public Health Service, National Institutes of Health, contract grant number GM065450, the Pharmacogenetic Research Network, U01, GM61388, and PhRMA Foundation Center of Excellence in Clinical Pharmacology Award. We thank Mr. Paul Doran and Affymetrix for the support of this study. We would also like to thank the referee for valuable comments.

APPENDIX

Derivative of $\rho_{Y_1Y_2 \bullet X_1,\ldots,X_p}$ with respect to a marginal correlation

To follow the notation used by Dwyer [23], let $Z_1 = Y_1$, $Z_2 = Y_2$ and $Z_{i+2} = X_i$. Let Σ denote the correlation matrix of $\underset{\sim}{Z} = \left(Z_1, Z_2, \ldots, Z_{p+2}\right)$ and $\langle\Sigma\rangle_{ij}$ denote the (i, j) element of Σ. It is easy to show that

$$\rho_{12\bullet 34,\ldots,(p+2)} = -\left\langle\Sigma^{-1}\right\rangle_{12}\left\langle\Sigma^{-1}\right\rangle_{11}^{-1/2}\left\langle\Sigma^{-1}\right\rangle_{22}^{-1/2}.$$

By results from Dwyer (1967), the derivative of $\rho_{12\bullet 34,\ldots,(p+2)}$ with respect to a matrix Σ is

$$\begin{aligned}\frac{\partial\rho_{12\bullet 34,\ldots,(p+2)}}{\partial\Sigma} &= \Sigma^{-1}K_{12}\Sigma^{-1}\left\langle\Sigma^{-1}\right\rangle_{11}^{-1/2}\left\langle\Sigma^{-1}\right\rangle_{22}^{-1/2} \\ &\quad -\tfrac{1}{2}\Sigma^{-1}K_{11}\Sigma^{-1}\left\langle\Sigma^{-1}\right\rangle_{12}\left\langle\Sigma^{-1}\right\rangle_{11}^{-3/2}\left\langle\Sigma^{-1}\right\rangle_{22}^{-1/2} \\ &\quad -\tfrac{1}{2}\Sigma^{-1}K_{22}\Sigma^{-1}\left\langle\Sigma^{-1}\right\rangle_{12}\left\langle\Sigma^{-1}\right\rangle_{11}^{-1/2}\left\langle\Sigma^{-1}\right\rangle_{22}^{-3/2}\end{aligned}$$

where K_{ij} is a $(p + 2)\times(p + 2)$ matrix with the (i, j) element equal to 1 and 0 elsewhere. Notice that each ρ_{ij} appears twice in Σ. Therefore,

$$\begin{aligned}\frac{\partial \rho_{12\bullet 34,\ldots,(p+2)}}{\partial \rho_{ij}} &= \left(\left\langle \Sigma^{-1} K_{12} \Sigma^{-1} \right\rangle_{ij} + \left\langle \Sigma^{-1} K_{12} \Sigma^{-1} \right\rangle_{ji}\right) \left\langle \Sigma^{-1} \right\rangle_{11}^{-1/2} \left\langle \Sigma^{-1} \right\rangle_{22}^{-1/2} \\ &\quad - \left\langle \Sigma^{-1} K_{11} \Sigma^{-1} \right\rangle_{ij} \left\langle \Sigma^{-1} \right\rangle_{12} \left\langle \Sigma^{-1} \right\rangle_{11}^{-3/2} \left\langle \Sigma^{-1} \right\rangle_{22}^{-1/2} \\ &\quad - \left\langle \Sigma^{-1} K_{22} \Sigma^{-1} \right\rangle_{ij} \left\langle \Sigma^{-1} \right\rangle_{12} \left\langle \Sigma^{-1} \right\rangle_{11}^{-1/2} \left\langle \Sigma^{-1} \right\rangle_{22}^{-3/2}.\end{aligned}$$

