Vol. 00 no. 00 2004
Pages 1–10

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

Dongxiao Zhu[a,b,*], Alfred O Hero[b] and Zhaohui S Qin[c]

[a]Bioinformatics Program, University of Michigan, Ann Arbor, MI 48109 [b]Departments of EECS, Biomedical Engineering and Statistics, University of Michigan, Ann Arbor, MI 48105 [c]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109

ABSTRACT

Motivation: Many exploratory microarray data analysis tools such as gene clustering and relevance networks rely on detecting pairwise gene co-expressions. 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 experimentally undetectable co-expressions.

Methods: 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 (Minimium Acceptable Strength,MAS) of the discovered co-expressions. Based on the estimation of pairwise gene profile correlation, the algorithm provides an initial co-expression discovery that controls only FDR, which is then followed by a second 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.

Results: We validated asymptotic null distributions of Pearson and Kendall correlation coefficients and the two-stage error-control procedure using simulated data. We then used yeast galactose metabolism data (Ideker et al. 2001) to illustrate the advantage of our method for clustering genes and constructing a relevance network. In gene clustering, the algorithm screens a seeded cluster of co-expressed genes with controlled FDR and MAS. In constructing the relevance network, the algorithm discovers a set of edges with controlled FDR and MAS.

Availability: The method has been implemented in an R package "GeneNT" that is freely available from: http://www-personal.umich.edu/~zhud .

Contact: zhud@umich.edu

Supplementary Information: Supplemental material can be found at: http://www-personal.umich.edu/~zhud

1 INTRODUCTION

The emergence and development of DNA miroarray 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 *et al*., 2000,Farkas *et al*., 2003) are two important explorative 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. Furthermore, discovering co-expressed gene pairs in lower eukaryote addresses gene functional prediction directly because co-expressed genes are known to be often co-expressed in pairs (Boutanaev *et al*., 2002).

Clearly, there is a demand 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 pondered in developing such a methodology, namely, which statistic(s) to use and what screening procedure to choose.

Several methods have been adopted to measure the strength of association between the expression profiles of gene pairs, such as: Euclidean distance (Tamayo *et al*., 1999), Pearson

[*]to whom correspondence should be addressed

correlation coefficient (Zhou *et al.*, 2002), coherence(Butte *et al.*, 2001), mutual information (Butte *et al.*, 2000), edge detection(Filkov *et al.*, 2002), and dominant spectral component analysis (Yeung *et al.*, 2004). Each of these methods have advantages and disadvantages. To select co-expressed gene pairs, the common practice is to calculate a sample correlation for each pair of gene and then to select the top pairs by correlation thresholding (Butte *et al.*, 2000,Zhou *et al.*, 2002, and Farkas *et al.*, 2003). This approach controls biological significance by screening only strongly correlated pairs. 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 only statistical significance: screen co-expressed gene pairs whose strength of association is different from zero using p-value thresholding. This approach does not control biological significance and can lead to screening-in some weakly correlated gene pairs that are difficult to verify by follow-up experiments such as real time RT-PCR.

Regarding which statistic(s) to use, the Pearson correlation coefficient has been one of the most popular choices because it is easy to calculate and its performance is comparable to more complex and computational intense methods (Yeung *et al.*, 2004, Kwon *et al.*, 2003). 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. We decided to explore the Pearson and Kendall correlation coefficient measures because their asymptotic distributions are known, as required by our two-stage screening procedure.

Regarding what screen procedure to choose, a two-stage statistical hypothesis testing scheme is applied in order to decide on whether the strength of association is statistically significant at the pre-specified MAS level. The test is non-standard because: 1) MAS is ordinarily greater than 0; 2) many comparisons have to be tested simultaneously. Our method is directly inspired by the two-stage screen methodology (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 mutants and one wild type strains incubated in either GAL-inducing or non-inducing media) with four replicates in each condition.

The outline of the paper is as follows. In section 2, we describe the proposed two-stage multicriteria approach. In section 3, we first show the approach indeed controls FDR at the pre-specified MAS level using synthetic data, and then illustrate it for yeast galactose metabolism data. In section 4, we discuss advantages of our method, model assumptions and restrictions.

2 METHODS

2.1 Measures of the strength of association

There are many possible discriminates for strength of association between two variables (generally denoted as 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 and Kendall correlation coefficients are special cases of the generalized correlation coefficient (Daniel, 1944), and are discussed below. 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:

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

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

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

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 $\overline{X_{g_i}} = N^{-1} \sum X_{g_i(n)}$, $\overline{X_{g_j}} = N^{-1} \sum X_{g_j(n)}$ are the 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: $\hat{\tau}_{i,j} = 2\sum\sum_{1\leq n\leq m\leq N} \frac{K_{nm}}{N(N-1)}$. Here K_{nm} is a indicator variable defined as $K_{nm} = \mathrm{sgn}(x_{g_{i(n)}} - x_{g_{i(m)}})\mathrm{sgn}(x_{g_{j(n)}} - x_{g_{j(m)}})$ for each set of pairs drawn from $\{X_{g_i}\}_{i=1}^{G}$ and $\{X_{g_j}\}_{j=1}^{G}$.

