A Statistical Framework to Infer Functional Gene Associations from Multiple Biologically Interrelated Microarray Experiments

[1]Siew-Leng Melinda Teng, [2]X. Jasmine Zhou, [3]Haiyan Huang

[1]Division of Biostatistics, University of California, Berkeley, USA;[2]Program in Molecular and Computational Biology, University of Southern California Los Angeles, USA;[3]Department of Statistics, University of California, Berkeley, USA
slteng@stat.berkeley.edu, xjzhou@usc.edu, hhuang@stat.berkeley.edu

June 2006

Abstract

Inferring functional gene relationships is a major step in understanding biological networks. With microarray data from an increasing number of biologically interrelated experiments, it now allows for more complete portrayals of functional gene relationships involved in biological processes. In current studies of gene relationships, the existence of dependencies between gene expressions from the biologically interrelated experiments, however, has been widely ignored. When not accounted for, these experimental dependencies can result in inaccurate inferences of functional gene relationships, and hence incorrect biological conclusions. This article proposes a statistical framework and a novel gene co–expression measure, named Knorm correlation, to address this problem. The most important aspect of the proposed model is its ability to decompose the interesting biological variations in gene expressions into two mutually independent components each arising from the genes and the experiments, in addition to variations due to random noises. As a result, the Knorm correlation can critically de-correlate the experimental dependencies before estimating the gene relationships, thus leading to improved accuracies in inferring functional gene relationships. Knorm correlation simplifies to the Pearson coefficient when experiments are uncorrelated. Using simulation studies, a yeast microarray and a human microarray dataset, we demonstrate the success of the Knorm correlation as a more accurate and reliable measure, and the adverse impact of experimental dependencies on the Pearson coefficient, in inferring functional gene relationships from interrelated and interdependent experiments.

Keywords: Covariance matrix; Co-expression measure; Experimental dependency; Functional gene relationships; Kronecker product.

1 Introduction

A major task in understanding biological networks is to infer functional relationships or associations between genes involved in biological processes. Given a set of genes subject to various experiments or experimental conditions (related to a biological process of interest), the functional gene relationships of these genes can be broadly described as the relationships between the gene responses observed across the experiments, which are often coordinated in a manner prescribed by the pathways of the biological process. Using microarrays, gene co–expression patterns can be used to infer these functional gene associations. This leads to the fundamental questions of how to quantify these gene co–expressions and how to interpret them in terms of functional gene associations or relationships.

Among various gene co–expression measures defined by parametric or non-parametric approaches (e.g. Brown et al. 2000; Fraley and Raftery 2000; Yeung and Ruzzo 2001; Ramoni, Sebastiani and Cohen 2002), the Pearson coefficient, which is simple to use and provides direct interpretations in terms of positive and negative regulatory associations, remains a widely used core technique for inferring functional gene relationships from microarray data (Kim et al. 2001; Li 2002; Zhou et al. 2005). However, in such analyses, the gene expressions were implicitly assumed to be uncorrelated across the experiments. Current gene co–expression measures do not deal with experimental dependencies in general.

The existence of experimental dependencies is, instead, a very real but widely ignored issue, especially when the experiments are (biologically) interrelated or interdependent. For example, considering three experiments that are carried out independently, where one experiment is on wild type yeast, another experiment involves the mutation of histone H3 and the third experiment involves the mutations of both histones H3 and H4, it is biologically reasonable to expect the expressions of genes are more likely to be highly correlated (dependent) in the latter two experiments, since the mutation of histone H3 occurs in both experiments that result in a state which alters the expressions of genes responsive to histone mutations in a similar way. Fig. 1 in Section 2 demonstrates such dependencies in a yeast microarray dataset (Sabet, Volo, Yu, Madigan and Morse 2004). When not accounted for, redundant genomic signals in gene expressions from many dependent experiments can overwhelm the important informative signals in the few uncorrelated experiments, leading to inaccurate estimates of gene associations, and hence incorrect biological conclusions. This undesirable effect is illustrated by simulation studies in Fig. 2 in Section 2 and Fig A1 in the Appendix, and further demonstrated by a yeast and a human microarray dataset analysis in Table 1 in Section 5. Hence from both simulation studies and real microarray datasets, there is an essential need to account for experimental dependencies in gene co-expression measures.

The problem of defining a gene co-expression measure to more accurately quantify gene associations in presence of experimental dependencies is made more challenging by the data structure in such analysis involving data from multiple experiments. A typical dataset consists of gene expressions for p genes from n experiments, with replicates of the p gene expressions in each experiment observed independently of the replicates of gene expressions in the other experiments. The number of replicates for each experiment may not be the same. As such, two types of variations are present in a gene expression:

the biological variations that contribute to the functional gene relationships, and the variations due to random noises. The biological variations are variations in gene expressions across the experiments as the genes respond in the biological process, and these variations are our parameters of interest. Variations due to random or measurement noises also exist within each gene and within each experiment. Hence, the statistical model to define the appropriate gene co-expression measure needs to model these variations and effects present in such data.

This article proposes a statistical framework to address the statistical problem, and a novel gene co–expression measure, named Knorm correlation, to answer the biological question. The statistical framework provides a model for the gene expressions data incorporating the biological and random variations, and the Knorm correlation provides a more accurate quantification of gene associations in presence of experimental dependencies. Because the proposed model has the ability to decompose the interesting biological variations in gene expressions into two mutually independent components each arising from the genes and the experiments, in addition to variations due to random noises, the Knorm correlation can therefore critically de-correlate the experimental dependencies before estimating the gene associations (correlations) and thus lead to improved accuracies in the gene correlation estimates. Briefly, gene expressions are modeled as responses from a linear additive model with random normal gene effects, experimental effects and gene-experiment interaction effects, and the dependencies between the interaction terms governed by the gene and experimental covariance matrices in a Kronecker product structured form. The gene expressions are then described by a multivariate normal distribution with a Kronecker product structured covariance matrix. Using a yeast microarray dataset as an illustrative example, this article also argues that the multivariate normal distribution with a Kronecker product structured covariance matrix is a natural and valid model for the gene expressions. The gene correlation estimator, Knorm correlation, is derived as the Maximum Likelihood Estimator (MLE), which is then used in conjunction with a bootstrapping technique when applied to real datasets. The Knorm correlation reduces to the Pearson coefficient when experiments are uncorrelated. The Kronecker product correlation structure bears several advantages. Besides maintaining the same experimental correlations across genes and the same gene correlations across experiments, this unique correlation structure also greatly reduces the number of covariance parameters. Using simulation studies, a yeast microarray and a human microarray dataset, we demonstrate the success of the Knorm correlation as a more accurate and reliable measure, and more importantly, the adverse impact of experimental dependencies on the Pearson coefficient, in inferring functional gene associations from interrelated and interdependent experiments.

The article is organized as follows. Section 2 presents empirical evidence of experimental dependencies in a typical microarray dataset. Section 3 introduces our statistical framework, Knorm correlation and estimation procedure. Microarray datasets used in our analyses are described in Section 4, along with some data preprocessing steps. Section 5 presents results of applying Knorm correlation in both simulation studies and real microarray datasets. Finally, Section 6 discusses some practical and technical issues encountered in practice.

2 Empirical Evidence Of Experimental Dependencies

The existence of dependencies between gene expressions across experiments is a real phenomenon, especially in analyses involving multiple biologically interrelated experiments. We use a publicly available yeast microarray dataset (Sabet et al. 2004), typical of the microarray datasets used in functional gene association studies, to illustrate such experimental dependencies. This dataset is generated to investigate the influence of histone modifications on gene regulation, and consists of eight experiments with two to three replicate arrays for each experiment. More dataset descriptions are provided in Section 4. Fig. 1 shows scatter plots of normalized gene expressions between different experiments, e.g. Expt 1 versus Expt 4 in Fig. 1(a) (for each experiment, one of the replicated arrays is randomly selected for plotting).

[Fig. 1 about here]

From Fig. 1, we see that the gene expressions in experiments 1 and 4 are almost uncorrelated, while gene expressions from experiments 3, 4 and 7 display roughly linear dependencies. This observation is consistent with the biological expectations of the experiments, since experiment 1 is on wild type yeast, whereas experiments 3, 4 and 7 are related to histone H3 mutations. It is therefore biologically reasonable to expect that the gene expressions in experiments 3 and 4 are more likely to be similar to each other than that between experiments 1 and 4, since mutations of histone H3 occurred in both experiments 3 and 4. The extent of dependency of gene expressions between experiments varies consistently with the biological expectations of the experiments (see Fig. 3 for experiment descriptions). The scatter plots of gene expressions between other experiments and between different replicated arrays from each experiment (not shown here) show similar stories as described above.

3 Statistical Framework

In this section, we introduce our statistical model for the gene expressions obtained from multiple experiments and the Knorm correlation.

3.1 Statistical Model

For a multivariate gene expression matrix $\mathbf{X}$ consisting of expressions of p genes (rows) in n experiments (columns), we assume the main effects from the genes and experiments are random and additive, and the gene-experiment interaction effects are random with dependencies among interaction terms governed by the gene and experimental covariance matrices (denoted by Σ^G and Σ^E respectively) in a Kronecker product structured form. Σ^G and Σ^E characterize the biological dependencies in the genes and experiments which contribute to the functional relationships. Here, Σ^E describes the dependencies between experimental covariates that affect the gene expressions for each gene in the matrix; Σ^G describes the dependencies between gene covariates for each experiment in the matrix. The random components in the gene effects and experiment effects are contributed by random noises, e.g. measurement errors.

