High throughput screening of co-expressed gene pairs with controlled False Discovery Rate (FDR) and Minimum Acceptable Strength (MAS) *

Dongxiao Zhu[1,2], Alfred O Hero[2], Zhaohui S Qin[3] and Anand Swaroop[4]

Address correspondence to: Dongxiao Zhu, Bioinformatics 2017H Palmer Commons, University of Michigan, Ann Arbor, MI 48109. Email: zhud@umich.edu. Tel. (734) 763-0564. FAX: (734) 763-8041.

Abstract

Many exploratory microarray data analysis tools such as gene clustering and relevance networks rely on detecting pairwise gene co-expression. Traditional screening of pairwise co-expression either controls biological significance or statistical significance, but not both. The former approach does not provide stochastic error control, and the later approach screens many co-expressions with excessively low correlation. We have designed and implemented a statistically sound two-stage co-expression detection algorithm that controls both statistical significance (False Discovery Rate, FDR) and biological significance (Minimum Acceptable Strength, MAS) of the discovered co-expressions. Based on estimation of pairwise gene correlation, the algorithm provides an initial co-expression discovery that controls only FDR, which is then followed by a second stage co-expression discovery which controls both FDR and MAS. It also computes and thresholds the set of FDR p-values for each correlation that satisfied the MAS criterion. Using simulated data, we validated asymptotic null distributions of the Pearson and Kendall correlation coefficients and the two-stage error-control procedure; we also compared our two-stage test procedure with another two-stage test procedure using Receiver Operating Characteristic (ROC) curve. We then used yeast galactose metabolism data to illustrate the advantage of our method for clustering genes and constructing a relevance network. The method has been implemented in an R package "GeneNT" that is freely available from the Comprehensive R Archive Network (CRAN): http://cran.r-project.org/

*[1]Bioinformatics Program, University of Michigan, Ann Arbor, MI 48109. [2]Departments of EECS and Statistics, University of Michigan, Ann Arbor, MI 48109. [3]Center for Statistical Genetics and Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109.[4]Departments of Ophthalmology, Visual Sciences and Human Genetics, University of Michigan, Ann Arbor, MI 48109.

Key words: False Discovery Rate Confidence Interval, Relevance Network, Error control, Gene pathway

1 Introduction

The emergence and development of DNA microarray technology (Affymetrix oligonucleotide expression arrays and cDNA arrays) enable researchers to interrogate gene expression levels simultaneously on the genome scale (Lockhart *et al*. 1996, Schena *et al*. 1995, DeRisi *et al*. 1997). The development of statistically sound and biologically meaningful techniques to analyze gene expression data is essential for transforming raw experimental data into scientific knowledge. Gene expression data have been subjected to a variety of statistical analyses, such as detecting differentially expressed genes (e.g. Tusher *et al*. 2001, Zareparsi *et al*. 2004), clustering genes/samples (e.g. Eisen *et al*. 1998, McLachlan *et al*. 2002), and cancer classification (e.g. Golub *et al*. 1999, Alizadeh *et al*. 2000).

Detection of co-expressed genes from microarray data has attracted much attention since many co-expressed genes are found to have functional relationships, e.g. lying in the same signal transduction pathway (Eisen *et al*. 1998, DeRisi *et al*. 1997). Hierarchical clustering Eisen *et al*. 1998 and relevance network construction (Butte and Kohane 2000, Farkas *et al*. 2003) are two important exploratory techniques. Both of these techniques are based on discovering pairs of co-expressed genes, which is one of the fundamental objectives in functional genomics and system biology. While not necessarily true in many higher Eukaryotes (Boutanaev *et al*. 2002), pairwise gene co-expression as prescribed by the standard two-component model (Nixon *et al*. 1986) characterizes gene co-expression in Bacteria, single-celled Eukaryotes, Archaea and higher Plants (Stock *et al*. 2000).

Clearly, there is a need for statistical methodology for high throughput screening of co-expressed gene pairs with stochastic error and strength of association controls. Two issues have to be considered in developing such a methodology, namely, choice of screening statistic and choice of screening acceptance and rejection criteria.

Regarding the choice of screening statistic, several methods have been adopted to measure the strength of association between expression profiles of gene pairs, such as: Pearson correlation coefficient (Zhou *et al*. 2002), coherence (Butte *et al*. 2001), mutual information (Butte and Kohane 2000), edge detection (Filkov *et al*. 2002), and dominant spectral component analysis (Yeung *et al*. 2004). Each of these methods has advantages and disadvantages. The Pearson correlation coefficient has been one of the most popular choices because it is easily computed and its performance is often comparable to more complex and computational intense methods (Yeung *et al*. 2004). However, the Pearson correlation coefficient can only capture linear relationships between gene

expression profiles. To circumvent this limitation, we propose to use the non-parametric Kendall rank correlation coefficient that is able to capture both linear and nonlinear associations between gene expression profiles. Both linear and nonlinear associations are very common in cellular gene expression profiles, for example, two functionally related enzymes with similar catalytic activities may have a linear correlation between their expression profiles; while the two enzymes with very different catalytic activities may have a nonlinear correlation. The Pearson and Kendall correlation coefficient measures are especially convenient because their asymptotic distributions are available, as required by our two-stage screening procedure to be described below.

Regarding the choice of screening acceptance criteria, one approach is to calculate a sample correlation for each pair of genes and then to select the top pairs by correlation thresholding (e.g. Zhou *et al*. 2002, and Farkas *et al*. 2003), e.g. those exceeding a minimum acceptable strength (MAS) level specified by the threshold. Without a statistical inference procedure, the observed weakly correlated gene pairs are more likely to due to chance, noise etc, and hence are less likely to be biologically relevant. For this reason the approach controls biological significance. However, it does not account for statistical sampling uncertainty and thus does not control error rate. Another approach (Lee *et al*. 2004) is to control statistical significance in addition to biological significance. It is implemented as a two-stage procedure: screen co-expressed gene pairs whose strength of association is different from zero using p-value thresholding, e.g. as determined by a specified level of family-wise error rate (FWER) or false discovery rate (FDR), followed by a "hard" correlation thresholding. The approach is able to control error rate at correlation level zero but not at any non-zero correlation level.

The purpose of combining correlation thresholding with p-value thresholding is to control sampling error (e.g. Type I, II errors or false discoveries) and systematic error (e.g. non-functionally relevant correlations) incurred in the screening process. Indeed, the sampling error alone can be controlled by adopting a regular hypothesis testing scheme, and the systematic error alone can be controlled by a correlation thresholding. However, a reliable procedure for simultaneously controlling sampling error and systematic error has not been well developed.

In this paper, a new two-stage statistical hypothesis testing scheme is applied in order to decide on whether the strength of association is statistically significant at the specified positive MAS level. Stage I screens statistically significant co-expression gene pairs whose strength of correlation is different from zero. It is then followed by Stage II, in which a "soft" correlation thresholding (FDR Confidence Interval, FDR-CI) instead of a "hard" thresholding is applied. Our method is directly inspired by the two-stage screen methodology of (Hero *et al*. 2004) that controls both False Discovery Rate (FDR) and Minimum Acceptable Difference (MAD) in detecting differentially expressed genes.

We demonstrate the application of our two-stage screening algorithm by constructing relevance networks and clustering co-expressed genes from yeast galactose metabolism data (Ideker *et al*. 2000). This data represents approximately 6200 gene expression levels on two-color cDNA microarrays collected over 20 physiological/genetic conditions (nine mutant and one wild type strains incubated in either GAL-inducing or non-inducing media). By applying our two-stage algorithm on this dataset, we achieved a high specificity (83% - 100%) in discovering genes in the galactose metabolism pathway as described in Section 4.