To make the estimated correlation robust against spurious outliers yet sensitive to strong similarities in expression patterns, we adopted a leave-one-out cross-validation technique, using the median estimate as a robust estimator of the correlation.

2.2 Hypothesis testing scheme

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

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

where $cormin$ is a the specified minimium acceptable strength of correlation. The sample correlation coefficient $\hat{\Gamma}_{i,j}$ ($\hat{\rho}_{i,j}$ or $\hat{\tau}_{i,j}$) could be used as a decision statistic 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 a random sample, 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 the sample could have been drawn from the population(s) being tested (or that a more improbable sample could be drawn) 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:

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

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

whrere Φ is the standard Gaussian cumulative density function, and $K = \sum\sum_{1\leq n\leq m\leq N} K_{nm}$. The above expressions are based on asymptotic Gaussian approximations (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 $\mathcal{G}$ 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 $\mathcal{G}$ 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, we adopt the FDR to control statistical significance of the selected gene pair correlations in our screening procedure (Hero *et al.*, 2004).

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

1. Specify FDR level α and MAS level $cormin$.
2. For each of $\mathcal{G}$ gene pairs, compute a list of PCER p-values: $p_1, p_2, ..., p_{\mathcal{G}}$ from $\{\hat{\rho}_{i,j}\}$ or $\{\hat{\tau}_{i,j}\}$.
3. Sort the list of PCER p-values in increasing order, i.e. $p_{(1)}, p_{(2)}, ..., p_{(\mathcal{G})}$.
4. Find the index k where $k = max\{k : p_k \leq \frac{k\alpha}{G\nu}\}$.
5. Set initial screening G_1 as those k gene pairs having p-values: $p_{(1)}, p_{(2)}, ..., p_{(k)}$.

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

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

1. Construct k diferent $(1-\alpha)\times 100\%$ PCER-CI's for ρ or τ of each gene pairs in G_1(Appendix 5.1).
2. Convert these PCER-CI's into k different $(1-\alpha)\times 100\%$ FDR-CI's using formula (Benjamini and Yekutieli, 2004): $I^g(\alpha) \rightarrow I^g(G_1\alpha/\mathcal{G})$.
3. Select the subset of G_2 of G_1 gene pairs whose FDR-CI's do not intersect $[-cormin, cormin]$.

Fig. 1. Two-stage direct screening algorithm.

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:

Stage I. Test the simple null hypothesis:

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

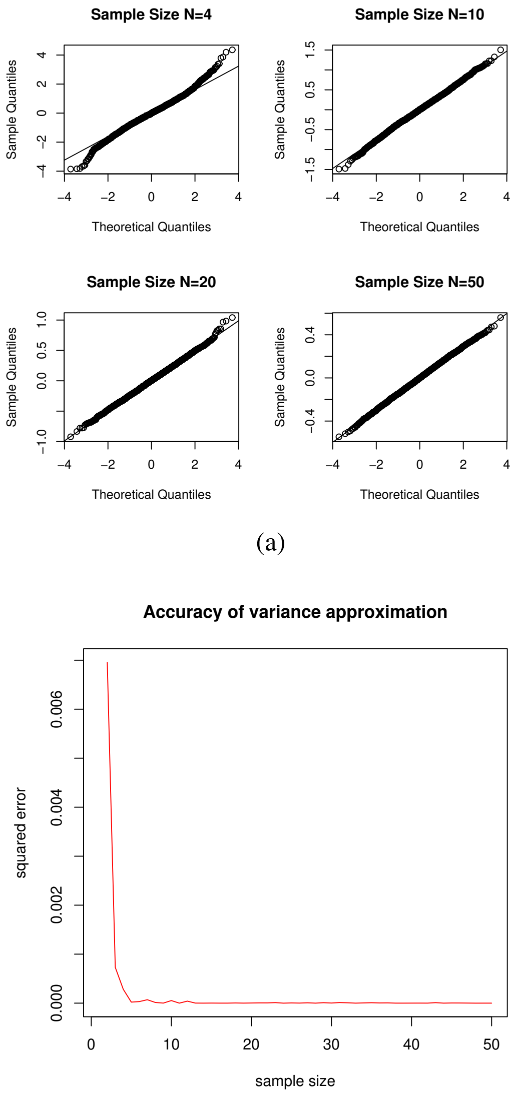
at FDR level α. The step-down procedure of Benjamini and Hochberg (Benjamini and Hochberg, 1995) is used. There are three ways of adjusting error rate: the single-step, step-down and step-up procedures. In single-step procedures, equivalent multiplicity adjustments are performed for all hypotheses, regardless of the distribution of PCER p-values giving strong control of FWER. Improvement in power, while preserving Type I error control, may be achieved by step-up and step-down procedures, in which rejection of a particular hypothesis is based not only on the total number of hypotheses, but also on the distribution of PCER p-values. Step-down procedures order the PCER p-values starting with the most significant, i.e. the smallest, while step-up procedures start with the least significant (Speed, T. ed).

