UNIVERSITY OF WISCONSIN

DEPARTMENT OF BIOSTATISTICS AND MEDICAL INFORMATICS

Technical Report **April 2011**
 # 218

An Empirical Bayesian Approach for Identifying Differential Co-expression in High-throughput Experiments

John A Dawson
Christina Kendziorski

UNIVERSITY OF WISCONSIN
DEPARTMENT OF BIOSTATISTICS
AND MEDICAL INFORMATICS
K6/446 Clinical Science Center
600 Highland Avenue
Madison, Wisconsin 53792-4675
(608) 263-1706

An Empirical Bayesian Approach for Identifying Differential Co-expression in High-throughput Experiments

John A. Dawson
Department of Statistics
University of Wisconsin
Madison, Wisconsin 53706

Christina Kendziorski *
Department of Biostatistics and Medical Informatics
University of Wisconsin
Madison, Wisconsin 53706
Phone: 608-262-3146
Fax: 608-265-7916

*Corresponding author

An Empirical Bayesian Approach
for Identifying Differential Co-expression in High-throughput Experiments

Abstract

A common goal of microarray and related high-throughput genomic experiments is to identify genes that vary across biological condition. Most often this is accomplished by identifying genes with changes in mean expression level, so called differentially expressed (DE) genes, and a number of effective methods for identifying DE genes have been developed. Although useful, these approaches do not accommodate other types of differential regulation. An important example concerns differential co-expression (DC). Investigations of this class of genes are hampered by the large cardinality of the space to be interrogated as well as by influential outliers. As a result, existing DC approaches are often underpowered, exceedingly prone to false discoveries, and/or computationally intractable for even a moderately large number of pairs. To address this, an empirical Bayesian approach for identifying DC gene pairs is developed. The approach provides a false discovery rate (FDR) controlled list of significant DC gene pairs without sacrificing power. It is applicable within a single study as well as across multiple studies. Computations are greatly facilitated by a modification to the EM algorithm and a procedural heuristic. Simulations suggest that the proposed approach outperforms existing methods in far less computational time; and case study results suggest that the approach will likely prove to be a useful complement to current DE methods in high-throughput genomic studies.

Keywords: co-expression, differential expression, empirical Bayes, gene expression, meta-analysis, microarray

1 Introduction

A common goal of microarray and related high-throughput genomic experiments is to identify genetic signatures that provide insight into understanding, diagnosing and/or treating disease. A multitude of effective methods have been designed for this purpose, almost all of which focus on identifying genes or gene sets showing average expression levels that vary across biological condition. Applications of the most effective approaches for identifying so called differentially expressed (DE) genes or gene sets have proven useful (for a review, see Newton et al. (2007); Barry et al. (2008); Yakovlev et al. (2010)). However, in spite of notable successes, major insights have resulted far less frequently than expected (Pollack, 2007; Zilliox and Irizarry, 2007). This is due to the immense complexity of most diseases, and in particular to the fact that manifestation of disease can result from a de- or re-regulation of genes that does not significantly affect each gene's *average* expression level.

An important example is given in a study of endometrial cancer (Kato et al., 2003), where the expression of two genes known to be involved in cellular proliferation and genome replication (Ki-67 and MCM3, respectively) demonstrated significant positive co-expression in normal cells, but not cancer cells, suggesting a deregulation between the two genes that potentially results in cancer development or maintenance. The identification of Ki-67 would not have been made if only the average levels of expression had been considered, since Ki-67 abundance did not change between the two groups. Chan et al. (2000) highlight a similar result in a study of ovarian cancer, where no co-expression between BCL-2 and p53 expression was found in normal ovaries, but significant negative co-expression in malignant ovaries is evidenced. Another example concerns a study of cell cycle regulation in islet (Keller et al., 2008) where investigators showed that p16 and a group of cyclins (genes that control progression of cells through the cell cycle) are negatively co-expressed in lean mice, but positively co-expressed in obese mice suggesting a re-regulation of the cell cycle pathway