Then, we have

$$\mathbf{X} = \mathbf{G} + \mathbf{E} + \boldsymbol{\Gamma}_{\mathrm{GE}} + \boldsymbol{\varepsilon}\,, \tag{1}$$

where $\mathbf{G}$ and $\mathbf{E}$ are random effects from the genes and experiments respectively, the interaction effects $\boldsymbol{\Gamma}_{\mathrm{GE}}$ are random with vectorized $\boldsymbol{\Gamma}_{\mathrm{GE}}$ distributed as a multivariate normal distribution with zero means and a covariance matrix $\boldsymbol{\Sigma}^{\mathrm{E}} \otimes \boldsymbol{\Sigma}^{\mathrm{G}}$, and $\boldsymbol{\varepsilon}$ represents small random normal noises with zero means. So $E(\mathbf{X}) = \mathbf{G} + \mathbf{E}$. Note that $\mathbf{X}$, $\mathbf{G}$, $\mathbf{E}$, $\boldsymbol{\Gamma}_{\mathrm{GE}}$, and $\boldsymbol{\varepsilon}$ are all matrices of dimension $p \times n$. Our parameter of interest in this model is the gene covariance matrix $\boldsymbol{\Sigma}^{\mathrm{G}}$.

The above modeling of the interaction term is motivated by the following consideration. Consider the ideal case with no random noises but only the biological variations $\boldsymbol{\Gamma}_{\mathrm{GE}}$ which are of our interest. After removing the fixed effects, projecting $\mathbf{X}$ onto the orthogonal gene and experimental eigenspaces determined by $\boldsymbol{\Sigma}^{\mathrm{G}}$ and $\boldsymbol{\Sigma}^{\mathrm{E}}$ would remove the dependencies among interaction terms and result in a matrix of independent random variables. If these random variables are $N(0,1)$, then when both covariance matrices are invertible, we have

$$\boldsymbol{\Lambda} = \mathbf{D}^{-1/2}\ \mathbf{U}^{\mathbf{T}}\ \left(\mathbf{X} - E\left(\mathbf{X}\right)\right)\ \mathbf{V}\ \mathbf{P}^{-1/2} \tag{2}$$

as a matrix of *i.i.d.* $N(0,1)$ random variables, where $\mathbf{P}$ is a diagonal matrix with diagonal elements being the eigenvalues of $\boldsymbol{\Sigma}^{\mathrm{E}}$ and the eigenvectors of $\boldsymbol{\Sigma}^{\mathrm{E}}$ make up the columns of $\mathbf{V}$ (i.e. $\boldsymbol{\Sigma}^{\mathrm{E}} = \mathbf{V}\,\mathbf{P}\,\mathbf{V}^{\mathrm{T}}$), $\mathbf{D}$ is a diagonal matrix with diagonal elements being the eigenvalues of $\boldsymbol{\Sigma}^{\mathrm{G}}$ and the eigenvectors of $\boldsymbol{\Sigma}^{\mathrm{G}}$ make up the columns of $\mathbf{U}$ (i.e. $\boldsymbol{\Sigma}^{\mathrm{G}} = \mathbf{U}\,\mathbf{D}\,\mathbf{U}^{\mathrm{T}}$). When the covariance matrices are singular, the pseudo–inverses of $\mathbf{P}$ and $\mathbf{D}$ can be used for projection. This will achieve a similar projection effect; the elements in the resulting $\boldsymbol{\gamma}$ are either independent $N(0,1)$ random variables or zeros. The number of zeros in $\boldsymbol{\gamma}$ are determined by the ranks of $\boldsymbol{\Sigma}^{\mathrm{G}}$ and $\boldsymbol{\Sigma}^{\mathrm{E}}$.

Following the above consideration and the assumption on $\boldsymbol{\Lambda}$, we can then naturally model $\mathbf{X}$ by

$$\mathbf{X} = E(\mathbf{X}) + \left(\mathbf{U}\mathbf{D}^{1/2}\right)\ \boldsymbol{\Lambda}\ \left(\mathbf{P}^{1/2}\mathbf{V}^{\mathbf{T}}\right) + \boldsymbol{\varepsilon}, \tag{3}$$

where $E(\mathbf{X}) = \mathbf{G} + \mathbf{E}$, and $\boldsymbol{\varepsilon}$ represents the small random normal noises with zero means as in equation (1). Under this model, if we ignore the error term, vectorized $\mathbf{X}$ then follows a multivariate normal distribution with mean $E(\mathbf{X}) = \mathbf{G} + \mathbf{E}$ and a covariance matrix $\boldsymbol{\Sigma}^{\mathrm{E}} \otimes \boldsymbol{\Sigma}^{\mathrm{G}}$, which is equivalent to equation (1) (see Appendix for a detailed proof). Equations (2) and (3), together with the arguments on $\boldsymbol{\gamma}$, provide a reasonable rationalization on the model described in equation (1). Justifications on the *i.i.d.* normal assumptions on $\boldsymbol{\gamma}$ are presented in a yeast microarray dataset analysis in Section 5.

To make the model identifiable, we further assume that $\left(\mathbf{G}\right)_{ij} = \mu_i$, $\left(\mathbf{E}\right)_{ij} = E_j = 0$ and that each experiment has a unit variance. The assumptions on experiment mean and variance are valid as the (RMA) normalized gene expressions have the same mean and variability in each experiment, and without loss of generality, we can set them to 0 and 1 respectively. With these assumptions, $E(X_{ij}) = \mu_i$, where X_{ij} is the (*i*, *j*)th element in $\mathbf{X}$. We should also note that the identifiability constraints can be different when different datasets are considered. In general, the identifiability constraints should be imposed based on the nature of the data and the purpose of analysis.

Our model provides a single framework to explain variations in a gene expression by two mutually independent biological variations each arising from the genes and the experiments, and the variations due to random noises. As such, the proposed model can concurrently model both gene and experimental dependencies, by which we also obtain a significant reduction of parameter space describing the correlation matrix of **X** from $\sim (np)^2$ to $\sim (n^2 + p^2)$. Another advantage of our model is that the Kronecker product structured covariance matrix maintains same experimental correlations across genes and same gene correlations across experiments, which agrees with the biological expectations of the data.

3.2 Parameter Estimation

Based on the unit variance assumption on experiments, we will use $\mathbf{R}^{\mathrm{E}}$, the experimental correlation matrix, to represent the experimental covariance matrix henceforth. Given a gene expression matrix $\mathbf{X}$, which is assumed to be generated from our model in equation (1), the MLEs of $\boldsymbol{\Sigma}^{\mathrm{G}}$, $\mathbf{R}^{\mathrm{E}}$, and $\boldsymbol{\mu}$ can be derived as

$$\hat{\boldsymbol{\Sigma}}^{\mathrm{G}} = \frac{1}{n}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{T}\right)\left(\mathbf{R}^{\mathbf{E}}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{T}\right)^{T}, \tag{4}$$

$$\hat{\mathbf{R}}^{\mathrm{E}} = \frac{1}{p}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{T}\right)^{T}\left(\boldsymbol{\Sigma}^{\mathbf{G}}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{T}\right), \tag{5}$$

$$\hat{\boldsymbol{\mu}} = \frac{\mathbf{X}\left(\mathbf{R}^{\mathbf{E}}\right)^{-1}\mathbf{1}}{\mathbf{1}^{T}\left(\mathbf{R}^{\mathbf{E}}\right)^{-1}\mathbf{1}}, \tag{6}$$

where **1** is a $n \times 1$ column vector of ones. Our parameter of interest is the gene correlation matrix $\mathbf{R}^{\mathbf{G}}$, which can be estimated as

$$\hat{\mathbf{R}}^{\mathrm{G}} = \mathbf{W}^{\mathbf{-1/2}}\,\hat{\boldsymbol{\Sigma}}^{\mathrm{G}}\,\mathbf{W}^{\mathbf{-1/2}}, \tag{7}$$

where **W** is a diagonal matrix with same diagonal elements in $\hat{\boldsymbol{\Sigma}}^{\mathbf{G}}$. The detailed derivations of the MLEs are presented in the Appendix.

It can be shown that the MLE of $\boldsymbol{\mu}$ derived from equation (4) is an unbiased and consistent estimator of $\boldsymbol{\mu}$, and that the MLE of $\mathbf{R}^{\mathbf{E}}$ and $\boldsymbol{\Sigma}^{\mathbf{G}}$ are consistent estimators when $\boldsymbol{\Sigma}^{\mathbf{G}}$ and $\mathbf{R}^{\mathbf{E}}$ are respectively known. A brief proof is provided in the Appendix.

3.2.1 Bootstrapping Procedure to Estimate Gene Correlations in Real Microarray Datasets. In real microarray datasets, we do not observe true matrix replicates (we only observe replicates of each single experiment/array) to derive our correlation estimators via the iterative procedure. Hence, we implement an additional bootstrapping procedure to construct the $p \times n$ data matrices from the replicates of each experiment. A $p \times n$ bootstrapped data matrix $\mathbf{X}$ is constructed by placing in the *j*th column of $\mathbf{X}$ a randomly selected replicate ($p \times 1$ vector of gene expressions for *p* genes) from the *j*th experiment, such that the bootstrapped matrix is a random sample from all observed arrays. By our model specified in equation (1), each bootstrapped matrix is sampled from the same probability distribution. We first obtain a reliable estimate of $\mathbf{R}^{\mathrm{E}}$ and then use it to estimate $\boldsymbol{\Sigma}^{\mathrm{G}}$. To estimate $\mathbf{R}^{\mathrm{E}}$, we (*i*) compute the experimental covariance matrix using the Pearson coefficient for each bootstrapped matrix, and (*ii*) take the average of the experimental covariance estimates over *B*. With $\hat{\mathbf{R}}^{\mathrm{E}}$, we estimate $\hat{\mathbf{R}}^{\mathrm{G}}$ as follows: (*i*) for the b^{th} bootstrapped matrix where $b = 1, \ldots, B$, obtain $\hat{\boldsymbol{\Sigma}}^{\mathrm{G,b}}$ using equation (2), where $\hat{\boldsymbol{\mu}}$ is estimated from equation (4) via $\hat{\mathbf{R}}^{\mathrm{E}}$, (*ii*) compute $\hat{\mathbf{R}}^{\mathrm{G,b}}$ using equation (7), and (*iii*) take the average over all estimates $\hat{\mathbf{R}}^{\mathrm{G,b}}$, i.e. $\hat{\mathbf{R}}^{\mathrm{G}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\mathbf{R}}^{\mathrm{G,b}}$. The gene correlations obtained from $\hat{\mathbf{R}}^{\mathrm{G,b}}$ are called Knorm correlations. We use *B*=500 in both the yeast and human datasets.