Stage II. Suppose G_1 pairs of genes pass the stage I procedure. In stage II, we first construct asymptotic PCER Confidence Intervals (PCER-CI's) :$I^g(\alpha)$ for each Γ (ρ or τ) in subset G_1, and convert into FDR Confidence Intervals(FDR-CI's) :$I^g(G_1\alpha/\mathcal{G})$ (Benjamini and Yekutieli, 2004). A gene pair in subset 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).

```
1. For each $\mathcal{G}$ gene pairs, compute a list of PCER $p$-values: $p_1, p_2, ..., p_{\mathcal{G}}$ using $\{\hat{\rho}_{i,j}\}$ or $\{\hat{\tau}_{i,j}\}$.
2. Sort the list of PCER $p$-values in the increasing order, i.e. $p_{(1)}, p_{(2)}, ..., p_{(\mathcal{G})}$.
for each gene pairs denoted as $\mathcal{G}_0$: $\mathcal{G}_0 \in \{x_{g_{i(n)}}, x_{g_{j(n)}}\}_{n=1}^{N}$,
    • Find the minimal $\alpha$ such that the PCER-CI does not intersect $[-cormin, cormin]$.
    • Compute the integer index $N(\alpha(\mathcal{G}_0)) = \sum_{k=1}^{\mathcal{G}} I(p(g_{(k)})k \leq \alpha(\mathcal{G}_0))$, where $I(A)$ is an indicator function.
      $I(A) = \begin{cases} 1 & \text{if } p(g_{(k)})k \leq \alpha(\mathcal{G}_0) \text{ is TRUE,} \\ 0 & \text{if Otherwise.} \end{cases}$    (3)
      The FDR $p$-value of the gene pair $\mathcal{G}_0$ is then simply $p_{(}g_i)$, where $i = N(\alpha(\mathcal{G}_0))$.
endfor
```

Fig. 2. Inverse screening algorithm.

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's the most stringent pair of criteria (α , $cormin$) would cause a particular subset of gene pairs to be declared as dependent. The inverse screening procedure is displayed in Fig 2.



(a)

(a)

Fig. 3. Verification of Gaussian null sampling distribution (a) and variance approximation (b). (a) QQ plot of transformed sampling distribution of Pearson correlation coefficient $\hat{\rho}$ versus Gaussian distribution. (b) Variance approximation of transformed sampling distribution of Pearson correlation coefficient $\hat{\rho}$.

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, 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 variance approximation is vital, which can be accessed by calculating squared error:($s.e.$ denotes standard error, and F_X denotes sampling distribution)

$$\hat{\sigma}^2_\rho = (s.e.(\tanh^{-1}(F_{\hat{\rho}})) - (N-3)^{-1/2})^2$$

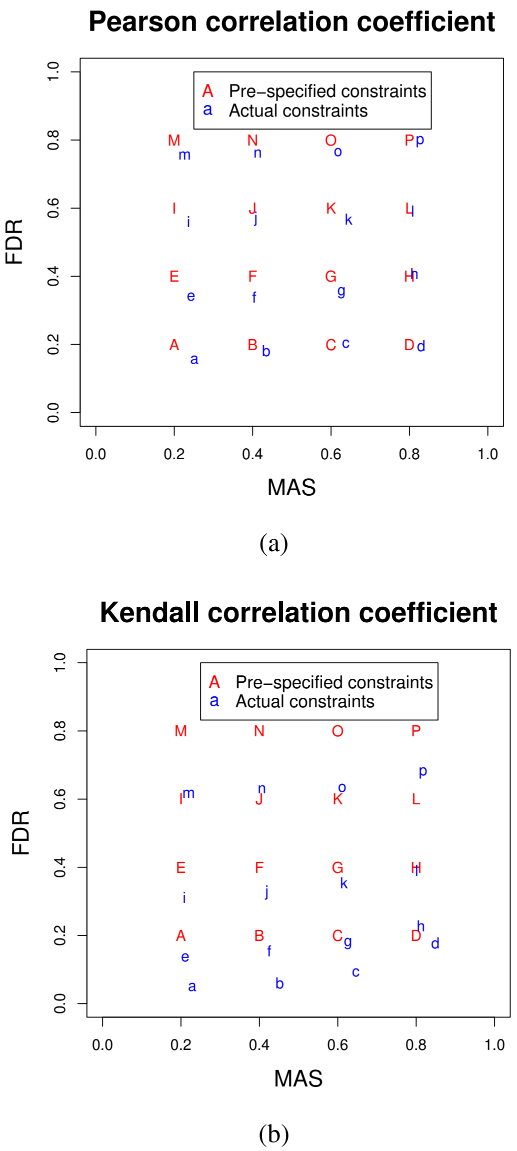
$$\hat{\sigma}_\tau = (s.e.(F_{\hat{\tau}}) - (\frac{2}{N(N-1)}\frac{2(N-2)}{N(N-1)^2}\sum_{i=1}^{N}(C_i - \overline{C}) + 1 - \hat{\tau}))^2$$

where the definition of C_i 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 ($N > 6$) from Fig 3b.