related to obesity. As in the other aforementioned examples, p16 and many of the cyclins were not shown to be DE between the lean and obese mice and would have therefore been missed had DE measures been applied in isolation. Numerous additional examples abound further suggesting that identification of other types of differential regulation, above and beyond DE measures, may increase one's ability to distinguish between groups and provide insight into their distinct etiologies (for a discussion and additional examples, see de la Fuente 2010). In particular, the identification of differentially co-expressed (DC) gene pairs may prove useful to this end (de la Fuente, 2010). As noted in de la Fuente (2010), the term co-expression often refers to some measure of correlation, and hereinafter we will use the term to refer specifically to Pearson's correlation unless otherwise noted.

The simplest methods for identifying DC gene pairs conduct pair-specific tests for selected pairs within a condition, identify those pairs that are strongly or significantly co-expressed, and define DC pairs as those co-expressed in one condition but not another. Approaches for doing so both within (Watson, 2006) and across (Choi et al., 2005) experiments have been developed. Although useful, these approaches sacrifice considerable power by conducting analyses separately within condition, they do not provide probabilistic statements regarding the likelihood that a particular pair is DC, and they cannot identify important types of DC pairs (e.g., those showing significant co-expression in both conditions that differs in magnitude or sign). These concerns are largely addressed by the approach of Lai et al. (2004) who propose a clever extension of the traditional F-test to accommodate not only changes in means but also correlations. The determination of exact thresholds is computationally prohibitive in their model, and as a result they propose an approach to approximate FDR, which is shown to be conservative in most cases. In addition, since the test statistic quantifies both DE and DC, selection of a pair provides no information about whether the pair is DE, DC, or both.

To address a number of these limitations, we here present an empirical Bayesian approach

for identifying DC gene pairs from a high-throughput experiment measuring expression in two or more conditions within a single study or across multiple studies. The approach provides an FDR controlled list of interesting pairs along with pair-specific posterior probabilities that can be used to identify particular types of DC. Section 2 details the underlying model and its assumptions with specific emphasis on computational efficiency and meta-analysis. The simulation studies presented in Section 3 suggest an improvement in power over comparable approaches with reasonable runtimes. Finally, case study results and a Discussion are presented in Sections 4 and 5, along with examples highlighting ways in which the derived posterior probabilities may be used in practice.

2 Methods

2.1 Description of the model

Consider normalized expression levels in a study indexed by s, profiled from m genes in n_s subjects, where the n_s subjects are partitioned into K conditions, each with n_s^k chips ($\sum n_s^k = n_s$). For every pair of genes, Fisher's Z-transformation (Fisher, 1928) is applied to sample correlations calculated within condition, $\rho_s^1, \ldots, \rho_s^K$, to yield $\vec{y_s} = (y_s^1, \ldots, y_s^K)$, where the pair subscript has for the moment been suppressed for simplicity of notation. As noted by Bartlett (1993), this transformation has several advantages, including symmetry, homogeneous and known variance, and approximate Normality when moderately large sample sizes are available. As a result, y_k, the transformed correlation for a pair within condition k, is assumed to arise from a Normal distribution with mean λ_s^k and variance $\frac{1}{n_s^k - 3}$.

The distribution of latent levels of correlation across pairs is also modeled in terms of Normal distributions, using the following density:

$$\psi_s(x) = \sum_{g=1}^{G_s} \left[w_{sg} \times \phi(\mu_{sg}, \tau_{sg}^2) \right] \tag{1}$$

This model specification accommodates fluctuation in the latent levels of correlation across pairs and allows for information sharing across pairs as well as conditions within the study. In practice, the one-component distribution is often too simplistic to describe the data while distributions with needlessly many components increase runtime without an accompanying increase in performance. Therefore, we will only consider $1 \leq G_s \leq 3$.