The outline of the paper is as follows. In Section 3, we describe the proposed two-stage multicriteria approach. In Section 4, we first show the approach indeed controls FDR at the specified MAS level using synthetic data, and then illustrate it for yeast galactose metabolism data. In Section 5, we discuss advantages of our method, model assumptions and restrictions.

2 Methods

2.1 Measures of the strength of association

There are many possible discriminants for strength of association between two variables, which we generally denote as a real number Γ. Under a Gaussian linear hypothesis, the Pearson correlation coefficient ρ is an appropriate metric. A robust distribution-free alternative is the Kendall rank correlation coefficient (Kendall's τ). The Pearson (Bickel and Doksum 2000) and Kendall correlation coefficients (Hollander and Wolfe 1999) are special cases of the generalized correlation coefficient (Daniel 1944). We define $\{g_p\}_{p=1}^G$ as the indices of G gene probes on the microarray; $\{X_{g_p}\}_{p=1}^G$ as normalized probe responses (random variables); and $\{\{x_{g_{p(n)}}\}_{p=1}^G\}_{n=1}^N$ as realizations of $\{X_{g_p}\}_{p=1}^G$ under N i.i.d. microarray experiments.

2.1.1 Pearson correlation coefficient.

The population Pearson correlation coefficient between random variables X_{g_i} and X_{g_j} (defined as long as $\mathrm{var}(X_{g_i})$, $\mathrm{var}(X_{g_j})$ are positive) is:

$$\rho(X_{g_i}, X_{g_j}) = \frac{\mathrm{cov}(X_{g_i}, X_{g_j})}{\sqrt{\mathrm{var}(X_{g_i})\mathrm{var}(X_{g_j})}}. \tag{1}$$

The sample Pearson correlation coefficient $\hat{\rho}$ is an asymptotically consistent unbiased estimator of ρ:

$$\hat{\rho}_{i,j} = \frac{S_{X_{g_i}, X_{g_j}}}{\sqrt{S_{X_{g_i}, X_{g_i}} S_{X_{g_j}, X_{g_j}}}}, \tag{2}$$

where $S_{X_{g_i},X_{g_i}}$, $S_{X_{g_j},X_{g_j}}$, and $S_{X_{g_i},X_{g_j}}$ are sample variances and covariances given by

$$S_{X_{g_i},X_{g_i}} = (N-1)^{-1}\sum_{n=1}^{N}(X_{g_i(n)} - \overline{X_{g_i}})^2,$$

$$S_{X_{g_j},X_{g_j}} = (N-1)^{-1}\sum_{n=1}^{N}(X_{g_j(n)} - \overline{X_{g_j}})^2,$$

$$S_{X_{g_i},X_{g_j}} = (N-1)^{-1}\sum_{n=1}^{N}(X_{g_i(n)} - \overline{X_{g_i}})(X_{g_j(n)} - \overline{X_{g_j}}),$$

and $\overline{X_{g_i}} = N^{-1}\sum_{n=1}^{N} X_{g_i(n)}$, $\overline{X_{g_j}} = N^{-1}\sum_{n=1}^{N} X_{g_j(n)}$ are sample means.

2.1.2 Kendall rank correlation coefficient.

Kendall's τ statistic is a measure of correlation that captures both linear and non-linear associations. The parameter τ is defined as $\tau = P_+ - P_-$, where, for any two independent pairs of observations $(x_{g_{i(n)}}, x_{g_{j(n)}})$, $(x_{g_{i(m)}}, x_{g_{j(m)}})$ from the population: $P_+ = P[(x_{g_{i(n)}} - x_{g_{i(m)}})(x_{g_{j(n)}} - x_{g_{j(m)}}) \geq 0]$ and $P_- = P[(x_{g_{i(n)}} - x_{g_{i(m)}})(x_{g_{j(n)}} - x_{g_{j(m)}}) < 0]$. An unbiased estimator of τ is given by the Kendall τ statistic:

$$\hat{\tau}_{i,j} = 2\sum\sum_{1\leq n\leq m\leq N}\frac{K_{nm}}{N(N-1)}, \tag{3}$$

here K_{nm} is a indicator variable defined as $K_{nm} = \text{sgn}(x_{g_{i(n)}} - x_{g_{i(m)}})\text{sgn}(x_{g_{j(n)}} - x_{g_{j(m)}})$ for each set of pairs drawn from $\{X_{g_i}\}_{i=1}^{G}$ and $\{X_{g_j}\}_{j=1}^{G}$.

2.2 Hypothesis testing scheme

To screen the strongly co-expressed pairs of G genes on each microarray, we will simultaneously test the $\Lambda = \binom{G}{2}$ pairs of composite hypotheses: $\{H_\lambda, K_\lambda : \lambda = (g_i, g_j)\}$.

$$H_\lambda : \Gamma_{g_i,g_j} \leq \textit{cormin} \quad \text{versus} \quad K_\lambda : \Gamma_{g_i,g_j} > \textit{cormin}, \quad \text{for} \quad g_i \neq g_j, \quad \text{and} \quad g_i, g_j \in (1,2,...G) \tag{4}$$

where *cormin* is the specified minimum acceptable strength of correlation. The sample correlation coefficient $\hat{\Gamma}_{i,j}$ ($\hat{\rho}_{i,j}$ or $\hat{\tau}_{i,j}$) could be thresholded to decide on pairwise dependency of two genes in the sample. When we must decide between the null hypothesis H_λ and the alternative hypothesis K_λ based on such a threshold test, there will generally be decision errors in the form of false positives (Type I errors: decide K_λ when H_λ is true) and false negatives (Type II errors: decide H_λ when K_λ is true). The Per Comparison Error Rate (PCER) is defined as the

number of Type I errors over the number of independent trials, i.e. the probability of Type I error. The p-value is the probability that a more improbable sample could have been drawn from the population(s) being tested given the assumption that the null hypothesis is true.

For N realizations of any pair of gene probe responses, $\{x_{g_{i(n)}}, x_{g_{j(n)}}\}_{n=1}^{N}$, we first calculate $\hat{\tau}_{i,j}$ or $\hat{\rho}_{i,j}$ respectively. For large N, the PCER p-values for $\rho_{i,j}$ or $\tau_{i,j}$ are:

$$p_{\rho_{i,j}} = 2\left(1-\Phi\left(\frac{\tanh^{-1}(\hat{\rho}_{i,j})}{(N-3)^{-1/2}}\right)\right) \tag{5}$$

$$p_{\tau_{i,j}} = 2\left(1-\Phi\left(\frac{K}{N(N-1)(2N+5)/18^{1/2}}\right)\right) \tag{6}$$

where Φ is the cumulative density function of a standard Gaussian random variable, and $K = \sum\sum_{1\le n\le m\le N} K_{nm}$. The above expressions are based on asymptotic Gaussian approximations to $\hat{\rho}_{i,j}$ (Bickel and Doksum 2000) and to $\hat{\tau}_{i,j}$ (Hollander and Wolfe 1999).

The PCER p-value refers to the probability of Type I error incurred in testing a single pair of hypothesis for a single pair of genes g_i, g_j. It is the probability that purely random effects would have caused g_i, g_j to be erroneously selected based on observing correlation between this pair of genes only. When considering the Λ multiple hypotheses for all possible pairs, two adjusted error rates have frequently been considered in microarray studies. These are family-wise error rate (FWER) and false discovery rate (FDR)(Benjamini and Hochberg 1995). The FWER is the probability that the test of all Λ pairs of hypotheses yields at least one false positive in the set of declared positive responses. In contrast, the FDR is the average proportion of false positives in the set of declared positive responses. The FDR is dominated by the FWER and is therefore a less stringent measure of significance. As in previous studies (Reiner *et al*. 2003), we adopt the FDR to control statistical significance of the selected gene pair correlations in our screening procedure.