3.1.2 Validating error control procedure. In order to validate our FDR and MAS error control procedure, we simulated pairwise gene expression data based on pre-specified 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 minimium true discovery of population correlation $\Gamma_{i,j}$ among the screened pairs. We pre-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 capital 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 small English alphabet (Blue). It can be observed that the actual FDR's (small alphabets) fall below the pre-specified constraint (capital alphabets) and the actual MAS's (small alphabets) fall above the pre-specified constraint (capital alphabets). The deviations of actual FDR's and MAS's from their pre-specified levels are due to the conservative asymptotic approximation. This will translate into a reduction of power in discovering co-expressed pairs at the specified levels.

3.2 Constructing a relevance network with controlled FDR and MAS

For the yeast galactose metabolism dataset, a subset of 997 genes were identified by Ideker et al using generalized likelihood ratio test (Ideker *et al.*, 2000). Genes having a likelihood statistic $\lambda \leq 45$ were selected as differentially expressed, whose mRNA levels differed significantly from reference under one or more perturbations. We used the average expression profiles over four replicates for subsequent analysis, which implicitly assumes that the between-replicates variances for a gene over different experimental conditions are equal.

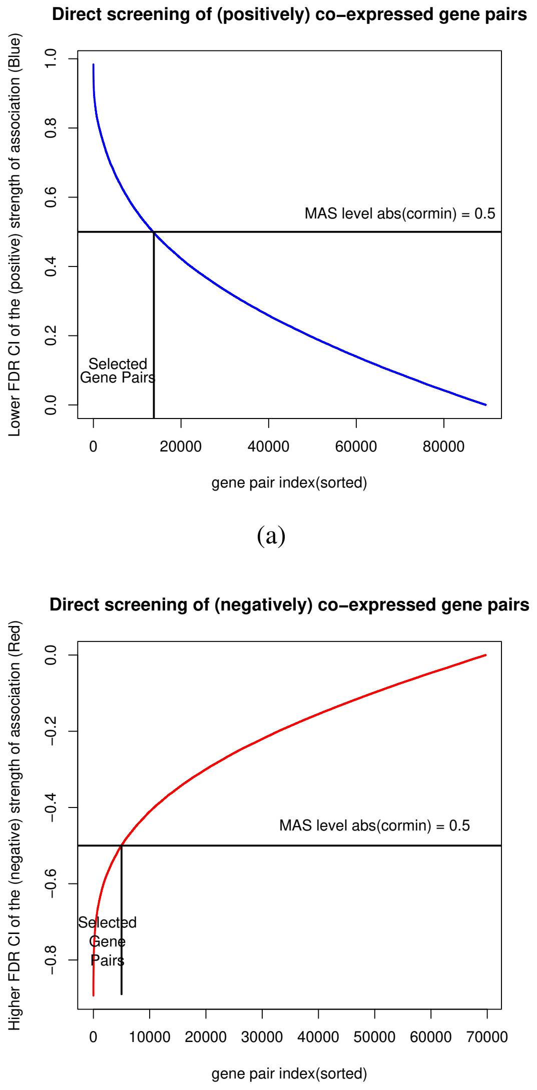


(a)

(b)

Fig. 4. Verification of two-stage error control procedure based on Pearson correlation coefficient (a) and Kendall correlation coefficient (b).

Fig 5a and Fig5b illustrate the direct implementation of the two-stage procedure to screen positively or negatively correlated gene pairs based on the Pearson correlation coefficient.