Because the above formula involves many manipulations of sparse matrices, we may further simplify the derivative as

$$\begin{aligned}\frac{\partial \rho_{12\bullet 34,\ldots,(p+2)}}{\partial \rho_{ij}} = \frac{\langle \sigma^{-1} \rangle_{12}}{\langle \sigma^{-1} \rangle_{11}^{1/2} \langle \sigma^{-1} \rangle_{12}^{1/2}} = \Bigg(&\frac{\langle \sigma^{-1} \rangle_{i1} \langle \sigma^{-1} \rangle_{2j}}{\langle \sigma^{-1} \rangle_{12}} + \frac{\langle \sigma^{-1} \rangle_{j1} \langle \sigma^{-1} \rangle_{2i}}{\langle \sigma^{-1} \rangle_{12}} \\ &- \frac{\langle \sigma^{-1} \rangle_{i1} \langle \sigma^{-1} \rangle_{1j}}{\langle \sigma^{-1} \rangle_{11}} - \frac{\langle \sigma^{-1} \rangle_{i2} \langle \sigma^{-1} \rangle_{2j}}{\langle \sigma^{-1} \rangle_{22}} \Bigg)\end{aligned}$$

REFERENCES

1. Akutsu, S.; Kuhara, T.; Maruyama, O.; Minyano, S. Identification of gene regulatory networks by strategic gene disruptions and gene over-expressions. Proc. Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, ACM-SIAM; 1998.
2. Chen, T.; Filkov, V.; Skiena, S. Identifying gene regulatory networks from experimental data. Proc. Third Annual International Conference on Computational Molecular Biology (RECOMB); 1999. p. 94-103.
3. Friedman N, Linial M, Nachman I, Pe'er. Using Bayesian networks to analyze expression data. Journal of Computational Biology. 2000; 7:601–620. [PubMed: 11108481]
4. Heckerman, D. In A tutorial on learning with Bayesian networks. Jordan, M., editor. MIT Press; Cambridge, MA:
5. Dempster A. Covariance selection. Biometrics. 1972; 28:157–175.
6. Whittaker, J. Graphical Models in Applied Multivariate Statistics. Wiley; New York: 1990.
7. Edwards, D. Introduction to Graphical Modeling. Springer; New York: 1995.
8. Cox, D.; Wermuth, N. Multivariate dependencies: models, analysis and interpretation. Chapman and Hall; London: 1996.
9. Eisen M, Spellman P, Brown P, Bostein D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA. 1998; 95:14863–14868. [PubMed: 9843981]
10. Stuart J, Segal E, Koller D, Kim S. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003; 302:249–255. [PubMed: 12934013]
11. Gibson G, Weir B. The quantitative genetics of transcription. Trends Genet. 2005; 21:616–623. [PubMed: 16154229]
12. Brem R, Storey J, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. Nature. 2005; 436:701–703. [PubMed: 16079846]
13. Zhu J, Lum P, Lamb J, Guhathakurta D, Edwards S, Thieringer R, Berger J, Wu M, Thompson J, Sachs A, Schadt E. An integrative genomics approach to the reconstruction of gene networks in segregating population. Cytogenet Genome Res. 2004; 105:363–374. [PubMed: 15237224]
14. Kulp D, Jagalur M. Causal inference of regulator-target pairs by gene mapping of expression phenotypes. BMC Genomics. 2006; 7:125. [PubMed: 16719927]
15. Li R, Tsaih S, Shockley K, Stylianou I, Wergedal J, Paigen B, Churchill G. Structural model analysis of multiple quantitative traits. PLOS Genetics. 2006; 2:1046–1057.
16. Olkin I, Finn J. Correlation Redux. Psychological Bulletin. 1995; 118:155–164.