2.3 Two-stage screening procedure

Select a level α of FDR and a level *cormin* of MAS significance levels. We use a modified version of the two-stage screening procedure proposed for gene screening by (Hero *et al*. 2004). This procedure consists of two stages, summarized in Fig 1.

Stage I. For each gene pair $\lambda = (g_i, g_j)$ in the set $\mathcal{G}$ of all $\Lambda = \binom{G}{2}$ gene pairs, test the simple null hypothesis:

$$H_\lambda : \Gamma_{g_i,g_j} = 0 \quad \text{versus} \quad K_\lambda : \Gamma_{g_i,g_j} \neq 0, \text{for} \quad g_i \neq g_j, \text{and} \quad g_i, g_j \in (1,2,...G) \tag{7}$$

Stage I (step-down): control of FDR at MAS = 0.

1. Specify FDR level α and MAS level *cormin*.
2. Compute a list of PCER p-values: $p_1, p_2, ..., p_\Lambda$ corresponding to $\Lambda = \binom{G}{2}$ pairs of composite hypotheses: $\{H_\lambda, K_\lambda : \lambda = (g_i, g_j)\}$ from $\{\hat{\rho}_{i,j}\}$ or $\{\hat{\tau}_{i,j}\}$.
3. Sort the list of PCER p-values in increasing order, i.e. $p_{(1)}, p_{(2)}, ..., p_{(\Lambda)}$.
4. Find the index k_0 where $k_0 = max\{k : p_{(k)} \leq \frac{k\alpha}{\Lambda\nu}\}$.
5. Set initial screening $\mathcal{G}_1$ as those $k_0 = \Lambda_1$ gene pairs having p-values: $p_{(1)}, p_{(2)}, ..., p_{(k_0)}$.

In step 4, $\nu = 1$ if the test statistics can be assumed statistically independent or positively dependent, where $\nu = \sum_{\lambda=1}^{\Lambda} \lambda^{-1}$ under the general dependency assumption.

Stage II: control of FDR and MAS = *cormin*.

1. Construct Λ_1 different $(1-\alpha)\times 100\%$ PCER-CI's $I^\lambda(\alpha)$ for ρ or τ of each gene pair in $\mathcal{G}_1$ (Appendix 5.1).
2. Convert these PCER-CI's into Λ_1 different $(1-\alpha)\times 100\%$ FDR-CI's using formula (Benjamini and Yekutieli 2004): $I^\lambda(\alpha) \rightarrow I^\lambda(\Lambda_1\alpha/\Lambda)$.
3. Select the subset $\mathcal{G}_2$ containing Λ_2 of Λ_1 gene pairs whose FDR-CI's do not intersect $[-cormin, cormin]$.

Figure 1: Two-stage direct screening procedure yields a subset $\mathcal{G}_2$ of all possible gene pairs $\mathcal{G}$ whose strength of association exceeds MAS level *cormin* at FDR level α.

at FDR level α. The step-down procedure of Benjamini and Hochberg (Benjamini and Hochberg 1995) is used to accomplish this.

Stage II. Suppose a number Λ_1 pairs of genes, denoted by the set $\mathcal{G}_1 \subset \mathcal{G}$, pass the Stage I procedure. In Stage II, we first construct asymptotic PCER Confidence Intervals (PCER-CI's): $I^\lambda(\alpha)$ for each Γ (ρ or τ) in subset $\mathcal{G}_1$. We convert these PCER-CI's into FDR Confidence Intervals (FDR-CI's): $I^\lambda(\alpha) \rightarrow I^\lambda(\Lambda_1\alpha/\Lambda)$ using the procedure in (Benjamini and Yekutieli 2004). A gene pair in subset $\mathcal{G}_1$ is declared to be both statistically significant and biologically significant if its FDR-CI does not intersect the MAS interval $[-cormin, cormin]$ (see Fig 5). The set of all such gene pairs is called $\mathcal{G}_2$.

1. Compute a list of PCER p-values: $p_1, p_2, ..., p_\Lambda$ corresponding to $\Lambda = \binom{G}{2}$ pairs of composite hypotheses: $\{H_\lambda, K_\lambda : \lambda = (g_i, g_j)\}$ from $\{\hat{\rho}_{i,j}\}$ or $\{\hat{\tau}_{i,j}\}$.
2. Sort the list of PCER p-values in increasing order, i.e. $p_{(1)}, p_{(2)}, ..., p_{(\Lambda)}$.

for any gene pair $\lambda_0 \in \{g_i, g_j\}_{i,j=1}^{G}$:

- Find the minimal $\alpha = \alpha(\lambda_0)$ such that the PCER-CI $I^{\lambda_0}(\alpha)$ does not intersect $[-cormin, cormin]$.
- Compute the integer index $N(\alpha(\lambda_0)) = \sum_{k=1}^{\Lambda} I(p_{(k)})k \leq \alpha(\lambda_0))$, where $I(A)$ is an indicator function of the truth of statement A. The FDR p-value of the gene pair λ_0 is then simply p_i, where $i = N(\alpha(\lambda_0))$.

endfor

Figure 2: Inverse screening procedure allows the FDR p-value of a gene pair's (λ_0) strength of association to be computed.

In many practical situations, the experimenter may not be comfortable in specifying a MAS or FDR criterion in advance. In this situation, it is useful to solve the inverse problem: what is the most stringent pair of criteria (α , *cormin*) that would cause a particular subset of gene pairs to be included in the screen $\mathcal{G}_2$. The inverse screening procedure is displayed in Fig 2.

3 Results

3.1 Validating the two-stage algorithm

3.1.1 Validating asymptotic null distribution.

Here we verify that the proposed two-stage algorithm controls FDR at a specified MAS level using simulated data. Since the p-values are based on asymptotic distribution approximations (eq. 5 and eq. 6), we display in Fig 3a the goodness of fit of the $\hat{\rho}$ sampling distribution to the Gaussian distribution using QQ plots. Note that there is good agreement to the Gaussian distribution for $N \geq 10$. Moreover, since the construction of confidence intervals requires estimation of sampling distribution variance, the accuracy of the variance approximation is vital. This

can be evaluated by the mean squared approximation error (*MSE*) for sample size N:

$$MSE_{\rho}^{(N)} = \Lambda^{-1} \sum_{1 \le i < j \le G} (S^{(N)}_{\tanh^{-1}(\hat{\rho}_{i,j})} - (N-3)^{-1/2})^2, \tag{8}$$

$$\begin{aligned} MSE_{\tau}^{(N)} = \Lambda^{-1} \sum_{1 \le i < j \le G} (S^{(N)}_{\hat{\tau}_{i,j}} - (\frac{2}{N(N-1)} \frac{2(N-2)}{N(N-1)^2} \\ \sum_{i=1}^{N} (C_r - \overline{C}) + 1 - \hat{\tau}))^2, \end{aligned} \tag{9}$$

where $S^{(N)}_{\tanh^{-1}(\hat{\rho}_{i,j})}$ and $S^{(N)}_{\hat{\tau}_{i,j}}$ denote standard errors of $\tanh^{-1}(\hat{\rho}_{i,j})$ and $\hat{\tau}_{i,j}$ at the sample size N. The definitions of C_r and $\bar{C}$ can be found in Appendix 5.1. The $\hat{\rho}$ variance approximations are seen to be in good agreement even for small sample sizes ($N > 10$) from Fig 3b.