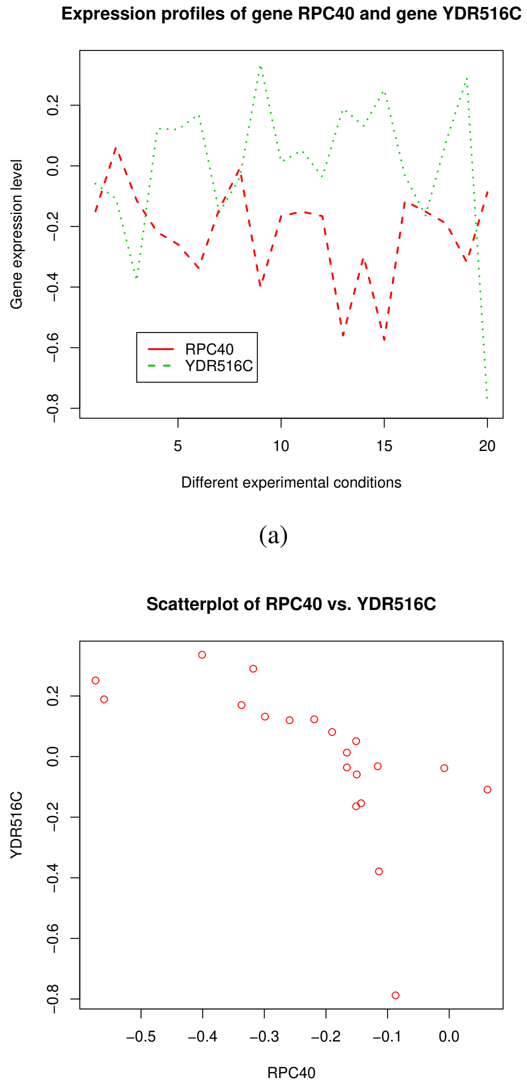


(a)

(b)

Fig. 5. Segments of lower bounds (a) and upper bounds (b) specifying 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.

The direct screening procedure is constrained by FDR criterion $\alpha = 0.05$ and MAS criterion $cormin = 0.5$. There were 159,287 out of 496,506 gene pairs having FDR ≤ 0.05, leaving 159,287 correlation coefficients for which FDR-CIs are constructed. A gene pair passes the the second stage screening if the FDR-CI does not intersect the interval $[-0.5, 0.5]$. 18,594 gene pairs are declared to be both ”biologically” and ”statistically” significant. Similarly, using Kendall correlation coefficient, there were 95,205 gene pairs that passed the stage I screen, and only 3,552 gene pairs passed the stage II screen constrained by the same MAS and FDR criteria (STable 1).



(a)

(b)

Fig. 6. A pair of non-linearly correlated genes.

Although for Gaussian distributed pairs the Kendall rank correlation coefficient has lower discovery power compared to Pearson correlation coefficient, it nevertheless is able to pull out many non-linearly correlated gene pairs that are missed by the Pearson correlation. For example, the link between gene ”RPC40” and gene ”YDR516C” passed both stage I and II screening ($\alpha = 0.01$, $cormin = 0.5$)

when using Kendall correlation coefficient ($\hat{\tau}$ =-0.7513333, FDR p-value = 0.0006150649, FDR-CI = [-0.9663466, -0.5363199]), but they failed to pass even the first screening using Pearson correlation coefficient $\hat{\rho}$ =-0.6263346 (FDR p-value = 0.01221224). From the scatter plot, we can observe the obvious non-linear correlation (Fig 6). The poor linear fit can be verified by fitting a simple linear regression model and observing $R^2 = 0.36$ (R^2 is the goodness of fit).

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). Therefore, our constructed networks are mixed networks in which edges are discovered using either Pearson correlation coefficient or Kendall correlation coefficient 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 7. The hub gene is the highly connected gene that dominates the network topology and is minimally sensitive to the network discovery criteria. We constructed five networks constraint by five pairs of constraints (FDR $\leq$ 0.05, $cormin = 0.5, 0.6, 0.7, 0.8, 0.9$) using both Pearson correlation coefficient and Kendall correlation coefficient. 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 which 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, and different types of proteins including transporters, enzymes, and regulators have to be synthesized upon induction (Wieczorke *et al.*, 1999). We ranked the ”hub genes” over five networks by calculating and sorting average rank of each ”hub gene” (Table 1, STable 2). Interestingly, the list of ”hub genes” contains many hypothetical Open Reading Frames (ORFs)(STable 2), which are presumably indispensable for galactose metabolism (Jeong *et al.*, 2001).

Fig 7 presents the discovered network topology with a FDR level of 0.05 (5% discovered edges are expected to be false positive) at the MAS level of $cormin = 0.9$. The network is composed of 91 connected vertices and 138 edges. Similiar to some other biological networks, the network marginal degrees appear power-law distributed, which is tested by verifying goodness of fit to the log-transformed power-law model ($R^2 = 0.95$) i.e., $\log P(K_i) = -\gamma \log K_i + \log \eta + \varepsilon_i$, $(i = 1, 2, , n)$. Here γ and η are shape and intercept parameters, ε_i is a residual fitting error. K_i is the degree and $P(K_i)$ is the corresponding probability.

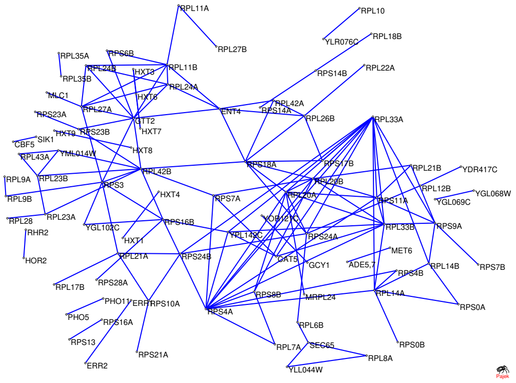


Fig. 7. Network topology visualization. The network is discovered by constraining FDR $\leq$ 5% at a MAS level of 0.9. No significant negative correlation is discovered at this level. The graph is drawn using Pajek (Batagelj and Mrvar, 1998).