17. Steiger J, Browne M. The comparison of interdependent correlations between optimal linear composites. Psychometrika. 1984; 49:11–24.
18. Anderson, T. An introduction to Multivariate Statistical Analysis. Wiley; New York: 1984.
19. Rao, C. Linear statistical inference and its application. John Wiley & Sons; New York: 1973.
20. Pearson K, Filon L. Mathematical contributions to the theory of evolution: IV. On the probable error of frequency constants and on the influence of random selection of variation and correlation. Philosophical Transactions of the Royal Society of London, Series A. 1898; 191:229–311.
21. Olkin, I.; Siotani, M. In Asymptotic distribution of functions of a correlation matrix. Ikeda, S., editor. Shinki Tsusho; Tokyo: p. 235-251.
22. Lord F. Automated hypothesis tests and standard errors for nostandard problems. The American Statistician. 1975; 29:56–59.
23. Dwyer P. Some applications to marker derivatives in multivariate analysis. American Statistical Association Journal. 1967; 62:607–625.
24. Meng X, Rosenthal R, Rubin D. Comparing correlated correlation coefficients. Psychological Bulletin. 1992; 111:172–175.
25. Hittner J, May K, Silver N. A Monte Carlo evaluation of tests for comparing dependent correlations. J Gen Psychol. 2003; 130:149–168. [PubMed: 12773018]
26. Kishino H, Waddell P. Correspondence analysis of genes and tissue types and finding genetic link from microarray data. Genome Informatics. 2000; 11:83–95. [PubMed: 11700590]
27. Heckerman D, Chickering D, Meek C, Rounthwaite R, Kadie C. Dependency networks for inference, collaborative filtering, and data visualization. Journal of Machine Learning Research. 2000; 1:49–75.
28. Dobra A, Hans C, Jones B, Nevins J, Yao G, West M. Sparse graphical models for exploring gene expression data. Multivariate Analysis. 2004; 90:196–212.
29. Nyholt D. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004; 74:765–769. [PubMed: 14997420]
30. Cheverud J. A simple correction for multiple comparisons in interval mapping genome scans. Heredity. 2001; 87:52–58. [PubMed: 11678987]
31. Jöreskog KG. Structural analysis of covariance and correlation matrices. Psychometrika. 1978; 43:443–477.
32. Jones B, West M. Covariance decomposition in undirected Gaussian graphical models. Biometrika. 2005; 92:779–786.
33. Butte A, Tamayo P, Slonim T, Golub T, Kohane I. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc. Natl. Acad. Sci. USA. 2000; 97:12182–12186. [PubMed: 11027309]
34. Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc B. 1996; 58:267–288.
35. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. Annals of Statistics. 2006; 34:1436–1462.
36. Zou H, Tibshirani R. Regularization and variable selection via the elastic net. J R Statist Soc B. 2005; 67:301–320.
37. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Appl. Genet. Mol. Biol. 2005; 4:32.

Table I

Type I error rates for testing the equality of marginal correlation $\rho_{Y_1Y_2}$ and partial correlation $\rho_{Y_1Y_2 \cdot X_1}$ when sample size is 50

Condition	$\rho_{Y_1X_1}$	$\rho_{Y_2X_1}$	$\rho_{Y_1Y_2}$	*r*	Fisher's z	*r* adjusted	Fisher's z adjusted
1	0	0.05	0	.0004	.0003	.0005	.0004
2	0.05	0.10	.7991	.0003	.0008	.0005	.0010
3	0.10	0.30	.5901	.0054	.0068	.0066	.0087
4	0.10	0.50	.3615	.0240	.0225	.0284	.0270
5	0.10	0.70	.2418	.0498	.0456	.0520	.0486
6	0.10	0.90	.1589	.0548	.0476	.0554	.0484
7	0.10	0.95	.1378	.0589	.0511	.0593	.0513
8	0.30	0.50	.8627	.0170	.0326	.0195	.0365
9	0.30	0.70	.6588	.0386	.0464	.0401	.0490
10	0.30	0.90	.4622	.0517	.0515	.0529	.0527
11	0.30	0.95	.4059	.0528	.0492	.0533	.0499
12	0.50	0.70	.9173	.0282	.0465	.0289	.0487
13	0.50	0.90	.7229	.0390	.0483	.0397	.0490
14	0.50	0.95	.6511	.0415	.0463	.0416	.0469
15	0.70	0.90	.9148	.0331	.0515	.0333	.0517
16	0.70	0.95	.8558	.0372	.0506	.0373	.0508
17	0.90	0.95	.9897	.0296	.0527	.0297	.0529