3.1.2 Validating the error control procedure.

In order to validate our FDR and MAS error control procedure, we simulated pairwise gene expression data based on known population covariances (Appendix 5.2). The actual FDR at a MAS level is calculated as a ratio of the number of screened gene pairs whose corresponding population correlation parameters $\Gamma_{i,j}$ are less than the MAS level specified, divided by the total number of screened gene pairs. The actual MAS is the minimum true discovery of population correlation $\Gamma_{i,j}$ among the screened pairs. We specified 16 pairs of (FDR,MAS) criteria (Four FDR levels: 0.2, 0.4, 0.6, 0.8; Four MAS levels: 0.2, 0.4, 0.6, 0.8), and each is plotted as a different upper case English alphabet (Red) in Fig 4. The 16 corresponding pairs of actual (FDR,MAS) criteria are also shown in Fig 4 using the same set of lower case English alphabets (Blue). It can be observed that generally the actual FDR's (lower case) fall below the specified constraint (upper case) and the actual MAS's (lower case) fall above the specified constraints (upper case). Any deviations of actual FDR's and MAS's from their specified levels are due to the conservative asymptotic approximation (Eqs (5) and (6)). Observe that use of Kendall correlation (Fig 4b) leads to greater overestimation of error rates than the Pearson correlation (Fig 4a). Overestimation of error rates will translate into a reduction of power in discovering co-expressed pairs at the specified levels.

3.2 Performance comparison

We compared the performance of the two two-stage algorithms using the Receiving Operator Characteristic (ROC) curve in which "sensitivity" is plotted against "1 - specificity". Let Λ_0 denotes the number of false hypotheses (true strength of pairwise association is smaller than or equal to the threshold *cormin*), and Λ_α denotes

the number of true hypotheses (true strength of pairwise association is greater than the threshold *cormin*). We counted false positives FP (falsely rejected hypotheses) and false negatives FN (falsely accepted hypotheses). The "sensitivity" (True positive rate, pTP) can be calculated as $pTP = 1 - E(FN/\Lambda_\alpha)$; and the "1 - specificity" (False positive rate, pFP) can be calculated as: $pFP = E(FP/\Lambda_0)$. The two-stage algorithm labelled as "FDR-only" in Fig. 5 denotes the FDR test followed by a "hard" correlation thresholding; and that labelled as "FDR-CI" denotes the FDR test followed by a "soft" correlation thresholding (FDR-CI). In Fig. 5, we observe overall better performance of "FDR-CI" test than the "FDR-only" test especially at low levels of correlation thresholding. For example, at the MAS level of 0.2 and the specificity level of 0.9, the "FDR-CI" method has a three-fold higher sensitivity ($pTP \approx 0.6$) than the "FDR-only" method ($pTP \approx 0.2$).

3.3 Constructing relevance networks with controlled FDR and MAS

For the yeast galactose metabolism dataset, a subset of 997 differentially expressed genes were identified by Ideker et al using a generalized likelihood ratio test procedure (Ideker *et al*. 2000). Genes having a likelihood ratio statistic $\lambda \leq 45$ were selected as differentially expressed, i.e. whose mRNA levels differed significantly from the reference under one or more treatments.

Figs. 6a and 6b illustrate the direct implementation of the two-stage procedure to screen positively or negatively correlated gene pairs based on the Pearson correlation coefficient. The direct screening procedure is constrained by FDR level $\alpha = 0.05$ and MAS level $cormin = 0.5$. Stage I of the screen discovered $\Lambda_1 = 153,983$ out of $\Lambda = \binom{997}{2} = 496,506$ gene pairs having FDR ≤ 0.05, leaving 153,983 correlation coefficients for which FDR-CI's must be constructed. Recall that gene pair passes the Stage II screening if the FDR-CI does not intersect the interval $[-0.5, 0.5]$. $\Lambda_2 = 18,135$ of the 153,983 gene pairs passed the Stage II screening and were declared to be both "biologically" and "statistically" significant. Similarly, using Kendall correlation coefficient, there were $\Lambda_1 = 95,205$ gene pairs that passed the Stage I screen, and only $\Lambda_2 = 3,552$ gene pairs passed the Stage II screen constrained by the same MAS and FDR criteria as above (STable 1).

Although for Gaussian distributed pairs the Kendall rank correlation coefficient has lower discovery power compared to the Pearson correlation coefficient, our screening procedure was nevertheless able to pull out many non-linearly correlated gene pairs that were missed by the Pearson correlation procedure. These non-linearly correlated gene pairs, just like those linearly correlated ones, may be biologically relevant too. For example, the link between gene "RPC40" and gene "YDR516C" passed both Stage I and II screening ($\alpha = 0.015$, $cormin = 0.5$) when using Kendall correlation coefficient ($\hat{\tau}$ =-7.5e-01, FDR p-value = 6.2e-04, FDR-CI = [-9.7e-01, -

5.4e-01]), but they failed to pass even the first screening when the Pearson correlation coefficient was used ($\hat{\rho}$ =-6.3e-01, FDR p-value = 1.2e-02). From the scatter plot, we can observe an obvious non-linear correlation for this gene pair (Fig 7). The poor linear fit can be verified by fitting a simple linear regression model and observing $R^2 = 0.36$. Biologically, the gene "RPC40" encodes RNA polymerase (I and III) subunit (transcription apparatus); although the specific function of gene "YDR516C" remains unclear, it is recently shown that it involves in transcriptional induction of the early meiotic-specific transcription factor IME1 Dwight *et al*. 2002. Both genes are thus components of transcription apparatus. Applying our two-stage algorithm based on Pearson correlation coefficient alone will miss the important functional relationship. Therefore, the Kendall correlation statistic can beat the Pearson correlation statistic in some instances and hence the two correlation statistics should be used together to capture functional relationships as many as possible.

Relevance networks are implemented as a graph where n nodes (genes) are connected by p sets of edges (co-expressions). Each of the p sets of edges are discovered using a different similarity measure (Butte *et al*. 2000, Butte and Kohane 2000). Therefore, our constructed networks are mixed networks with $p = 2$ in which edges are discovered using either Pearson correlation coefficients or Kendall correlation coefficients constrained by the same set of (FDR,MAS) criteria. In relevance networks, genes that are of considerable interest to the biologist are "hub genes" such as RPL33A and RPS4A in Fig 8. Hub genes are best connected genes that dominate a large part of the network topology (Jeong *et al*. 2001, Barabàsi 2004). We constructed five such networks that are constrained by five pairs of constraints (FDR ≤ 0.05, $cormin = 0.5, 0.6, 0.7, 0.8, 0.9$). Most of the "hub genes" in each discovered network fall into two categories: "RPL" and "RPS". The former encodes "Ribosome Protein Large (60S) subunit," and the latter encodes "Ribosome Protein Small (40S) subunit". Both of these categories are structural components of the ribosome that is responsible for protein biosynthesis. Protein biosynthesis plays the central role in galactose metabolism because galactose is not a primary carbon source for yeast, when switching from primary carbon sources (glucose) to secondary carbon source (e.g. galactose), many different types of proteins including transporters, enzymes, and regulators have to be synthesized to be able to degrade the secondary carbon source Wieczorke *et al*. 1999. We ranked the "hub genes" by calculating and sorting average rank of each "hub gene" over five networks (Table 1, STable 2). The list of "hub genes" (STable 2) are presumably indispensable for galactose metabolism (Jeong *et al*. 2001).

Fig. 8 presents the discovered network topology with a FDR level of 5% (5% discovered edges are expected to be false positive) at the MAS level of $cormin = 0.9$. The network is composed of 89 connected vertices and 132 edges. Similar to some other biological networks, the network marginal degree distributions appear to

be of the form of a power-law. This was tested by verifying goodness of fit to the log-transformed power-law model ($R^2 = 0.95$) i.e., $\log P(D_i) = -\gamma \log D_i + \log \eta + \varepsilon_i$ (Barabàsi 2004). Here γ and η are shape and intercept parameters, i is the index of a gene in the network, ε_i is a residual fitting error, D_i is the number of edges (degree) of ith gene and $P(D_i)$ is the corresponding probability.