4 Data And Preprocessing

4.1 Microarray Datasets

We use two publicly available microarray datasets that are typical in current studies of functional gene relationships. These two datasets each consists of gene expressions obtained from biologically interrelated experiments underlying a biological process of interest.

4.1.1 Yeast microarray dataset. This dataset comes from a study by Sabet et al. (2004) to investigate the influence of histone modifications on gene regulation, consisting of gene expressions from eight experiments with two to three replicate arrays for each experiment. Experiment descriptions are provided in Fig. 3. This dataset is accessible through the NCBI Gene Expression Omnibus Database by the accession number GDS772.

3.1.2 Human microarray dataset. This dataset is generated by Lund et al. (2005) to study mechanisms regulating CD4+ cell polarization. CD4+ lymphocytes were induced to differentiate into Th1 and Th2 through treatment with IL–12 or IL–4 in the presence of TGFbeta. The dataset consists of 16 experiments conducted using 5 related treatments at three time points besides the untreated cells. There are two to four replicated microarrays for each experiment, resulting in a total of 34 microarrays. Experiment descriptions are shown in Fig. 5.

4.2 Data Preprocessing

The raw data from each microarray dataset are first normalized using the robust multi-array average (RMA) method developed by Irizarry et al. (2003). We next proceed to select a set of genes for our analyses, to be used to assess the performance of correlation estimators in inferring functional gene associations. Since the main purpose of this gene set is to allow the comparison of estimator performance and not as an attempt to identify genes significantly differentially expressed over the experiments, we use a set of Gene Ontology (GO) annotated genes with high expression variations across the experiments. These genes are more likely to be genes responsive to the biological process, e.g. for the histone mutation due to the high expression changes, and therefore should be more biologically interesting in the context of the experimental datasets. Being GO annotated, it allows a way of verifying the inferred functional gene associations via the GO. From the GO annotated genes, we identify genes with high gene expression variations as follows: (*i*) for each experiment, rank the genes by their average expression over replicates, (*ii*) for each gene, obtain the difference between the maximum and minimum rank across the experiments, (*iii*) a gene is identified as highly variably expressed if this difference exceeds a specified threshold. We chose the top 20% of such genes (532 genes) for the yeast microarray dataset, and the top 10% of such genes (526 genes) for the human microarray dataset. Note that this selection procedure is employed to select a set of genes (likely to be responsive to the experiments) based on which to assess the performance of correlation estimators in inferring functional gene associations. Other gene sets can also be used for this purpose.

5 Results

In this section, we first apply our proposed correlation measure, Knorm correlation, to a simulation dataset, a yeast microarray dataset, and a human microarray dataset to evaluate its performance. To biologically evaluate the gene associations inferred from the real datasets, we assess the gene functional similarity based on GO Biological Process annotation. Since there is no gold standard gene association measure, we use the Pearson coefficient as a comparison benchmark because of its widespread use and similar interpretations to our Knorm correlation in terms of gene associations. The results demonstrate the success of our proposed method in practical applications.

5.1 Application of Knorm Correlation to a Simulation Dataset

In this simulation study, we demonstrate the increased accuracies of Knorm correlation estimates over that of the Pearson coefficients in presence of increasing column (e.g. experiment) dependencies in two *correlated* row vectors (e.g. genes). At each p% dependency level (with p=1,…,100), we first generate 1000 *i.i.d.* column vectors of dimension 2, each from a bivariate normal distribution with zero means, unit variances and a correlation of 0.17, and then assign the first 1000p% vectors to be the same as the first vector (while remaining the last 1000(1-p)% independent vectors unchanged). Putting these 1000 column vectors of dimension 2 into a matrix, we now obtain two row vectors of dimension 1000 with a true row correlation of 0.17, and p% of the vector components being identical. We then compute both the Pearson coefficient and Knorm correlation of the two row vectors, and plot the estimates in blue and red respectively in Fig. 2.

[Fig. 2 about here]

The Knorm correlation was computed using equation (4) with the column correlation matrix known by the construction procedure of the row vectors at the p% dependency level. Fig. 2 shows the effectiveness of Knorm correlation. The Knorm correlation estimate is closer to the true correlation of 0.17 and has a much smaller variance until we reach about an 80% dependency, than the Pearson coefficient which rapidly fails in accuracy after an approximate 5% dependency between the row vector components. We also have similar observations for simulation studies with different values of true correlations, both negative and positive (besides the value 0.17), and we only present the simulation study with a true correlation of 0.17 here as an illustrative example.

Another simulation study demonstrating the effectiveness of an iterative procedure for estimating both the row (e.g. gene) and column (e.g. experiment) correlation matrices in the case of an unknown column correlation matrix is presented in the Appendix. This iterative procedure is suggested by the MLEs of the row and column correlation matrices in equations (4)–(7). However, in applying the Knorm correlation to real datasets, we need to modify the iterative estimation procedure and implement a bootstrapping procedure to estimate both the gene and experimental correlation matrices, since the real datasets do not have replicates of gene expression matrix as opposed to simulation data. We refer the reader to Section 3.2 for the motivation and detailed descriptions of the bootstrapping procedure.

5.2 Application of Knorm Correlation to a Yeast Microarray Dataset

We present our results of applying Knorm correlation to a public yeast microarray dataset (Sabet et al. 2004). We apply the Knorm correlation to GO annotated genes with high expression variations across the eight biologically related experiments selected by a procedure described in Section 4.2. These genes, being more likely to be responsive genes for the histone mutation due to the high expression changes and therefore should be more biologically interesting in the context of this experimental dataset, is used to assess the performance of both the Pearson and Knorm correlation in inferring gene-gene associations

The estimated correlations between the eight experiments, shown in Fig. 3, are favorably consistent with the biological expectations. These estimated experimental correlations agree with the scatter plots shown in Fig. 1. The experimental correlation matrix describes the dependencies between experimental covariates that are assumed to affect the gene expressions in the same manner for all responsive genes.

[Fig. 3 about here]

Based on the estimated experimental correlation matrix shown in Fig. 3, we estimate the gene correlations using Knorm correlation. The Pearson coefficients are also computed using the gene expressions that are averaged over the replicates within each experiment (a common approach in practice). The functional associations between genes are then predicted based on the sign and magnitude of their correlation estimates. The magnitude reflects the extent of a gene pair's synchronous response to the experiments. A positive sign indicates a parallel response while a negative sign suggests an opposite response. We first assess the performance of the gene correlation estimates using GO annotations. We consider genes as being annotated functionally related if they are in the same GO node at level 6 or more below the root. We compute the percentage of functionally related gene pairs from among those with the highest Knorm correlation or Pearson coefficient (in absolute value). Knorm correlation reports consistently higher percentages of annotated functionally related gene pairs than those obtained by the Pearson coefficient (see Table 1).

[Table 1 about here]

Out of the top 10, 30, 50 and 100 gene pairs in estimated correlations, 30%, 36.7%, 38% and 27% (respectively) gene pairs identified by Knorm are known to be functionally related by GO annotations whereas only 10%, 20%, 26% and 21% (respectively) are functionally related gene pairs for Pearson coefficient. This suggests that Knorm correlation is more effective in inferring functional gene associations than the Pearson coefficient. The distinction is especially strong for the gene pairs with highly ranked correlations. In general, the higher the correlation estimate, the more likely the inferred functional association is true. It is worthwhile to note that the percentages of functionally related gene pairs from both the proposed method and Pearson approach in Table 1 decrease and the percentage differences become stable as more top gene pairs are considered.

Many functionally related gene pairs with high Knorm correlation but low Pearson coefficient are supported by literature. For example, MCM1 and SWI5 yield a Knorm correlation of 0.47, but a Pearson coefficient of only –0.08. Since reduced acetylation of histone amino termini is known to be associated with reduced transcription levels of SWI5 (Deckert and Struhl 2002; Shimizu, Takahashi, Lamb, Shindo and Mitchell 2003) and MCM1 is known to be a direct regulator of SWI5 (Kumar et al. 2000; Lee et al 2002), the expressions of SWI5 and MCM1 are expected to show positive correlation in this dataset where histone amino termini have been deleted or modified. The Knorm correlation has confirmed this expectation. As another example, both CKA1 and PMC1 are involved in maintaining cell ion homeostasis and yeast growth. Knorm correlation gives a positive estimate of 0.62 between these two genes, which reflects their related roles, while the Pearson coefficient gives an estimate of only –0.02. Scatter plots presenting the gene expressions of the above two gene pairs before and after removing the experimental dependencies are shown in Fig. 4, demonstrating the necessity of de-correlating experimental dependencies for a more accurate correlation estimation. As another example, both HSF1 and CTK3 are involved in the regulation of transcription from RNA polymerase II promoter, with an expected positive correlation, which was revealed by Knorm correlation (0.53), but not by the Pearson coefficient (–0.03).

5.3 Application of Knorm Correlation to a Human Microarray Dataset

We next apply Knorm correlation to GO annotated genes, selected by a procedure described in Section 4.2, for a human dataset presented by Lund et al. (2005). We obtain both the Pearson coefficient and Knorm correlation estimates for each gene pair.