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 most stable gene under varying constraints of (FDR,MAS).

Gene Name	Average Rank
RPL42B	4.2
RPS3	5.8
RPL14A	7.0
RPS16B	7.6
GTT2	8.4
RPS4A	9.8
RPL33A	11.8
RPL23B	15.8
RPS7A	16
RPL27A	17.4

3.3 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, we applied the two-stage algorithm to cluster co-expressed genes with controlled FDR and MAS, and $vice\ versa$. We sought to demo its application in metabolic pathway discovery by ”rediscovering” 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 ”rediscover” the known genes but also to discover some unknown genes in the pathway.

We select gene ”GAL7” as the ”seed gene” which encodes the UDP-glucose-hexose-1-phosphate uridylyltransferase (EC 2.7.7.12). The enzyme catalyzes the transformation of Galactose-1-P into Glucose-1-P, and the latter enters the glycolysis pathway through relocating the phosphate group. Many genes lying in the galactose metabolic pathway are ”rediscovered” by our technique under the relative stringent criteria ($\alpha = 0.05, cormin = 0.2$) (Fig 8). Transcription factor genes (GAL4 and GAL80) are not ”rediscovered” together with transporter genes and enzyme genes because the experiment could not capture time-delayed co-expressions simultaneously. The algorithm also discovers some unknown genes that are hypothesized to be relevant to galactose metabolism (STable 3). The pathway ”rediscovery” based on other ”seed genes” in the pathway such as GAL1 and GAL10 gave similar results (STable 4).

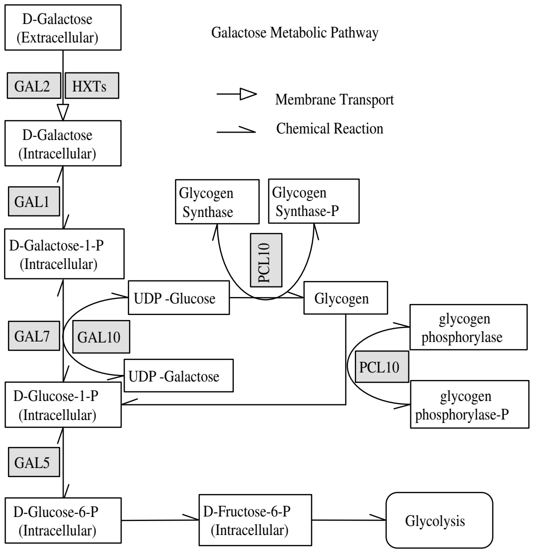

D-Galactose (Extracellular)
Galactose Metabolic Pathway
GAL2
HXTs
Membrane Transport
Chemical Reaction
D-Galactose (Intracellular)
GAL1
Glycogen Synthase
Glycogen Synthase-P
PCL10
D-Galactose-1-P (Intracellular)
UDP -Glucose
Glycogen
GAL7
GAL10
glycogen phosphorylase
UDP -Galactose
PCL10
D-Glucose-1-P (Intracellular)
glycogen phosphorylase-P
GAL5
D-Glucose-6-P (Intracellular)
D-Fructose-6-P (Intracellular)
Glycolysis

Fig. 8. Diagram of up-to-date galactose metabolic pathway. The shaded squares denote the genes whose gene products lie in the galactose metabolic pathway “rediscovered” by our algorithm.

3.4 Performance comparison

In Table 2 and Table 3, we compare the performance of the proposed screening algorithm, labeled ‘Two-stage FDR-CI,” with two other commonly used algorithms, called ‘(Thresholded) FDR” and ‘(Thresholded) MAS”. All three algorithms aim to control MAS at a level of $cormin = 0.5$. The ‘Two-stage FDR-CI” and “ Thresholded FDR” algorithms aim to control FDR at a level of $\alpha = 0.05$ in addition to MAS. Both of these latter algorithms were implemented as two-stage algorithms with common stage I, which is to select pairs of genes G_1 that pass the test of association with $cormin = 0$ at a FDR level of 5%. The second stage of the ‘Two-stage FDR-CI” algorithm selects G_2 as a subset of G_1 at the pre-specified FDR-CI level of 5%. Stage 2 of the ‘Thresholded FDR” algorithm simply selects a subset of G_1 having a strength of association greater than 0.5. The single-stage ‘Thresholded MAS” algorithm selects a subset of the original 496,506 gene pairs by setting the threshold $\hat{\Gamma}_{i,j} \geq 0.5$.