Table II

Type I error rates for testing the equality of marginal correlation $\rho_{Y_1Y_2}$ and partial correlation $\rho_{Y_1Y_2 \bullet X_1}$ when sample size is 500

Condition	$\rho_{Y_1X_1}$	$\rho_{Y_2X_1}$	$\rho_{Y_1Y_2}$	*r*	Fisher's z	*r* adjusted	Fisher's z adjusted
1	0	0.05	0	.0007	.0007	.0011	.0011
2	0.05	0.10	.7991	.0047	.0054	.0063	.0071
3	0.10	0.30	.5901	.0391	.0400	.0412	.0417
4	0.10	0.50	.3615	.0484	.0485	.0495	.0497
5	0.10	0.70	.2418	.0489	.0486	.0491	.0488
6	0.10	0.90	.1589	.0490	.0483	.0490	.0485
7	0.10	0.95	.1378	.0486	.0476	.0486	.0478
8	0.30	0.50	.8627	.0439	.0461	.0444	.0470
9	0.30	0.70	.6588	.0506	.0512	.0508	.0514
10	0.30	0.90	.4622	.0510	.0509	.0511	.0510
11	0.30	0.95	.4059	.0517	.0514	.0517	.0515
12	0.50	0.70	.9173	.0465	.0483	.0465	.0483
13	0.50	0.90	.7229	.0501	.0507	.0501	.0508
14	0.50	0.95	.6511	.0481	.0482	.0481	.0483
15	0.70	0.90	.9148	.0454	.0461	.0454	.0461
16	0.70	0.95	.8558	.0500	.0520	.0500	.0520
17	0.90	0.95	.9897	.0452	.0475	.0452	.0475

Table III

Type I error rates for testing the equality of marginal correlation $\rho_{Y_1Y_2}$ and partial correlation $\rho_{Y_1Y_2 \cdot X_1X_2X_3X_4}$ when sample size is 50

Condition	$\rho_{Y_1X_1}$	$\rho_{Y_2X_1}$	$\rho_{Y_1Y_2}$	*r*	Fisher's z	*r* adjusted	Fisher's z adjusted
1	0	0.05	0	.0015	.0007	.0027	.0021
2	0.05	0.10	.7963	.0006	.0012	.0021	.0044
3	0.10	0.20	.7843	.0060	.0095	.0118	.0191
4	0.10	0.30	.5551	.0219	.0224	.0291	.0322
5	0.10	0.40	.3882	.0465	.0419	.0512	.0492
6	0.10	0.45	.3142	.0487	.0431	.0518	.0469
7	0.20	0.30	.8996	.0156	.0283	.0199	.0395
8	0.20	0.40	.7110	.0333	.0407	.0370	.0463
9	0.20	0.45	.5995	.0426	.0463	.0451	.0495
10	0.30	0.40	.9231	.0259	.0451	.0282	.0494
11	0.30	0.45	.8291	.0358	.0486	.0367	.0504
12	.0.40	.0.45	.9750	.0339	.0489	.0347	.0502

Table IV

Type I error rates for testing the equality of marginal correlation $\rho_{Y_1Y_2}$ and partial correlation $\rho_{Y_1Y_2 \cdot X_1X_2X_3X_4}$ when sample size is 500