Fig. 5 shows our estimated correlations between the 16 experiments, which are favorably consistent with the biological expectations. Our estimated correlation matrix effectively captures the dependencies (*i*) between biologically similar experimental conditions, e.g. untreated cells, and experiments conducted at two hours after treatment when the treatment effect was not yet obvious, (*ii*) between experiments with the same treatment at different time points, e.g. antiCD3+antiCD28+IL–12 experiments at two hours, six hours, and 48 hours after treatment, and (*iii*) between experiments with different but similar treatments at the same time points.

[Fig. 5 about here]

In validating the potential functional associations of top gene pairs (ranked by absolute correlation estimates), we use the GO annotations to evaluate the results. We see from Table 1 that our Knorm correlation again reports favorably higher percentages of annotated functionally related gene pairs than those obtained by the Pearson coefficient, especially for the very highly ranked gene pairs. Like the consistently high percentages observed for the Knorm correlation in the yeast dataset, the percentages for the human dataset shown in Table 1 again reinforce the effectiveness of Knorm correlation in inferring functional gene associations. We should also note that the percentages in human dataset are consistently lower than those in yeast dataset, which can be attributed to the poor annotations in human genome.

Similar to the yeast results, many gene pairs predicted to have functional relevance by Knorm correlation but not Pearson coefficient are validated by experiments in the literature. For example, APEX1 is a rate–limiting enzyme in DNA base excision repair. MSH6 is a primary DNA mismatch repair gene. A recent study reported that the expression of APE protein leads to the suppression of DNA mismatch repair and that the MSH6 protein was markedly reduced in the APE–expressing cells (Chang et al 2005). Agreeing with these previous findings, our method reports a negative correlation of –0.41 (4.5th percentile) between the two genes, which the Pearson coefficient fails to capture with a value of –0.18 (34th percentile). Another example is the gene pairs RB1 and CDKN1A. It has been reported that the retinoblastoma protein RB1 is a cooperating factor for the transcription factor MITF to activate the expression of the cyclin–dependent kinase inhibitor gene CDKN1A, which contributes to the cell cycle exit and activation of the differentiation program (Carreira et al. 2005). In accordance with this fact, Knorm correlation yields a correlation estimate of 0.45 between RB1 and CDKN1A, in contrast to a value of –0.05 provided by the Pearson coefficient.

5.4 Model Justification

A key assumption in our probability model is the *i.i.d.* normal assumption on the elements in $\mathbf{\Lambda}$ in equation (2). As an attempt to justify that our probability model is a reasonable model in practice, we examine the qq-plot of the elements in $\hat{\mathbf{\Lambda}}$, estimated from the yeast dataset analysis in Section 5.2, against a standard normal distribution, in addition to performing a Kolmogorov-Smirnov (K-S) test on the elements in $\hat{\mathbf{\Lambda}}$. $\hat{\mathbf{\Lambda}}$ is computed by equation (2) using the mean and covariance matrices estimated for the yeast dataset. Fig. 6 shows a randomly selected qq-plot among those obtained from 500 replicated expression matrices constructed through the bootstrapping procedure.

[Fig. 6 about here]

The qq-plot in Fig. 6 is clearly suggestive of a standard normal distribution for the elements in $\hat{\mathbf{\Lambda}}$ (with a P-value of 0.2 for the K-S test). Overall we observe good qq-plots with an average p-value of 0.34 for the K-S tests. The same study on the human dataset (Lund et al. 2005) yields similar observations. Therefore our *i.i.d.* N(0, 1) assumptions on $\boldsymbol{\gamma}$ are reasonably valid for the yeast and human microarray datasets, especially considering that the random noises $\boldsymbol{\varepsilon}$ in model equation (1) are confounded with $\mathbf{R}^{\mathbf{E}}$ and $\mathbf{\Sigma}^{\mathbf{G}}$ which can make the model justification more difficult.

6 Discussion

This article has introduced a naturally intuitive statistical framework and novel correlation measure to estimate the row correlation matrix for matrix data when the columns are no longer uncorrelated. When applied to microarray data, the matrix is the gene expression matrix with the (*i*,*j*)th element being the expression of gene *i* in experiment *j*, and rows correspond to genes and columns correspond to experiments. Our method can more precisely capture functional gene associations in terms of their involvement in a biological process when experiments are correlated in two typical microarray datasets. Knorm correlation is derived by modeling the gene expression matrix through a multivariate normal distribution with a Kronecker product structured covariance matrix.

In practice, we face several challenges when applying the approach to real datasets. First is the specificity of the gene set to the biological process of interest. Equations (4)–(7) show that the estimations of gene and experimental correlation matrices (dependencies) are intertwined. Using different sets of genes could yield different estimates of experimental correlation matrices. If a gene set consists of many unresponsive genes, the gene or experimental dependencies may be greatly obscured by irrelevant noises. In our applications to the two mircoarray datasets, the gene sets were selected to be genes highly likely to be responsive to the experiments for the purpose of assessing the performance of the gene correlation estimators in inferring functional gene associations. However, to be able to elicit more and to accurately determine a set of functional gene associations specific to a biological process, it is critical to select an appropriate gene set that highly relates to the biological process of interest. The second challenge is the presence of random noise or non–biological variations. In practice, a gene expression measurement consists of not only the interesting biological signals, it also consists of measurement errors and/or biologically irrelevant noises. By design (often with decisions beyond our control for publicly available datasets), these interesting biological signals are often confounded with biologically irrelevant variations (e.g. measurement errors), which makes it difficult to estimate the biological dependencies of interest. Our gene correlation estimates become conservative because of these non-biological noises. The larger the noises, the more conservative our correlation estimates will be. But more importantly, the Knorm correlation estimate will still retain the sign of the true correlation, and hence is still able to provide an accurate inference of functional gene association. Third is the inference of directional relationships from the inferred functional gene associations. Correlation only provides a first step in inferring functional gene relationships; it provides a measure whether the genes are associated with one another in the biological process of interest. After gene associations have been established, a set of functionally related genes can then be identified (as carried out in the yeast and human microarray dataset analyses), and if of further interest, other technologies may be employed to determine their directional relationships.

There are also several technical issues when applying our approach to real datasets. First, replicates from each experiment are observed instead of replicates of full gene expression matrices. We addressed this problem by implementing the bootstrapping procedure (see Section 3.2), i.e. constructing a gene expression matrix by putting in each column of the matrix a randomly selected replicate from the corresponding experiment. The bootstrapping procedure, in effect, resamples gene expression matrices from the multivariate normal distribution defined in equation (1). Second, we have much more genes than experiments ($p >> n$) in real datasets, which can affect the quality of parameter estimates through the iterative procedure suggested by equations (4)-(7). One possible way to reduce the parameter space of the gene covariance matrix is to explore special matrix structures, like the idea in Lasso (Tibshirani 1996).

We have shown that considering experimental dependencies is important in making more accurate functional gene association inferences. Our applications to yeast and human datasets yield promising and biologically meaningful results. It is reasonable to expect that Knorm correlation can improve the accuracy of biological inferences made from those experiments which are currently (and incorrectly) assumed to be uncorrelated.

APPENDIX A: IMPACT OF EXPERIMENTAL DEPENDENCIES ON PEARSON COEFFICIENTS

In addition to the results presented in Fig. 2 in Section 5.1, we further investigate the adverse impact of experimental (column) dependencies on the Pearson coefficient of two *uncorrelated* genes across eight experiments (i.e. two row vectors of dimension 8). In this simulation study, we simulate gene expression matrices with rows corresponding to genes, and columns corresponding to experiments. Three different correlation matrices are used to describe the column dependencies: in Fig. A1(a), the column correlation matrix is an identity matrix to simulate for row vectors with independent vector components; in Fig. A1(b), the column correlation matrix consists of a mixture of zero and positive elements to simulate for row vectors with moderately positively correlated components; in Fig. A1(c), the column correlation matrix consists of elements in a range of 0.8 to 1.0 to simulate for row vectors with highly positively correlated components. Each histogram in Fig. A1 consists of 5000 Pearson coefficients, each computed from a pair of row vectors that are independently generated by a common multivariate normal distribution with zero means, unit variances and a specified correlation matrix as described above.

[Fig. A1 about here]

From Figs. A1(a)–A1(c), we see a change in the distribution of the Pearson coefficients with increasing dependencies between the vector components. Fig A1(a) shows a histogram representing the true distribution of Pearson coefficients between the two uncorrelated vectors. The distributions of the Pearson coefficients in Figs A1(b) and A1(c) become more skewed toward the larger absolute correlation values as dependencies between the components increase. Fig. A1 is a clear demonstration of the adverse impact that experimental dependencies can have on the Pearson coefficients.

APPENDIX B: EFFECTIVENESS OF THE ITERATIVE PROCEDURE FOR ESTIMATING BOTH ROW AND COLUMN CORRELATION MATRICES

The MLEs of the row and column correlation matrices (denoted by $\mathbf{R}^{\mathbf{E}}$ and $\mathbf{R}^{\mathbf{G}}$) in equations (4)–(7) suggest an iterative estimation procedure. This simulation study demonstrates the effectiveness of the iterative procedure for estimating $\mathbf{R}^{\mathbf{G}}$ and $\mathbf{R}^{\mathbf{E}}$ when the column correlation matrix $\mathbf{R}^{\mathbf{E}}$ is unknown. Here, we simulate gene expressions for 100 genes in 15 experiments; we generate five replicated matrices of dimension 100×15, where each vectorized matrix transpose follows a multivariate normal distribution with zero means, unit variances, and a correlation matrix $\mathbf{R}^{\mathbf{G}} \otimes \mathbf{R}^{\mathbf{E}}$. The gene correlation matrix, $\mathbf{R}^{\mathbf{G}}$, has about 50% of its elements ranging between –0.45 and 0.45. In Fig. A2(a), the experimental correlation matrix, $\mathbf{R}^{\mathbf{E}}$, is an identity matrix (i.e. experiments are correlated), whereas in Figs. A2(b), A2(d) and A2(e), $\mathbf{R}^{\mathbf{E}}$ is such that 9 out of 15 experiments are positively correlated. The reported Pearson coefficients are computed as the average of the five Pearson coefficients computed in each replicate by treating the experiments as independent observations.