The number of screened and discovered gene pairs for the three algorithms is indicated in the first two columns of Table 2 and Table 3. The maximum and median of the FDR p-values of the discovered gene pairs are indicated in the third and fourth columns for each algorithm. The last column indicates the average length of the FDR-CI’s on correlation coefficients of the discovered gene pairs. We conclude from Table 2 and Table 3 that the proposed ‘Two-stage FDR-CI” algorithm outperforms the other algorithms in terms of (1) maintaining the FDR requirement that false positives not exceed 5%(column 4); (2) ensuring a substantially lower median FDR p-value than the others (column 5); (3) discovering genes that have tighter (on the average) confidence intervals on biologically significant (e.g. > 0.5) correlation coefficients (column 6).

Table 2. Performance comparison for three algorithms based on Pearson correlation coefficient for selecting gene pairs with a MAS level of 0.5. (Thresholded) MAS and (Thresholded) FDR are significantly worse in terms of statistical significance (p-value) than the proposed (Two-stage) FDR-CI algorithm (columns 4 and 5). Furthermore, the average length of the CIs on ρ or τ of the discovered gene pairs are shorter for the (Two Stage) FDR-CI algorithm than for the other algorithms (column 6).

	# Screened	# Discovered	Max(Pv)	Meidan(Pv)	AvgFDRCI
MAS	496,506	97,786	0.06659022	0.006930025	0.717358
FDR	159,287	89,554	0.04999962	0.005163348	0.6179784
FDR-CI	159,287	18,594	2.55205e-07	9.304143e-09	0.3282658

4 DISCUSSION

In this paper, we presented a two-stage algorithm for screening co-expressed gene pairs that controls both biological and statistical significance. For those discovered co-expressions, our method also provides an ‘accuracy” assessment of the strength of association by constructing FDR-CI’s for the

Table 3. Performance comparison for three algorithms based on Kendall's τ for selecting gene pairs with a MAS level of 0.5.

	# Screened	# Discovered	Max(Pv)	Meidan(Pv)	AvgFDRCI
MAS	496,506	31,151	0.01955337	0.006432614	0.6309374
FDR	95,205	31,151	0.01955337	0.006432614	0.6309374
FDR-CI	95,205	3,552	0.001410414	0000431815	0.4051204

strength of each edge. Indeed, for the typically small sample size microarray data, a simultaneous confidence interval is necessary in addition to characterize reliability of the reported strength of association. We illustrated two potential applications of our algorithm to discovering relevance network and to clustering genes, in which the algorithm provides the error rate control at a biological detectable level.

The algorithm is sufficiently general to be applied to many different correlation measures (e.g. Pearson and Kendall correlation coefficients in accordance to the Gaussian microarray data and non-Gaussian microarray data) and hence to be extended to different frameworks such as Gaussian Graphic Models (GGM) in which partial correlation coefficient is used as the correlation measure (Whittaker, 1990). Different groups have developed different appraches to infer GGM from small sample size microarray data (Wang *et al.*, 2003,Schfer and Strimmer, 2004, Adrian *et al.*, 2004). Schafer and Strimmer recently presented a procedure that is based on the bootstrap estimator of the partial correlation coefficient (Schfer and Strimmer, 2004). Our two-stage algorithm has been extended to the GGM framework to control biological significance in addition to statistical significance, and the implementations are included in the R package ”GeneNT” (availible from http://www-personal.umich.edu/ zhud).

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

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 general dependency procedure ($\nu = \frac{1}{\sum_{k=1}^{\mathcal{G}} k^{-1}}$) 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).

REFERENCES

Adrian, D., Chris, H. et.al.(2004) Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*,**90**, 196-212.

Alizadeh, A., Eisen, M. et.al.(2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*,**403**, 503-11.

Altschul, S., Gish, W. et.al. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.

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

Batagelj, A., Mrvar, A.(1998) Pajek - Program for Large Network Analysis. *Connections*, **21**, 47-57.

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

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

Benjamini, Y., Yekutieli, D. (2004) False Discovery Rate adjusted multiple confidence intervals for selected parameters. Submitted to *Journal of American Statistical Association*.

Boutanaev, A., Kalmykova, A. et al. (2001) Large clusters of co-expressed genes in the Drosophila genome. *Nature*, **420**, 666-9.

Butte, A., Tamayo, P. et al. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci*, **97**, 12182-6.

Butte, A., Bao, L. et al. (2001) Comparing the similarity of time-series gene expression using signal processing metrics. *J Biomed Inform*, **34**, 396-405.

Christie, K., Weng, S. et al. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from $Saccharomyces\ cerevisiae$ and related sequences from other organisms. *Nucleic Acids Res*, **32**, D311-4.

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

DeRisi, J., Iyer, V. et.al. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.

Eisen, M., Spellman, P. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci*, **95**, 14863-8.

Filkov, V., Skiena, S. et.al. (1997) Methods for analysis of microarray time-series data. *Journal of Computational Biology*, **9**, 317-330.