3.4 Clustering co-expressed genes

Inspired by the Basic Local Alignment Search Tool (BLAST) (Altschul *et al*. 1990), and based on the "guilt-by-association" assumption (Eisen *et al*. 1998), we applied the two-stage screening procedure to cluster co-expressed genes with controlled FDR and MAS. We sought to demo its application in metabolic pathway discovery by "re-discovering" the extensively studied galactose metabolic pathway, which consists of at least three types of genes including transporter genes (GAL2, HXTs etc), enzyme genes (GAL1, GAL7, GAL10 etc) and transcription factor genes (GAL4, GAL80, GAL3 etc). Some other genes are also involved in galactose metabolism but their roles are not entirely clear (Rohde *et al*. 2000, Ideker *et al*. 2001). Therefore, our aims are not only to validate our procedure by rediscovering known co-expressed genes pairs, but also to discover some unknown genes in the pathway.

We selected gene "GAL10" as the "seed gene" which encodes the UDP-glucose-4-epimerase (EC 5.1.3.3) (Fig. 9). We set a relatively stringent criterion ($\alpha = 0.05$, $cormin = 0.6$), and $cormin = 0.6$ is widely used in the literature (e.g. Zhou *et al*. 2002, Farkas *et al*. 2003). We discovered six genes (GAL10, GAL7, GCY1, GAL1, GAL2 and YOR121C) (STable 3). Five of six genes are known to be lying in the pathway as shown in shaded squares in Fig. 9, which leads to a specificity of at least 83%. The sixth gene "YOR121C" is a hypothetical ORF for which no functional annotation is currently available. Our results provide strong motivation to experimentally characterize this gene's biological function. Known transcription factor genes (GAL4 and GAL80) were not discoverable from this microarray experiment as the GAL4 and GAL80 expressions are time shifted and only one time sample was included. The pathways discovered using other "seed genes" in the pathway such as GAL1 and GAL7 gave similar results (STable 4).

4 Discussion

In this paper, we presented a two-stage procedure for screening co-expressed gene pairs that controls both biological and statistical significance of the discovered strength of association. For the discovered co-expressions,

our method also provides an "accuracy" assessment of the strength of association by constructing confidence intervals for the strength of each edge. Indeed, for the typically small sample size microarray data, a simultaneous confidence interval is useful to characterize reliability of the reported strength of association. Correlation thresholding is becoming standard practice in gene co-expression analyses (e.g. Butte and Kohane 2000, Butte *et al*. 2000, Zhou *et al*. 2002, Farkas *et al*. 2003,Lee *et al*. 2004), yet "hard" thresholding lowers the discriminative power of the FDR based test (Fig. 5). Our "soft thresholding " procedure is able to control error rate and maintain discriminative power (Fig. 4). The method requires a tight confidence interval on correlation, which may be difficult to obtain for small sample sizes. However, we have shown that our algorithm provides error rate control at a biologically relevant level with relatively few samples(20 samples for Fig. 3b, Fig. 4).

The algorithm is sufficiently general to be applied to many different correlation measures, e.g. Spearman's or Hotelling's dependency statistics. The algorithm can also be extended to different frameworks such as Gaussian Graphic Models (GGM) in which partial correlation coefficients are used as the dependency measures (Whittaker 1990). Different groups have developed approaches to infer GGM from small sample size microarray data (Wang *et al*. 2003, Schafer and Strimmer 2004, Dobra *et al*. 2004). Schafer and Strimmer recently presented a procedure that is based on the bootstrap estimator of the partial correlation coefficient (Schafer and Strimmer 2004). Most of the pairwise partial correlations discovered by their procedure are very close to zero. On one hand, these ultra weak correlations screened by the FDR based inference procedure are "true correlation" from a pure statistical point of view. On the other hand, the "true correlation" may be caused by a variety of factors other than functional relationship, such as positional and spatial artifacts of gene co-expression along chromosomes Kluger *et al*. 2003. Thus it seems necessary to combine such statistical testing with a "soft" thresholding to achieve high sensitivity and specificity (Fig. 5). This paper has presented such a method to simultaneously minimize the discovered proportion of the functionally irrelevant "true correlations" and maximize that of functionally relevant ones. Our two-stage algorithm has been extended to the GGM framework and implementations are included in our R package "GeneNT" (available from http://www.cran.org).

The scope of application of our statistical analysis is explicitly that of randomly sampled, complete observational data (Dobra *et al*. 2004). In this paper, we are not concerned with developing models of causal gene networks(Dobra *et al*. 2004). This would require a different experimentation and interventation approach to understand directional influences, rather than the simple observational random sampling paradigm adopted here (Dobra *et al*. 2004).

Finally we note that the two-stage procedures can be applied under the independency/positive dependency

or the general dependency assumptions (Benjamini and Hochberg 1995, Benjamini and Yekutieli 2001). The implementation of the general dependency procedure ($\nu = \sum_{\lambda=1}^{\Lambda} \lambda^{-1}$) causes loss of discovery power. The assumption of independence may not be critical in the discovery of relevance networks since biological networks are typically very sparse (Yeung *et al*. 2002).

Acknowledgements

This work was partially supported by grants from the National Institute of Health (EY01115), The Foundation Fighting Blindness, Sramek Foundation and Research to Prevent Blindness.

References

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X., Powell,J.I., Yang,L., Marti,G.E., Moore,T., Hudson,J. Jr., Lu,L., Lewis,D.B., Tibshirani,R., Sherlock,G., Chan,W.C., Greiner,T.C., Weisenburger,D.D., Armitage,J.O., Warnke,R., Levy,R., Wilson,W., Grever,M.R., Byrd,J.C., Botstein,D., Brown,P.O. and Staudt,L.M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.

Altschul,S., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.

Barabàsi,A. 2004. Network biology: understanding the cell's functional organization. *Nat.Rev.Genet.*, **5**, 101-113.

Batagelj,A. and Mrvar,A. 1998. Pajek - Program for large network analysis. *Connections*, **21**, 47-57.

Benjamini,Y. and Hochberg,Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met*, **57**, 289-300.

Benjamini,Y. and Yekutieli,D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, **29**, 1165-1188.

Benjamini,Y. and Yekutieli,D. 2004. False discovery rate adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, **100**, 71-80.

Boutanaev,A., Kalmykova,A., Shevelyov,Y.Y. and Nurminsky,D.I. 2002. Large clusters of co-expressed genes in the *Drosophila* genome. *Nature*, **420**, 666-669.

Bickel,P.J. and Doksum,K.A. 2000. Mathematical statistics: basic ideas and selected topics. 2nd Edition. *Prentice Hall*, Upper Saddle River, NJ, USA.

Butte,A. and Kohane,I.S. 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, **5**, 415-426.

Butte,A., Tamayo,P. Slonim,D., Golub,T.R. and Kohane,I.S. 2000. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci USA*, **97**, 12182-12186.

Butte,A., Bao,L., Reis,B.Y., Watkins,T.W. and Kohane,I.S. 2001. Comparing the similarity of time-series gene expression using signal processing metrics. *J Biomed Inform*, **34**, 396-405.

Daniel, H. 1944. The relation between measures of correlation in the universe of sample permutations. *Biometrika*, **33**, 129-135.

DeRisi,J., Iyer,V., and Brown,P.O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.

Dobra,A., Hans,C., Nevins,R., Yao,G. and West,M. 2004. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90**, 196-212.