[Fig. A2 about here]

Fig. A2(e) clearly shows that the gene correlation estimates by the iterative procedure (using 10 iterations) are more accurate than the Pearson estimates shown in Fig. A2(d). Figs. A2(b) and A2(e) also show that the iterative estimates (with the experimental dependencies unknown) are close to those estimated using the known true experimental correlation matrix. In addition, we also see that the iterative approach, shown in Fig. A2(b), achieves comparable gene correlation estimates as would the Pearson approach in the uncorrelated experiments case, shown in Fig. A2(a). Fig. A2(a) shows the estimation variations in the Pearson coefficients in this simulation dataset.

APPENDIX C: DERIVATION OF KRONECKER PRODUCT STRUCTURED COVARIANCE MATRIX OF GENE EXPRESSION MATRIX X UNDER OUR MODEL

Theorem S1. Given a $p \times n$ matrix $\mathbf{\Lambda}$ of *i.i.d.* elements with mean 0 and unit variances, the covariance matrix of $\mathbf{X} = \mathbf{U}\ \mathbf{D}^{1/2}\ \mathbf{\Lambda}\ \mathbf{P}^{1/2}\mathbf{V}^{\mathrm{T}}$ is $\mathbf{\Sigma}^{\mathrm{G}} \otimes \mathbf{\Sigma}^{\mathrm{E}}$, where $\mathbf{\Sigma}^{\mathrm{G}} = \mathbf{U}\ \mathbf{D}\ \mathbf{U}^{\mathrm{T}}$ and $\mathbf{\Sigma}^{\mathrm{G}} = \mathbf{V}\ \mathbf{P}\ \mathbf{V}^{\mathrm{T}}$ are the singular value decompositions of $\mathbf{\Sigma}^{\mathrm{G}}$ and $\mathbf{\Sigma}^{\mathrm{E}}$ respectively.

Proof. Letting $\mathbf{\Omega} = \mathbf{\Lambda}\ \mathbf{P}^{1/2}\ \mathbf{V}^{\mathrm{T}}$, we have $\mathbf{X} = \mathbf{U}\ \mathbf{D}^{1/2}\ \mathbf{\Omega}$. Since $\mathbf{\Lambda}$ is a $p \times n$ matrix of *i.i.d.* elements with unit variances, the covariance matrix of $\mathbf{\Lambda}$ is $\mathbf{I}_{\mathbf{pn}}$, or equivalently $\mathrm{Cov}\left(\mathrm{vec}(\mathbf{\Lambda})\right) = \mathbf{I}_{\mathbf{pn}}$, where $\mathbf{I}_{\mathbf{pn}}$ is a $(pn) \times (pn)$ identity matrix.

Now we consider the covariance matrix of $\mathbf{\Omega}$. Let Ω_{ij} be the (*i*,*j*)th element in $\mathbf{\Omega}$, $\mathbf{e}_{\mathbf{i}}$ be a *p*-dimensional column vector of zeroes except a value of 1 at the *i*th element and $\mathbf{f}_{\mathbf{j}}$ be a *n*-dimensional column vector of zeroes except a value of 1 at the *j*th element. Then we have

$$\begin{aligned}
\mathrm{Cov}\left(\Omega_{ij}, \Omega_{i'j'}\right) &= \mathrm{Cov}\left(\mathbf{e}_i^{\mathrm{T}}\ \mathbf{\Lambda}\ \mathbf{P}^{1/2}\ \mathbf{V}^{\mathrm{T}}\ \mathbf{f}_j,\ \mathbf{e}_{i'}^{\mathrm{T}}\ \mathbf{\Lambda}\ \mathbf{P}^{1/2}\ \mathbf{V}^{\mathrm{T}}\ \mathbf{f}_{j'}\right) \\
&= \mathbf{f}_j^{\mathrm{T}}\ \mathbf{V}\ \mathbf{P}^{1/2}\ \mathrm{Cov}(\mathbf{e}_i^{\mathrm{T}}\ \mathbf{\Lambda}, \mathbf{e}_{i'}^{\mathrm{T}}\ \mathbf{\Lambda})\ \mathbf{P}^{1/2}\ \mathbf{V}^{\mathrm{T}}\ \mathbf{f}_{j'} \\
&= \begin{cases} \mathbf{f}_j^{\mathrm{T}}\ \mathbf{V}\ \mathbf{P}^{1/2}\mathbf{P}^{1/2}\ \mathbf{V}^{\mathrm{T}}\ \mathbf{f}_{j'} & \text{when } i = i' \\ 0 & \text{when } i \neq i' \end{cases} \\
&= \begin{cases} \mathbf{f}_j^{\mathrm{T}}\ \mathbf{\Sigma}^{\mathrm{E}}\ \mathbf{f}_{j'} & \text{when } i = i' \\ 0 & \text{when } i \neq i' \end{cases} \\
&= \begin{cases} \left(\mathbf{\Sigma}^{\mathrm{E}}\right)_{jj'} & \text{when } i = i' \\ 0 & \text{when } i \neq i' \end{cases}
\end{aligned} \tag{A1}$$

Therefore, the covariance matrix of $\mathbf{\Omega}$ is $\mathbf{I}_{\mathbf{p}} \otimes \mathbf{\Sigma}^{\mathrm{E}}$. Furthermore, letting X_{ij} to be the (*i*,*j*)th element in $\mathbf{X}$, we have

$$\begin{aligned}\mathrm{Cov}\left(X_{ij}, X_{i'j'}\right) &= \mathrm{Cov}\left(\mathbf{e}_i^{\mathbf{T}}\ \mathbf{U}\ \mathbf{D}^{1/2}\ \boldsymbol{\Omega}\ \mathbf{f}_j,\ \mathbf{e}_{i'}^{\mathbf{T}}\ \mathbf{U}\ \mathbf{D}^{1/2}\ \boldsymbol{\Omega}\ \mathbf{f}_{j'}\right) \\ &= \mathbf{e}_i^{\mathbf{T}}\ \mathbf{U}\ \mathbf{D}^{1/2}\ \left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)_{jj'} \mathbf{I}_{\mathbf{p}}\ \mathbf{D}^{1/2}\ \mathbf{U}\ \mathbf{e}_{i'} \\ &= \left(\mathbf{e}_i^{\mathbf{T}}\ \boldsymbol{\Sigma}^{\mathbf{G}}\ \mathbf{e}_{i'}\right)\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)_{jj'} \\ &= \left(\boldsymbol{\Sigma}^{\mathbf{G}}\right)_{ii'}\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)_{jj'}\end{aligned} \qquad \text{(A2)}$$

Thus the covariance matrix of $\mathbf{X} = \mathbf{U}\ \mathbf{D}^{1/2}\ \boldsymbol{\Omega}$ is $\boldsymbol{\Sigma}^{\mathbf{G}} \otimes \boldsymbol{\Sigma}^{\mathbf{E}}$. ■

APPENDIX D: DERIVATION OF MLEs IN EQUATIONS (4)–(6)

Theorem S2. Let the covariance matrices $\boldsymbol{\Sigma}^{\mathbf{G}}$ and $\boldsymbol{\Sigma}^{\mathbf{E}}$ be invertible. Given that $\mathrm{vec}(\mathbf{X})$ follows a multivariate normal distribution with mean $\mathrm{vec}(E(\mathbf{X})) = \mathrm{vec}(\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}})$ and covariance matrix $\boldsymbol{\Sigma}^{\mathbf{G}} \otimes \boldsymbol{\Sigma}^{\mathbf{E}}$, where $\mathbf{1}$ is a column vector of ones, the Maximum Likelihood Estimators (MLEs) of $\boldsymbol{\Sigma}^{\mathbf{G}}$, $\boldsymbol{\Sigma}^{\mathbf{E}}$ and $\boldsymbol{\mu}$ are given in equations (4)–(6) in Section 3 respectively.

Proof. By the assumed multivariate normal model, the log-likelihood function of an observed $\mathbf{X}$ is

$$l(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma}^{\mathbf{E}},\boldsymbol{\Sigma}^{\mathbf{G}}) = -\frac{p}{2}\log|\boldsymbol{\Sigma}^{\mathbf{E}}| - \frac{n}{2}\log|\boldsymbol{\Sigma}^{\mathbf{G}}| - \frac{1}{2}tr\left(\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)^{\mathbf{T}}\left(\boldsymbol{\Sigma}^{\mathbf{G}}\right)^{-1}\right).$$

Then the first partial derivatives of $l(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma}^{\mathbf{E}},\boldsymbol{\Sigma}^{\mathbf{G}})$ with respect to $\boldsymbol{\Sigma}^{\mathbf{G}}$, $\boldsymbol{\Sigma}^{\mathbf{E}}$ and $\boldsymbol{\mu}$ are

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\Sigma}^{\mathbf{G}}} &= -\frac{n}{2}\left(\frac{\partial}{\partial \boldsymbol{\Sigma}^{\mathbf{G}}}\log|\boldsymbol{\Sigma}^{\mathbf{G}}|\right) - \frac{1}{2}\left(\frac{\partial}{\partial \boldsymbol{\Sigma}^{\mathbf{G}}} tr\left(\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)^{\mathbf{T}}\left(\boldsymbol{\Sigma}^{\mathbf{G}}\right)^{-1}\right)\right) \\ &= -\frac{n}{2}\left(\boldsymbol{\Sigma}^{\mathbf{G}}\right)^{-1} + \frac{1}{2}\left(\left(\boldsymbol{\Sigma}^{\mathbf{G}}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)^{\mathbf{T}}\left(\boldsymbol{\Sigma}^{\mathbf{G}}\right)^{-1}\right),\end{aligned}$$