Farkas, I., Jeong, H. et.al. (2003) They topology of transcription regulatory network in the yeast, $Saccharomyces\ cerevisiae$. *Physica A*, **318**, 601-612.

Golub, T., Slonim, D. et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-7.

Hero, A., Fleury, G. et. al. (2004) Multicriteria gene screening for analysis of differential expression with DNA microarrays. *EURASIP Journal on Applied Signal Processing*, **1**, 43-52.

Hollander, A., Wolfe, D.(2001) Nonparametric statistical methods. *Wiley-Interscience*, Hoboken, NJ, USA.

Ideker, T., Thorsson, V. et. al. (2001) Testing for differentially expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, **7**, 805-817.

Ideker, T., Thorsson, V. et. al. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929-34.

Jeong, H., Mason, S. et. al. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41-42.

Kwon, A., Holger, H. (2003) Inference of transcriptional regulation relationships from gene expression data. *Bioinformatics*, **19**, 905-912.

Lee, H., Hsu, A. et al.(2004) Coexpression analysis of human genes across many microarray data sets. *Nature*, **14**, 1085-1094.

Lockhart, D., Dong H. et.al.(1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675-1680.

McLachlan, G., Bean, R. et.al. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413-422.

Rohde, J., Trinh, J. et al.(2000) Multiple signals regulate GAL transcription in yeast. *Mol Cell Biol*, **20**, 3880-6.

Schena, M., Shalon, D. et. al. (1995) Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, **270**, 467-470.

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

Speed, T. ed. (2003) Statistical analysis of gene expression microarray data. *Chapman & Hall/CRC Press*, Boca Raton, Fla, USA.

Tamayo, P., Slonim, D. et al.(1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci*, **96**, 2907-2912.

Tusher, V., Tibshirani, R. et. al. (2001) Significance analysis of microarrays applied to the the ionizing radiation response. *Proc Natl Acad Sci*, **98**, 5116-5121.

Wang, J., Myklebost, O. et.al.(2003) MGraph: graphical models for microarray data analysis. *Bioinformatics*, **19**, 2210-1.

Wieczorke, R., Krampe, S. et al.(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. et al. (2004) Dominant spectral component analysis for transcriptional regulations using microarray time-series data. *Bioinformatics*, **20**, 742-9.

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

Zareparsi, S., Hero, A. et.al.(2004) Seeing the unseen: Microarray-based gene expression profiling in vision. *Invest Ophthalmol Vis Sci.*, **45**, 2457-2462.

Zhou, X., Kao, M. et.al. (2002) Transitive functional annotation by shortest path analysis of gene expression data . *Proc Natl Acad Sci*, **99**, 12783-12788.

5 APPENDIX

5.1 Construct PCER-CI for ρ and τ

1a. 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: $I^g(\alpha)$ on each true Pearson correlation coefficient ρ of the set 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 $P(N(0,1) > z_{\alpha/2}) = \alpha/2$.

1b. The asymptotic PCER $(1-\alpha) \times 100\%$ Confidence Interval: $I^g(\alpha)$ on each true Kendall correlation coefficient τ of the set G_1 is constructed as follows:

i. Compute $C_i = \sum_{t=1, t\neq i}^{N} Q((X_i, Y_i), (X_t, Y_t))$, for $i = 1, 2, ..., N.$, where $Q((a,b),(c,d))$ is given by:

$$Q((a,b),(c,d)) = \begin{cases} 1 & \text{if } (d-b)(c-a) > 0, \\ 0 & \text{if } (d-b)(c-a) = 0, \\ -1 & \text{if } (d-b)(c-a) < 0. \end{cases} \qquad (4)$$

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

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

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

5.2.1 *Pearson correlation coefficient ρ .*

i. Specify a covariance matrix $\mathbf{V}$ and a mean vector μ.

ii. Form the Cholesky decomposition of $\mathbf{V}$, i.e. find the lower triangular matrix L such that $\mathbf{V} = LL^T$.

iii. Simulate a vector $\mathbf{z}$ with independent $N(0,1)$ elements.

A vector simulated from the required multivariate normal distribution is then given by $\mu + L\mathbf{z}$.

5.2.2 *Kendall's τ .*

i. Specify a value for τ.

ii. Simulate an $N \times N$ indicator matrix M given τ as follows:

$$M[n,m]_{1 \leq n < m \leq N} = \begin{cases} 1 & \text{if Bernulli}(\frac{1+\tau}{2}) \text{ is TRUE}, \\ -1 & \text{if Otherwise.} \end{cases} \qquad (5)$$

iii. Simulate i.i.d pairs (X_i, Y_i) $(i = 1, 2, ..., N)$ according to M matrix and definition

$$Q((a,b),(c,d)) = \begin{cases} 1 & \text{if } (d-b)(c-a) > 0, \\ -1 & \text{if } (d-b)(c-a) < 0. \end{cases} \qquad (6)$$

No tied observations are generated.