Dwight, S.S., Harris, M.A., Dolinski, K., Ball, C.A., Binkley, G., Christie, K.R., Fisk, D.G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D. and J.M. Cherry. 2002. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acid Research*, **30**, 69-72.

Eisen,M., Spellman,P., Brown,P.O., Botstein,D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, **95**, 14863-14868.

Filkov,V., Skiena,S. and Zhi,J. 2002. Methods for analysis of microarray time-series data. *Journal of Computational Biology*, **9**, 317-330.

Farkas,I., Jeong,H., Vicsek.,T., Barabasi,A.L. and Oltvai,Z.N. 2003. The topology of transcription regulatory network in the yeast, *Saccharomyces cerevisiae*. *Physica A*, **318**, 601-612.

Golub,T., Slonim,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Hero,A.O., Fleury,G., Mears,A. and Swaroop,A. 2004. Multicriteria gene screening for analysis of differential expression with DNA microarrays. *EURASIP Journal on Applied Signal Processing*, **1**, 43-52.

Hollander,A. and Wolfe,D. (1999) Nonparametric statistical methods. *Wiley-Interscience*, Hoboken, NJ, USA.

Ideker,T., Thorsson,V., Siegel, A.F. and Hood, L.E. 2000. Testing for differentially expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, **7**, 805-817.

Ideker,T., Thorsson,V., Ranish,J.A., Christmas,R., Buhler,J., Eng,J.K., Bumgarner,R., Goodlett,D.R., Aebersold,R. and Hood,L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929-934.

Jeong,H., Mason,S., Barabasi,A.L. and Oltvai,Z.N. 2001. Lethality and centrality in protein networks. *Nature*, **411**, 41-42.

Kluger,Y., Yu, H., Qian, J., and Gerstein. M. 2003. Relationship between gene co-expression and probe localization on microarray slides. *BMC Genomics*, **4**, 49.

Lee,H., Hsu,A., Sajdak,J., Qin,J. and Pavlidis,P. 2004. Coexpression analysis of human genes across many microarray data sets. *Genome Res*, **14**, 1085-1094.

Lockhart,D., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675-1680.

McLachlan,G., Bean,R. and Peel,D. 2002. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413-422.

Nixon,T., Ronson,C. and Ausubel,F.M. 1986. Two-component regulatory systems responsive to environmental stimuli share strongly conserved domains with the nitrogen assimilation regulatory genes ntrB and ntrC. *Proc Natl Acad Sci USA*, **83**, 7850-7854.

Reiner,A., Yekutieli,D. and Benjamini,Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 386-375.

Rohde,J., Trinh,J. and Sadowski,I. 2000. Multiple signals regulate GAL transcription in yeast. *Mol Cell Biol*, **20**, 3880-3886.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-470.

Schafer,J., and Strimmer,K. 2004. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **1**, 1-13.

Stock,M., Victoria,L. and Goudreau,P.N. 2000. Two-component signal transduction. *Annual Review of Biochemistry*, **69**, 183-215.

Tusher,V., Tibshirani,R. and Chu,G. 2001. Significance analysis of microarrays applied to the the ionizing radiation response. *Proc Natl Acad Sci USA*, **98**, 5116-5121.

Wang,J., Myklebost,O. and Hovig,E. 2003. MGraph: graphical models for microarray data analysis. *Bioinformatics*, **19**, 2210-2211.

Wieczorke,R., Krampe,S., Weierstall,T., Freidel,K., Hollenberg,C.P. and Boles,E. 1999. Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in *Saccharomyces cerevisiae*. *FEBS Lett*, **464**, 123-128.

Whittaker,J. 1990. Graphic models in applied multivariate statistics. *Wiley*, New York, USA.

Yeung,L., Szeto,L., Liew,A.W. and Yan,H. 2004. Dominant spectral component analysis for transcriptional regulations using microarray time-series data. *Bioinformatics*, **20**, 742-749.

Yeung,M., Tegner,J. and Collins,J.J. 2002. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci USA*, **99**, 6163-6168.

Zareparsi,S., Hero,A.O., Zack,D.J., Williams,R. and Swaroop,A. 2004. Seeing the unseen: Microarray-based gene expression profiling in vision. *Invest Ophthalmol Vis Sci.*, **45**, 2457-2462.

Zhou,X., Kao,M. and Wong,W.H. 2002. Transitive functional annotation by shortest path analysis of gene expression data. *Proc Natl Acad Sci USA*, **99**, 12783-12788.

5 Appendix

5.1 Construct PCER-CI for ρ

Based on the fact that z ($z = \tanh^{-1}(\hat{\rho})$) is the monotonic function of $\hat{\rho}$, the asymptotic PCER $(1-\alpha)\times 100\%$ Confidence Interval: $I^{\lambda}(\alpha)$ on each true Pearson correlation coefficient ρ of the set $\mathcal{G}_1$ is: $\tanh(z - \frac{z_{\alpha/2}}{(N-3)^1/2}) \leq \rho \leq (z + \frac{z_{\alpha/2}}{(N-3)^1/2})$, where $P(N(0,1) > z_{\alpha/2}) = \alpha/2$.

5.2 Construct PCER-CI for τ

The asymptotic PCER $(1-\alpha)\times 100\%$ Confidence Interval: $I^{\lambda}(\alpha)$ on each true Kendall correlation coefficient τ of the set $\mathcal{G}_1$ is constructed as follows:

- Compute $C_r = \sum_{\substack{t=1 \\ t\neq r}}^{N} Q((X_r,Y_r),(X_t,Y_t))$, for $r = 1,2,...,N.$, where $Q((a,b),(c,d))$ is given by:

$$Q((a,b),(c,d)) = \begin{cases} 1 & \text{if } (d-b)(c-a) > 0, \\ 0 & \text{if } (d-b)(c-a) = 0, \\ -1 & \text{if } (d-b)(c-a) < 0. \end{cases} \tag{10}$$

- Let $\bar{C} = \frac{1}{N}\sum_{r=1}^{N} C_r$ and define $\hat{\sigma}_\tau = \frac{2}{N(N-1)}\frac{2(N-2)}{N(N-1)}\sum_{i=1}^{N}$
  $(C_r - \bar{C})^2 + 1 - \hat{\tau}^2]$

- $I^{\lambda}(\alpha) : \hat{\tau} - z_{\alpha/2}\hat{\sigma}_\tau \leq \tau \leq \hat{\tau} + z_{\alpha/2}\hat{\sigma}_\tau$.

5.3 Simulation of pairwise vectors based on pre-specified population covariances

5.3.1 Pearson correlation coefficient ρ

.

- Specify a covariance matrix $\mathbf{V}$ and a mean vector μ.
- Form the Cholesky decomposition of $\mathbf{V}$, i.e. find the lower triangular matrix L such that $\mathbf{V} = LL^T$.
- Simulate a vector $\mathbf{z}$ with independent $N(0,1)$ elements.
- A vector simulated from the required multivariate normal distribution is then given by $\mu + L\mathbf{z}$.

5.3.2 Kendall's τ

.