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\Sigma}^{\mathbf{E}}} &= -\frac{p}{2}\left(\frac{\partial}{\partial \boldsymbol{\Sigma}^{\mathbf{E}}}\log|\boldsymbol{\Sigma}^{\mathbf{E}}|\right) - \frac{1}{2}\left(\frac{\partial}{\partial \boldsymbol{\Sigma}^{\mathbf{E}}} tr\left(\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)^{\mathbf{T}}\left(\boldsymbol{\Sigma}^{\mathbf{G}}\right)^{-1}\right)\right) \\ &= -\frac{p}{2}\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)^{-1} + \frac{1}{2}\left(\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)^{\mathbf{T}}\left(\boldsymbol{\Sigma}^{\mathbf{G}}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)^{-1}\right),\end{aligned}$$

$$\frac{\partial l}{\partial \boldsymbol{\mu}} = \left(\boldsymbol{\Sigma}^{\mathbf{G}}\right)^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}\right)\left(\boldsymbol{\Sigma}^{\mathbf{E}}\right)^{-1}\mathbf{1}.$$

As the normal distribution belongs to the exponential family and its log-density function is concave, the MLEs can be obtained by equating the above derivatives to zero and solving for $\boldsymbol{\Sigma}^{\mathbf{G}}$, $\boldsymbol{\Sigma}^{\mathbf{E}}$ and $\boldsymbol{\mu}$, by which, we then obtain the MLEs of $\boldsymbol{\Sigma}^{\mathbf{G}}$, $\boldsymbol{\Sigma}^{\mathbf{E}}$ and $\boldsymbol{\mu}$ as given in equations (4)–(6) in Section 3. Note that the $\boldsymbol{\Sigma}^{\mathbf{E}}$ here is equivalent to the $\mathbf{R}^{\mathbf{E}}$ in equation (5) as $\boldsymbol{\Sigma}^{\mathbf{E}}$ is assumed to have unit variances in the main paper. ■

Remark**:** When $\mathbf{\Sigma}^{\mathrm{G}}$ and $\mathbf{\Sigma}^{\mathrm{E}}$ are not invertible, we can use their pseudo inverses to estimate the parameters by equations (4)–(6). Though the MLEs of the parameters will not be unique in this case, they are nevertheless solutions that satisfy the MLE estimating equations.

APPENDIX E: PROOF THAT THE MLEs IN EQUATIONS (4)–(6) ARE CONSISTENT ESTIMATORS

Theorem S3. Let $\mathbf{X}$ be a random matrix satisfying $\operatorname{vec}\left(\mathbf{X}\right) \sim N\left(\operatorname{vec}(\boldsymbol{\mu}\mathbf{1}^{\mathrm{T}}),\ \mathbf{R}^{\mathrm{E}} \otimes \mathbf{\Sigma}^{\mathrm{G}}\right)$. Let $\mathbf{X_1},\ \ldots,\mathbf{X_m}$ be *i.i.d.* observations of $\mathbf{X}$. Then when the covariance matrices are invertible, the MLEs of $\mathbf{\Sigma}^{\mathrm{G}}$, $\mathbf{R}^{\mathrm{E}}$ and $\boldsymbol{\mu}$ are consistent estimators.

Note that $\mathbf{R}^{\mathrm{E}}$ is equivalent to $\mathbf{\Sigma}^{\mathrm{E}}$, as $\mathbf{\Sigma}^{\mathrm{E}}$ is assumed to have unit variances in Section 3.

Proof. First, we derive the MLEs of $\mathbf{\Sigma}^{\mathrm{G}}$, $\mathbf{R}^{\mathrm{E}}$ and $\boldsymbol{\mu}$ using similar arguments as in the previous section. Given *i.i.d.* observations $\mathbf{X_1},\ \ldots,\mathbf{X_m}$, the log-likelihood function is

$$
\begin{aligned}
&l(\mathbf{X_1},\ldots,\mathbf{X_m};\boldsymbol{\mu},\mathbf{R}^{\mathrm{E}},\mathbf{\Sigma}^{\mathrm{G}}) \\
&\quad = -\frac{mp}{2}\log|\mathbf{\Sigma}^{\mathrm{E}}| - \frac{mn}{2}\log|\mathbf{\Sigma}^{\mathrm{G}}| - \frac{1}{2}\sum\nolimits_{k=1}^{m} tr\left(\left(\mathbf{X_k} - \boldsymbol{\mu}\mathbf{1}^{\mathrm{T}}\right)\left(\mathbf{R}^{\mathrm{E}}\right)^{-1}\left(\mathbf{X_k} - \boldsymbol{\mu}\mathbf{1}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{\Sigma}^{\mathrm{G}}\right)^{-1}\right).
\end{aligned}
$$

Equating the first partial derivatives with respect to $\mathbf{\Sigma}^{\mathrm{G}}$, $\mathbf{R}^{\mathrm{E}}$ and $\boldsymbol{\mu}$ to zero and solving the equations for $\mathbf{\Sigma}^{\mathrm{G}}$, $\mathbf{R}^{\mathrm{E}}$ and $\boldsymbol{\mu}$, we then have the following MLEs:

$$
\begin{aligned}
\hat{\mathbf{\Sigma}}^{\mathrm{G}} &= \frac{1}{mn}\sum\nolimits_{k=1}^{m}\left(\mathbf{X}_k - \boldsymbol{\mu}\mathbf{1}^T\right)\ \left(\mathbf{R}^{\mathrm{E}}\right)^{-1}\ \left(\mathbf{X}_k - \boldsymbol{\mu}\mathbf{1}^T\right)^T; \\
\hat{\mathbf{R}}^{\mathrm{E}} &= \frac{1}{mp}\sum\nolimits_{k=1}^{m}\left(\mathbf{X}_k - \boldsymbol{\mu}\mathbf{1}^T\right)^T\ \left(\mathbf{\Sigma}^{\mathrm{G}}\right)^{-1}\ \left(\mathbf{X}_k - \boldsymbol{\mu}\mathbf{1}^T\right);, \\
\hat{\boldsymbol{\mu}} &= \frac{1}{m}\frac{\sum\nolimits_{k=1}^{m}\mathbf{X}_k\ \left(\mathbf{R}^{\mathrm{E}}\right)^{-1}\ \mathbf{1}}{\mathbf{1}^T\ \left(\mathbf{R}^{\mathrm{E}}\right)^{-1}\ \mathbf{1}}.
\end{aligned}
$$

Next, we prove that these MLEs are consistent estimators. By the arguments in Section 3, we can express $\mathbf{X}$ as $\mathbf{X} = \boldsymbol{\mu}\mathbf{1}^{\mathrm{T}} + \mathbf{U}\ \mathbf{D}^{1/2}\ \mathbf{\Lambda}\ \mathbf{P}^{1/2}\ \mathbf{V}^{\mathrm{T}}$ when we ignore the random noises, where $\mathbf{\Lambda}$ is a matrix of *i.i.d.* elements with mean zero and unit variances. Then we have

$$
\begin{aligned}
\hat{\boldsymbol{\mu}} &= \frac{1}{m}\frac{\sum_{k=1}^{m}\mathbf{X}_k\left(\mathbf{R}^{\mathbf{E}}\right)^{-1}\mathbf{1}}{\mathbf{1}^T\left(\mathbf{R}^{\mathbf{E}}\right)^{-1}\mathbf{1}} \\
&= \frac{1}{m}\frac{\sum_{k=1}^{m}\left(\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}+\mathbf{U}\,\mathbf{D}^{1/2}\,\boldsymbol{\Lambda}_{\mathbf{k}}\,\mathbf{P}^{1/2}\,\mathbf{V}\right)\left(\mathbf{R}^{\mathbf{E}}\right)^{-1}\mathbf{1}}{\mathbf{1}^T\left(\mathbf{R}^{\mathbf{E}}\right)^{-1}\mathbf{1}} \qquad \text{(A3)} \\
&= \boldsymbol{\mu}+\frac{\mathbf{U}\,\mathbf{D}^{1/2}\left(\frac{1}{m}\sum_{k=1}^{m}\boldsymbol{\Lambda}_{\mathbf{k}}\right)\mathbf{P}^{1/2}\,\mathbf{V}\left(\mathbf{R}^{\mathbf{E}}\right)^{-1}\mathbf{1}}{\mathbf{1}^T\left(\mathbf{R}^{\mathbf{E}}\right)^{-1}\mathbf{1}}.
\end{aligned}
$$

From the last line in equation (A3), it is clear that $\hat{\boldsymbol{\mu}}$ is a consistent estimator of $\boldsymbol{\mu}$ since when m goes to infinity, $\frac{1}{m}\sum_{k=1}^{m}\boldsymbol{\Lambda}_{\mathbf{k}}\to\mathbf{0}$ in distribution.