Condition	ρ_{13}	ρ_{23}	$\rho_{12}(=\rho_{12.3456})$	*r*	Fisher's z	*r* adjusted	Fisher's z adjusted
1	0	0.05	0	.0020	.0013	.0051	.0050
2	0.05	0.10	.7963	.0173	.0158	.0231	.0245
3	0.10	0.20	.7843	.0419	.0417	.0455	.0460
4	0.10	0.30	.5551	.0424	.0422	.0440	.0444
5	0.10	0.40	.3882	.0495	.0489	.0501	.0500
6	0.10	0.45	.3142	.0505	.0498	.0505	.0505
7	0.20	0.30	.8996	.0470	.0482	.0478	.0498
8	0.20	0.40	.7110	.0494	.0503	.0500	.0511
9	0.20	0.45	.5995	.0486	.0490	.0489	.0494
10	0.30	0.40	.9231	.0480	.0494	.0484	.0499
11	0.30	0.45	.8291	.0505	.0526	.0506	.0527
12	0.40	0.45	.9750	.0501	.0515	.505	.0517

Table V

Type I error rates of the two-stage procedure

	Model 1		Model 2	
	Union	**Intersection**	**Union**	**Intersection**
σ^2=10	.002	.000	.039	.025
σ^2=15	.001	.001	.089	.039
σ^2=20	.001	.002	.346	.050

Table VI

Power of the two-stage procedure

a) Model 3

	Union					Intersection				
	Probability of detecting 0, 1, 2, or 3 covariates				Power	Probability of detecting 0, 1, 2, or 3 covariates				Power
	0	1	2	3		0	1	2	3	
σ^2=10	.066	.313	.428	.193	.550	.503	.373	.114	.010	.320
σ^2=15	.011	.125	.418	.446	.848	.144	.394	.348	.114	.660
σ^2=20	.000	.040	.352	.608	.960	.017	.183	.460	.340	.882

b) Model 4

	Union					Intersection				
	Probability of detecting 0, 1, 2, or 3 covariates				Power	Probability of detecting 0, 1, 2, or 3 covariates				Power
	0	1	2	3		0	1	2	3	
σ^2=10	.046	.288	.440	.226	.469	.603	.327	.061	.009	.268
σ^2=15	.002	.071	.371	.556	.738	.283	.457	.205	.055	.479
σ^2=20	.000	.004	.195	.801	.783	.104	.407	.383	.106	.601

Stat Med. Author manuscript; available in PMC 2009 August 30.

Table VII

Correlation coefficients and p-values for test of equality of marginal and partial correlations

	PSMB1	PSMB2	PSMB5
PSMB1		.519	.408
PSMB2	.390 (3.71e-06)		.533
PSMB5	.182 (2.28e-10)	.411 (4.84e-06)	

Values in upper triangle are marginal pairwise correlations. Values in lower triangle are partial correlations for pairs of phenotypes conditional on the remainder phenotype and values in the parentheses are p-values for test of equality of marginal and partial correlations.

Table VIII

Analyses of correlation of PSMB1 and PSMB2

Gene	SNP Location	dbSNP	Minor allele frequency	$r_{i,PSMB1}$(1)	$r_{i,PSMB2}$(2)	$r_{PSMB1,PSMB2}$(3)	$r_{PSMB1,PSMB2 \cdot i}$(4)	Corr(5)	P-value(6)
PSMB1	5'-FR(-1043)	rs2179373	0.384	-0.055	-0.017	0.519	0.519	0.996	.978
PSMB1	Ex1(31)	n/a(7)	0.465	-0.065	-0.007	0.519	0.519	0.995	.838
PSMB1	IVS2(49)	rs2076319	0.382	-0.009	-0.040	0.519	0.519	0.997	.972
PSMB1	IVS4(35)	n/a(7)	0.052	-0.069	-0.034	0.519	0.518	0.996	.800
PSMB1	IVS4(128)	rs1474642	0.138	-0.107	-0.082	0.519	0.515	0.992	.424
PSMB2	5'-FR(-477)	rs676614	0.500	0.004	-0.017	0.519	0.519	0.998	.889
PSMB5	5'-FR(-693)	rs8020463	0.434	-0.039	0.040	0.519	0.521	0.995	.520
PSMB5	3'-FR(1042)	rs941718	0.349	-0.056	0.031	0.519	0.521	0.994	.493
PSMB5	3'-FR(1103)	rs941717	0.455	-0.042	0.051	0.519	0.522	0.994	.449

(1)The correlation between the *i*th marker and gene expression level PSMB1.