- Specify a value for τ.
- Simulate an $N \times N$ indicator matrix M given τ as follows:

$$M[n,m]_{1\leq n<m\leq N} = \begin{cases} 1 & \text{if Bernulli}(\frac{1+\tau}{2}) \text{ is TRUE,} \\ -1 & \text{if Otherwise.} \end{cases} \tag{11}$$

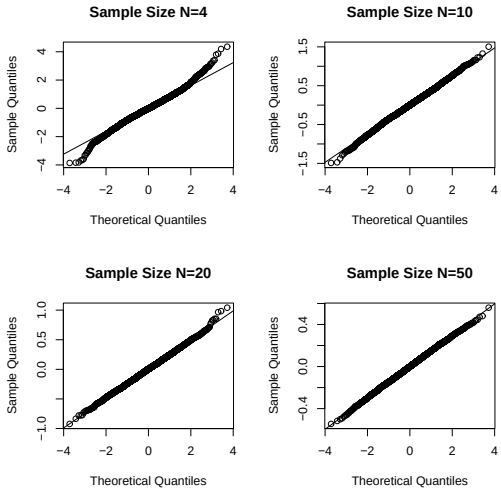
- Simulate i.i.d pairs (X_r, Y_r) $(r = 1,2,...,N)$ according to M matrix and definition

$$Q((a,b),(c,d)) = \begin{cases} 1 & \text{if } (d-b)(c-a) > 0, \\ -1 & \text{if } (d-b)(c-a) < 0. \end{cases} \tag{12}$$

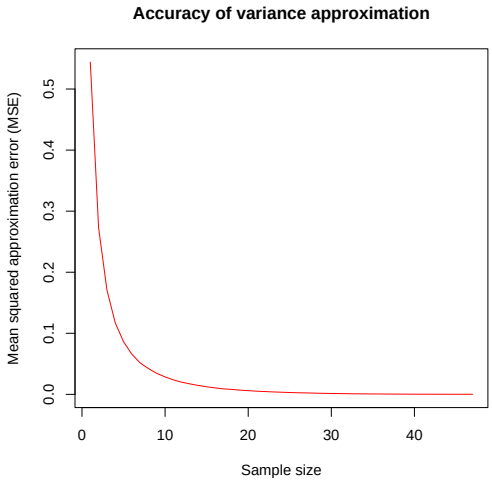
No tied observations are generated. Alternatively, $\hat{\tau}$ can be directly calculated from the indicator matrix M without generating the i.i.d pairs (eq. 3).

Table 1: Top ten "hub genes". The rank of each gene is the average rank over five different networks. Each of five networks is constrained by a different pair of (FDR,MAS) criteria. The highest ranked gene is the most connected and stable gene under varying constraints of (FDR,MAS).

Gene Name	Average Rank
RPL42B	4.2
RPS16B	6.2
RPL14A	7.4
RPS3	7.4
GTT2	8.0
RPS4A	9.8
RPL33A	11.6
RPL23B	15.4
RPS7A	15.8
RPS4B	17.2

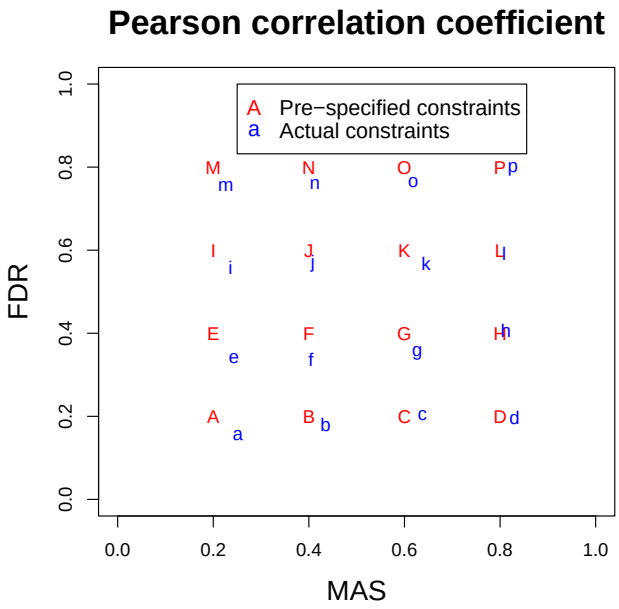


(a)

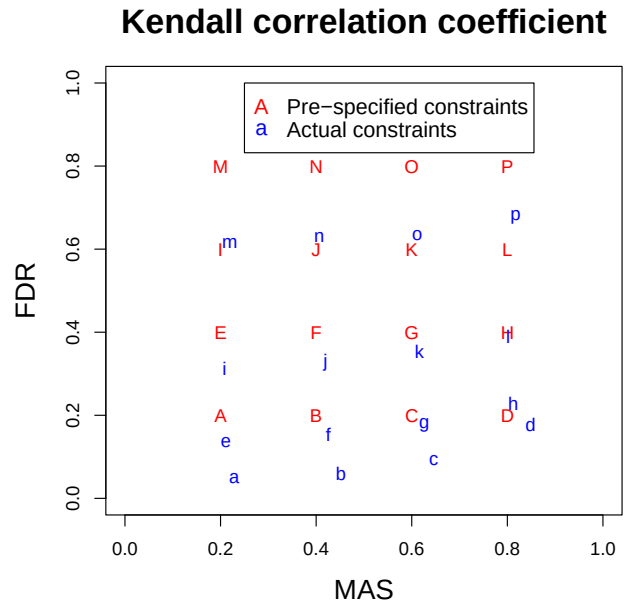


(b)

Figure 3: Verification of Gaussian null sampling distribution and variance approximation for Pearson correlation coefficient (eq. 8). (a) *QQ* plot of transformed sampling distribution of Pearson correlation coefficient $\hat{\rho}$ versus Gaussian distribution. (b) Mean squared approximation errors (MSE) of the variances of transformed sample Pearson correlation coefficients $\hat{\rho}$.



(a)



(b)

Figure 4: Verification of two-stage error control procedure based on Pearson correlation coefficient (a) and Kendall correlation coefficient (b). Sample size $N = 20$.

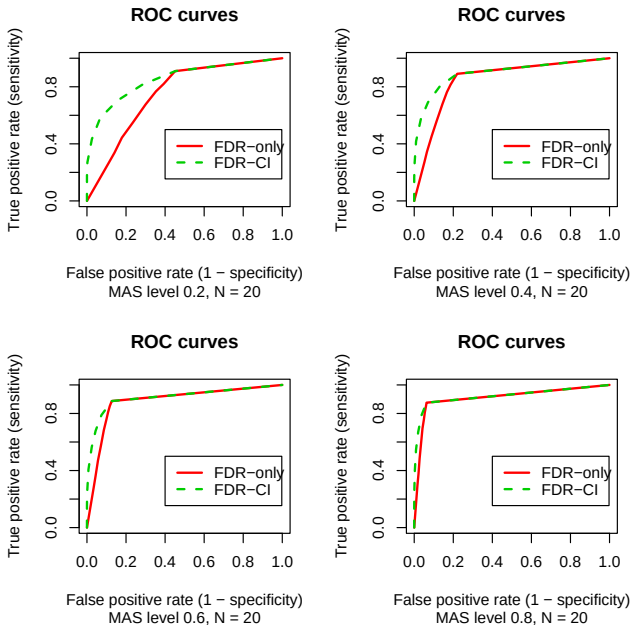
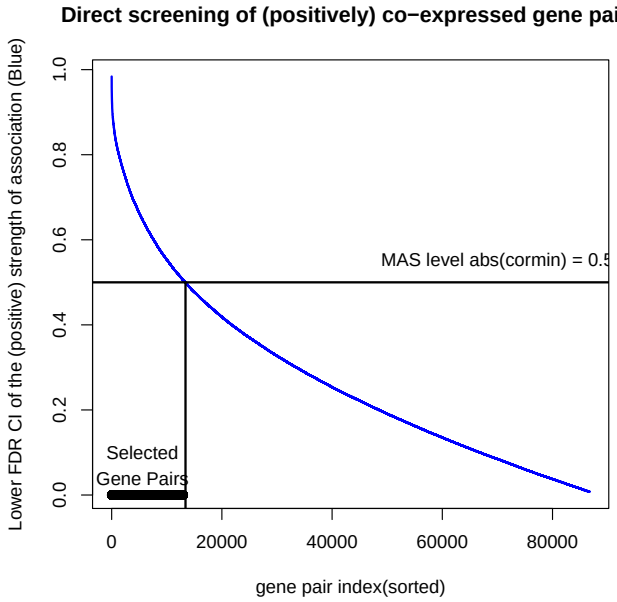
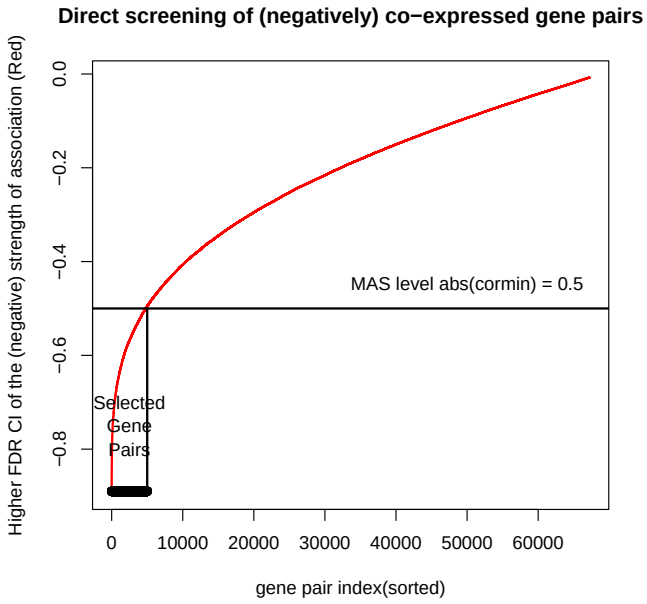