Finally we prove the consistency properties of $\hat{\boldsymbol{\Sigma}}^{\mathbf{G}}$ and $\hat{\mathbf{R}}^{\mathbf{E}}$. By $\mathbf{X}=\boldsymbol{\mu}\mathbf{1}^{\mathbf{T}}+\mathbf{U}\,\mathbf{D}^{1/2}\,\boldsymbol{\Lambda}\,\mathbf{P}^{1/2}\,\mathbf{V}^{\mathbf{T}}$, we have

$$
\begin{aligned}
\hat{\boldsymbol{\Sigma}}^{\mathbf{G}} &= \frac{1}{mn}\sum_{k=1}^{m}(\mathbf{X}_{\mathbf{k}}-E(\mathbf{X}_{\mathbf{k}}))\left(\mathbf{R}^{\mathbf{E}}\right)^{-1}(\mathbf{X}_{\mathbf{k}}-E(\mathbf{X}_{\mathbf{k}}))^{\mathbf{T}} \\
&= \frac{1}{n}\mathbf{U}\mathbf{D}^{1/2}\left(\frac{1}{m}\sum_{k=1}^{m}\boldsymbol{\Lambda}_{\mathbf{k}}\,\boldsymbol{\Lambda}_{\mathbf{k}}^{\mathbf{T}}\right)\mathbf{D}^{1/2}\mathbf{U}^{\mathbf{T}}.
\end{aligned} \qquad \text{(A4)}
$$

When m goes to infinity, it is clear that $\frac{1}{n}\left(\frac{1}{m}\sum_{k=1}^{m}\boldsymbol{\Lambda}_{\mathbf{k}}\boldsymbol{\Lambda}_{\mathbf{k}}^{\mathbf{T}}\right)\to\mathbf{I}_{\mathbf{p}}$ in distribution, and therefore $\frac{1}{n}\mathbf{U}\mathbf{D}^{1/2}\left(\frac{1}{m}\sum_{k=1}^{m}\boldsymbol{\Lambda}_{\mathbf{k}}\,\boldsymbol{\Lambda}_{\mathbf{k}}^{\mathbf{T}}\right)\mathbf{D}^{1/2}\mathbf{U}^{\mathbf{T}}\to\mathbf{U}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{U}^{\mathbf{T}}=\Sigma^{\mathrm{G}}$ in distribution. So $\hat{\boldsymbol{\Sigma}}^{\mathbf{G}}\to\boldsymbol{\Sigma}^{\mathbf{G}}$ in distribution as m goes to infinity, or equivalently, $\hat{\boldsymbol{\Sigma}}^{\mathbf{G}}$ is a consistent estimator of $\boldsymbol{\Sigma}^{\mathbf{G}}$.

Using similar arguments as above, we can prove that $\hat{\mathbf{R}}^{\mathbf{E}}$ in equation (5) is a consistent estimator of $\mathbf{R}^{\mathbf{E}}$. ■

REFERENCES

1. Brown, M. P. S, Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Furey, T. S., Jr. Ares, M. & Haussler, D. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 262–267.
2. Yeung, K. Y. & Ruzzo, W. L. (2001) *Bioinformatics* **17(9)**, 763–774.
3. Fraley, C. & Raftery, A. E. (2000) University of Washington, Center for Statistics, and the Social Sciences, Working paper No. 11.
4. Ramoni, M., Sebastiani, P. & Cohen , P. R. (2002) *Machine Learning* **47(1)**, 91 – 121.
5. Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N. & Davidson, G. S. (2001) *Science* **293(5537)**, 2087 – 2092.
6. Zhou, X. J., Kao, M. C., Huang, H., Wong, A., Nunez–Iglesias, J., Primig, M., Aparicio, O. M., Finch, C. E., Morgan, T. E. & Wong, W. H. (2005) *Nature Biotechnology* **23(2)**, 238–243.
7. Li, K–C. (2002). *Proc. Natl. Acad. Sci. USA* **99(26),** 16875–16880.
8. Sabet, N., Volo, S., Yu, C., Madigan, J. P. & Morse, R. H. (2004) *Mol. Cell. Biol.* **24(20)**, 8823 –8833.
9. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. & Speed, T. P. (2003) *Nucleic Acids Res.* **31(4)** e15.
10. Deckert, J. & Struhl, K. (2002) *Mol. Cell Biol.* **22(18)**, 6458–6470.
11. Shimizu, M., Takahashi, K., Lamb, T. M., Shindo, H. & Mitchell, A. P. (2003) *Nucleic Acids Res.* **31(12)**, 3033–3037.
12. Kumar, R., Reynolds, D. M., Shevchenko, A., Shevchenko, A., Goldstone, S. D. & Dalton, S. (2000) *Curr. Biol.* **10(15)**, 896–906.
13. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar–Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K. & Young, R. A. (2002) *Science* **298(5594)**, 799–804.
14. Lund, R., Ahlfors, H., T., Kainonen, E., Lahesmaa, A. M. Dixon, C. & Lahessmaa, R. (2005) *Eur. J. Immunol.* **35(11)**, 3307–3319.
15. Chang, I. Y., Kim, S. H., Cho, H. J., Lee, D. Y., Kim, M. H., Chung, M. H. & You, H. J. (2005) *Nucleic Acids Res.* **33(16)**, 5073–5081.
16. Carreira, S., Goodall, J., Aksan, I., La Rocca, S. A., Galibert, M. D., Denat, L., Larue, L. & Goding, C. R.(2005) *Nature* **433(7027)**, 764–769.

Tibshirani, R. (1996) *Journal of Royal Statistical Society* **B(58)**, 267–288.

TABLES

Table 1. Percentages of known functionally related gene pairs among the top ranking gene pairs identified by Knorm correlation and the Pearson approach for the yeast and human microarray datasets separately. The gene pairs are ranked based on the absolute values of the correlation estimates.

No. of top ranking gene pairs	Yeast microarray dataset		Human microarray dataset	
	Knorm correlation	Pearson coefficient	Knorm correlation	Pearson coefficient
Top 10	30.0	10.0	10.0	10.0
Top 30	36.7	20.0	13.3	3.3
Top 50	38.0	26.0	8.0	4.0
Top 100	27.0	21.0	5.0	2.0
Top 200	23.0	21.0	3.5	3.0
Top 300	22.0	20.3	4.7	3.7
Top 400	22.3	21.8	4.0	3.8
Top 500	21.8	21.8	4.2	3.4
Top 1000	20.0	19.5	3.3	3.1

FIGURES, FIGURE TITLES AND LEGENDS

Figures in the main paper

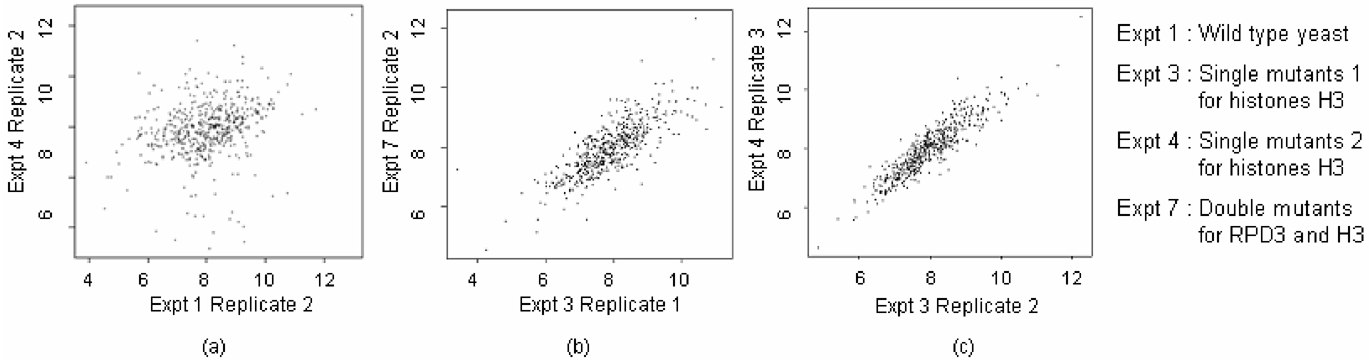

Fig. 1. Scatter plots of gene expressions of 532 GO annotated yeast genes (with high expression variations across experiments) between different experiments in a yeast histone mutation dataset. Axes represent gene expression values.

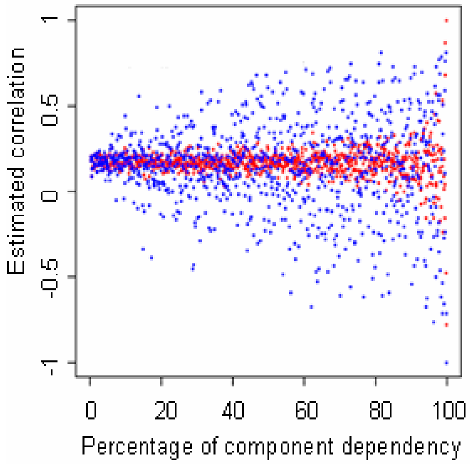

Fig. 2. Correlation estimates of two simulated vectors by Knorm correlation (in red) and Pearson coefficients (in blue) in the presence of vector component dependencies at different levels. X-axis indicates the dependency level; Y-axis represents the estimated correlation. The true correlation value is 0.17.

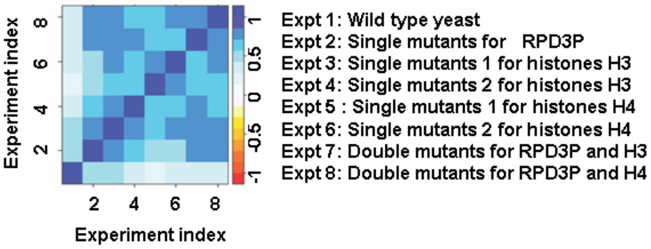

Fig. 3. Heatmap of our estimated experimental correlation matrix for the yeast dataset.

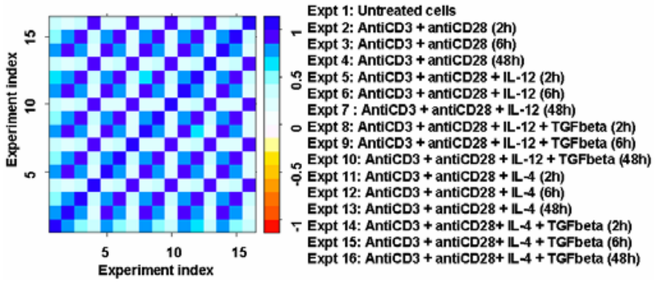

Fig. 5. Heatmap of our estimated experimental correlation matrix for the human microarray dataset.