(2)The correlation between the *i*th marker and gene expression level PSMB2.

(3)The correlation between gene expression levels PSMB1 and PSMB2.

(4)The correlation between gene expression levels PSMB1 and PSMB2 conditional on the *i*th marker.

(5)Correlation between the marginal and partial correlations.

(6)P-values to test the equality of the marginal and partial correlations.

(7)A novel SNP that is not in the National Center for Biotechnology Information (NCBI) database.

Table IX

Analyses of correlation of PSMb1 and PSMb5[(1)]

Gene	SNP Location	dbSNP	Minor allele frequency	$r_{i,PSMB1}$	$r_{i,PSMB2}$	$r_{PSMB1,PSMB2}$	$r_{PSMBl,PSMB2'i}$	Corr	P-value
PSMB1	5'-FR(-1043)	rs2179373	0.384	-0.055	-0.015	0.408	0.408	0.997	.949
PSMB1	Ex1(31)	n/a	0.465	-0.065	-0.019	0.408	0.408	0.996	.926
PSMB1	IVS2(49)	rs2076319	0.382	-0.009	-0.002	0.408	0.408	0.998	.996
PSMB1	IVS4(35)	n/a	0.052	-0.069	-0.012	0.408	0.408	0.996	.964
PSMB1	IVS4(128)	rs1474642	0.138	-0.107	-0.031	0.408	0.407	0.992	.887
PSMB2	5'-FR(-477)	rs676614	0.500	0.004	-0.002	0.408	0.408	0.998	.968
PSMB5	5'-FR(-693)	rs8020463	0.434	-0.039	0.094	0.408	0.414	0.990	.377
PSMB5	3'-FR(1042)	rs941718	0.349	-0.056	0.170	0.408	0.424	0.974	.149
PSMB5	3'-FR(1103)	rs941717	0.455	-0.042	0.147	0.408	0.419	0.981	.246

[(1)]See footnotes of Table VIII.

Stat Med. Author manuscript; available in PMC 2009 August 30.

Table X

Analyses of correlation of PSMB2and PSMB5[(1)]

Gene	SNP Location	dbSNP	Minor allele frequency	$r_{i,PSMB1}$	$r_{i,PSMB2}$	$r_{PSMB1,PSMB2}$	$r_{i,PSMB1,PSMB2'i}$	Corr	P-value
PSMB1	5'-FR(-1043)	rs2179373	0.384	-0.017	-0.015	0.533	0.533	0.998	.885
PSMB1	Ex1(31)	n/a[(7)]	0.465	-0.007	-0.019	0.533	0.533	0.998	.983
PSMB1	IVS2(49)	rs2076319	0.382	-0.040	-0.002	0.533	0.533	0.997	.861
PSMB1	IVS4(35)	n/a[(7)]	0.052	-0.034	-0.012	0.533	0.533	0.997	.966
PSMB1	IVS4(128)	rs1474642	0.138	-0.082	-0.031	0.533	0.532	0.995	.893
PSMB2	5'-FR(-477)	rs676614	0.500	-0.017	-0.002	0.533	0.533	0.998	.961
PSMB5	5'-FR(-693)	rs8020463	0.434	0.040	0.094	0.533	0.532	0.994	.822
PSMB5	3'-FR(1042)	rs941718	0.349	0.031	0.170	0.533	0.535	0.991	.738
PSMB5	3'-FR(1103)	rs941717	0.455	0.051	0.147	0.533	0.532	0.987	.884

[(1)]See footnotes of Table VIII.