Figure 5: ROC curves of "FDR-CI" test procedure and "FDR-only" test procedure based on Pearson correlation statistic

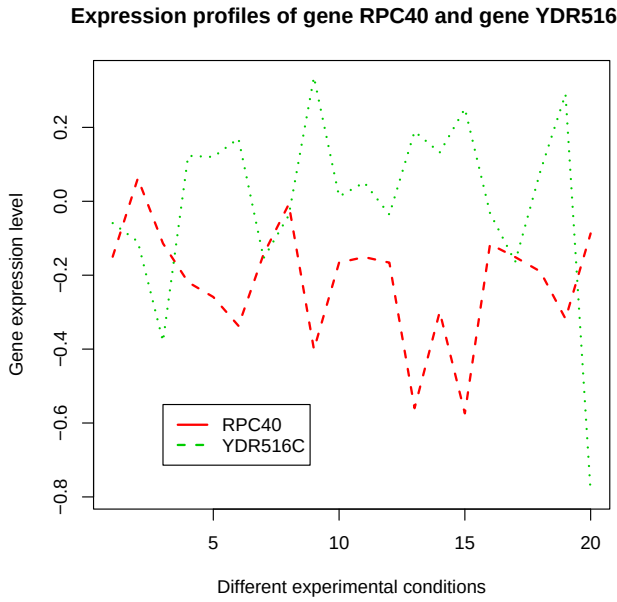


(a)

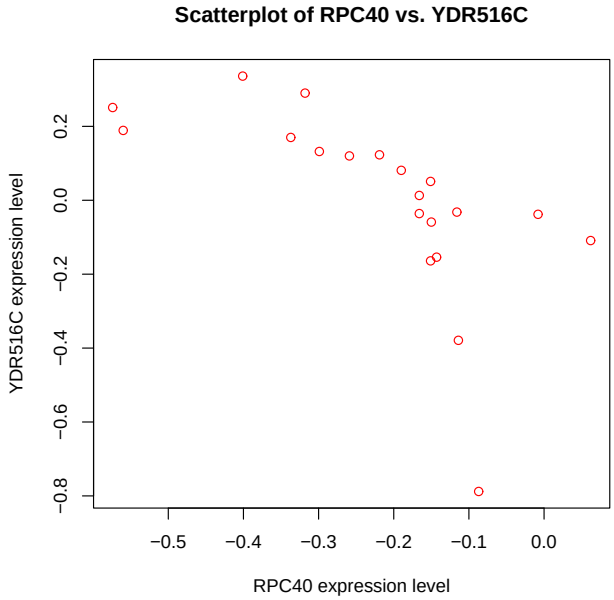


(b)

Figure 6: Curves specify lower endpoints (a) and upper endpoints (b) of the 5% FDR-CI's on the positive Pearson correlation coefficients (a) and negative Pearson correlation coefficients (b) for the galactose metabolism study. Only those gene pairs whose FDR-CI's do not intersect $[-cormin, cormin]$ are selected by the second stage of screening. When the MAS strength of association criterion is $cormin = 0.5$, these gene pairs are obtained by thresholding the curves as indicated.



(a)



(b)

Figure 7: A pair of non-linearly correlated genes.

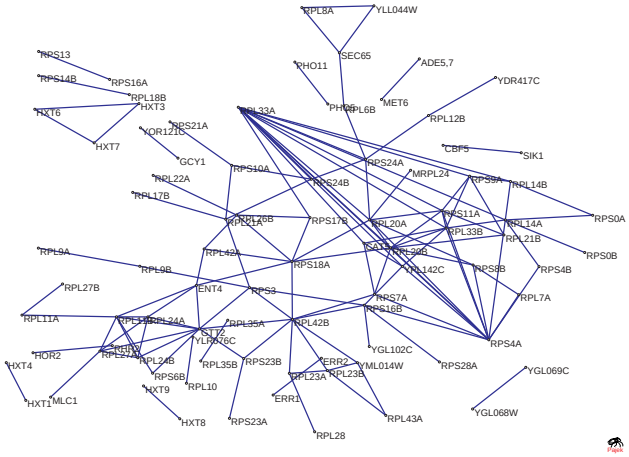


Figure 8: Network topology visualization. The network is discovered by constraining FDR $\leq 5\%$ at a MAS level of 0.9. No significant negative correlation is discovered at this level. The graph is drawn using Pajek (Batagelj and Mrvar 1998).

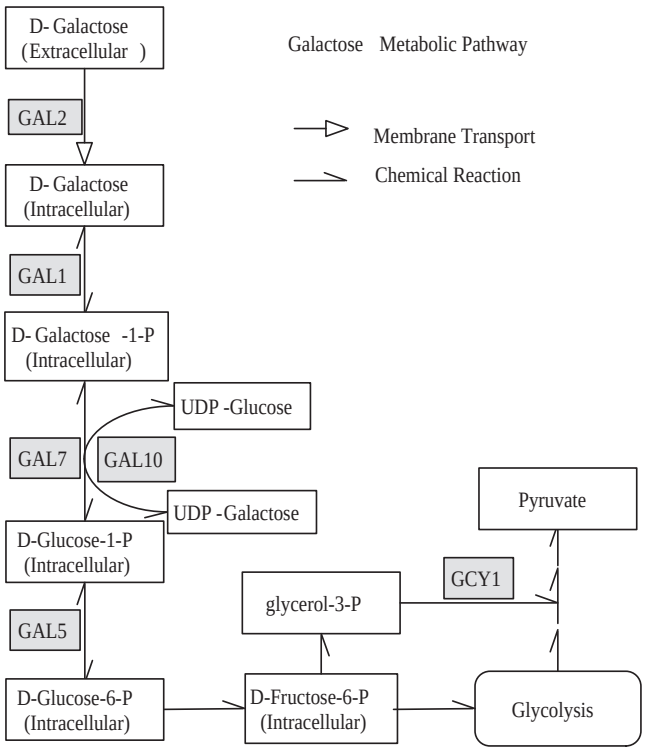


Figure 9: Diagram of the structural module of the galactose metabolic pathway. The shaded boxes denote the five out of six genes whose gene products lie in the galactose metabolic pathway "rediscovered" by our algorithm.