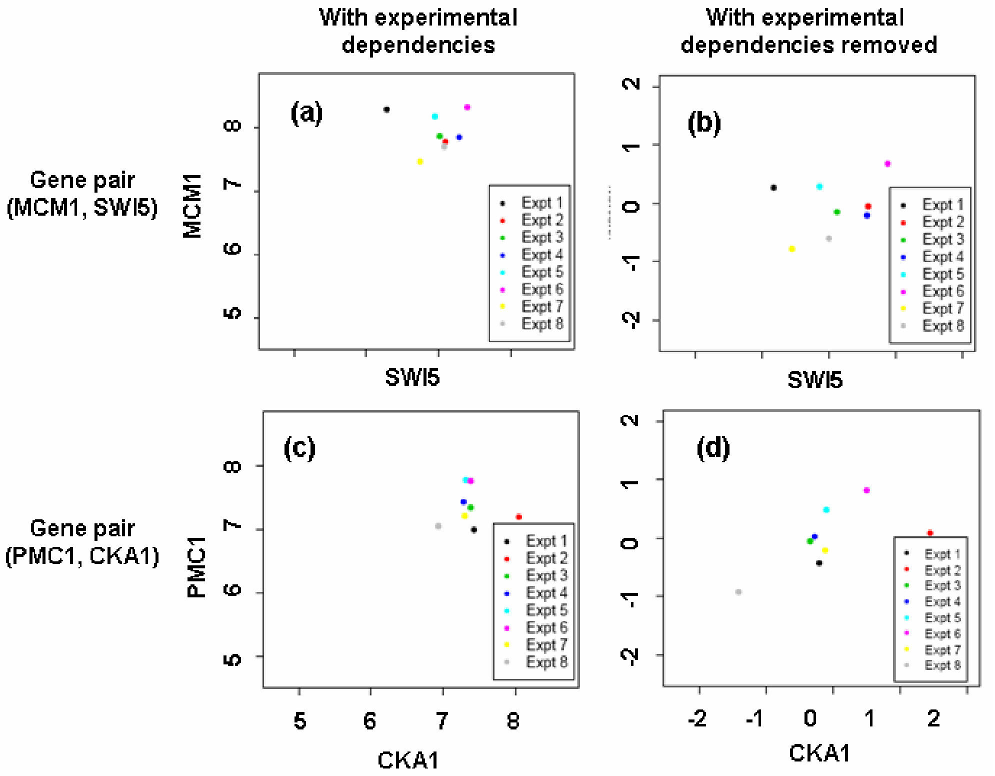

Fig. 4. Scatter plots of gene expressions for gene pairs (MCM1, SWI5) and (CKA1, PMC1), before and after the experimental dependencies are removed. Each of the two pairs is known to be functionally related. (a) and (c) are scatter plots for gene expressions **before** removing the experimental dependencies; each point represents the averaged expressions across replicates in each experiment. (b) and (d) are scatter plots for gene expressions **after** removing the experimental dependencies; each point represents the averaged centered transformed expressions across 500 bootstrapped replicates under the proposed approach. Axes represent normalized gene expressions. For genes MCM1 and SWI5, (b) clearly shows a positive correlation of 0.47 after removing the experimental dependencies, in contrast to the Pearson coefficient of –0.08 before removing the experimental dependencies in (a). Similarly, for genes CKA1 and PMC1, (d) clearly illustrates a positive correlation of 0.62 after experimental dependencies are removed in contrast to the Pearson coefficient of –0.08 with experimental dependencies.

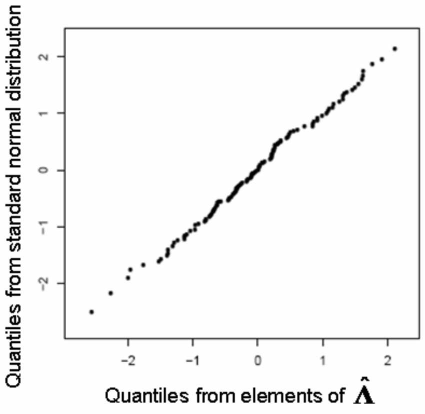

Fig. 6. QQ-plot of elements in $\hat{\mathbf{\Lambda}}$, estimated from a randomly selected expression matrix constructed through a bootstrapping procedure for the yeast dataset described in Section 3 against a standard normal distribution.

Figures in the Appendices

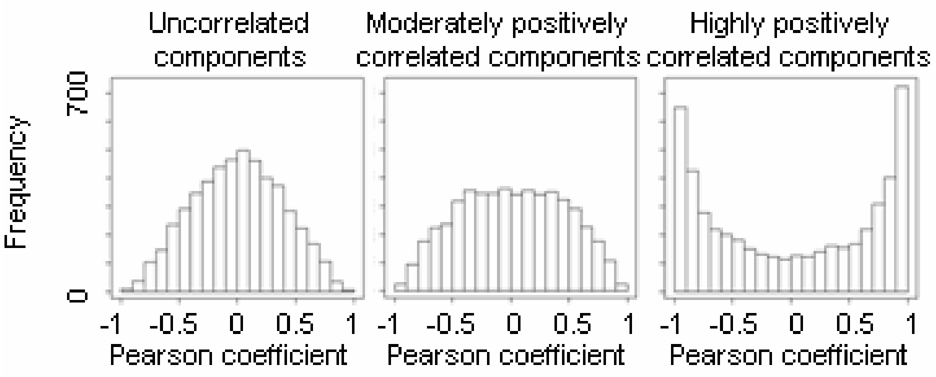

Fig. A1. Adverse impact of increasing component dependencies on the distribution of the Pearson coefficients for a pair of uncorrelated vectors. Each histogram consists of Pearson coefficients estimated from 5000 random pairs of uncorrelated vectors.

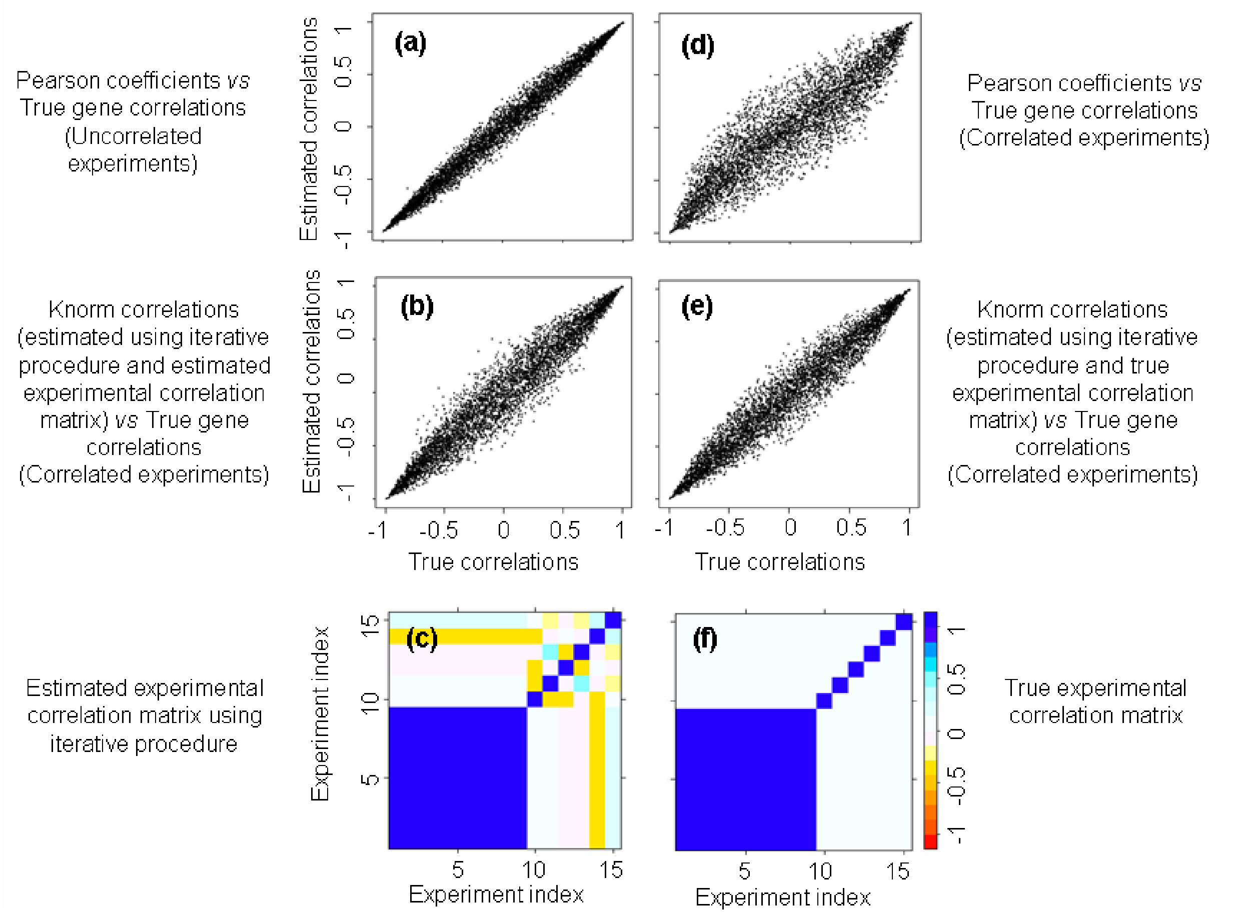

Fig. A2. Scatter plots comparing gene correlation estimates by our approach and Pearson approach to the true gene correlations. **(a)** Pearson coefficients vs. true correlations between genes for a simulated dataset with *un*correlated experiments. **(b)** Knorm correlation estimates (estimated using our estimated experimental correlation matrix, shown in (c), and iterative procedure) vs. true correlations between genes for a simulated dataset with correlated experiments. **(c)** Our iterative estimate of experimental correlation matrix for a simulated dataset with correlated experiments. **(d)** Pearson coefficients vs. true correlations between genes for a simulated dataset with correlated experiments. **(e)** Knorm correlation estimates (estimated using the *true* experimental correlation matrix, shown in (f), and iterative procedure) vs. true correlations between genes for a simulated dataset with correlated experiments. **(f)** *True* experimental correlation matrix for simulated datasets with correlated experiments. The same simulation dataset is used for (b), (c) and (e), and the same gene correlation matrix is used in all datasets for the plots.