Of primary interest is identifying those pairs for which λ_s^k differs across conditions, or, more generally, defining the DC class. For example, when $K = 2$, there is a single way in which a pair could be classified as DC ($\lambda_s^1 \neq \lambda_s^2$), referred to hereinafter as a DC class. When $K = 3$, there are 4 DC classes: $\lambda_s^1 \neq \lambda_s^2 = \lambda_s^3$; $\lambda_s^2 \neq \lambda_s^1 = \lambda_s^3$; $\lambda_s^3 \neq \lambda_s^1 = \lambda_s^2$; and a DC class where λ_s^1, λ_s^2 and λ_s^3 are all distinct. The number of EC/DC classes (there is always a single EC class) increases with increasing K, as prescribed by the Bell exponential number (Bell, 1934); L will be used to refer to this quantity and the EC/DC classes will be indexed by $l = 1, \ldots, L$.

The preceding yields the following multivariate setup for a correlation vector $\vec{y_s} = (y_s^1, \ldots, y_s^K)$ in study s:

$$\vec{y_s} | \vec{\lambda_s} \sim MVN(\vec{\lambda_s}, \Sigma_s) \tag{2}$$

$$\vec{\lambda_s} | l, \{\mu_{sg}\}, \{\tau_{sg}\}, \{w_{sg}\} \sim \sum_{g=1}^{G_s} \left[w_{sg} \times MVN(\vec{\mu_{sg}}, \Xi_{sg}^l) \right] \tag{3}$$

where $\vec{\lambda_s} = (\lambda_s^1, \ldots, \lambda_s^K)$, the K-vector $\vec{\mu_{sg}} = (\mu_{sg}, \ldots, \mu_{sg})$, Σ_s is a K-by-K diagonal matrix with diagonal entries $d_{kk} = \frac{1}{n_s^k - 3}$, and the Ξ_{sg}^l are (possibly singular) matrices particular to the EC/DC class of the gene pair in question. For example, when there are two conditions

the Ξ^1_{sg} (EC case) and Ξ^2_{sg} (DC case) are respectively given by

$$\Xi^1_{sg} = \begin{pmatrix} \tau^2_{sg} & \tau^2_{sg} \\ \tau^2_{sg} & \tau^2_{sg} \end{pmatrix} \qquad and \qquad \Xi^2_{sg} = \begin{pmatrix} 0 & \tau^2_{sg} \\ \tau^2_{sg} & 0 \end{pmatrix}$$

Combining Equations (2) and (3) gives the joint conditional distribution of $\vec{y_s}$ and $\vec{\lambda_s}$ under a specific EC/DC class l and in turn the intermediate marginal distribution of $\vec{y_s}$ for that class can be obtained by integrating over $\vec{\lambda_s}$. Since the specified model is a mixture of conjugate quantities and hence conjugate itself, as Σ_s and all Ξ^l_{sg} are known, the intermediate marginal result is immediate:

$$\vec{y_s}|l, \{\mu_{sg}\}, \{\tau_{sg}\}, \{w_{sg}\} \sim \sum_{g=1}^{G_s} \left[w_{sg} \times MVN(\vec{\mu_{sg}}, \Sigma_s + \Xi^l_{sg}) \right] \tag{4}$$

It is worth noting that the density of a $MVN(\vec{\mu_{sg}}, \Sigma_s + \Xi^l_{sg})$ can be evaluated as the product of the densities of one or more multivariate Normals with a particular covariance structure, $D + uu'$, where D is a diagonal J-by-J matrix ($J \leq K$) and u is a J-vector, both containing only positive entries. This fact contributes greatly to computational efficiency, since any particular multivariate Normal density with covariance as given above can be stated without using determinants or inverses, via application of the Matrix Determinant Lemma (Harville, 1997) and the Sherman-Morrison formula (Bartlett, 1951) and hence evaluated using only linear algebra. To highlight this point (and simplify notation a bit later on) the likelihood obtained from Equation (4) is stated as

$$L(\vec{y_s}|l, \{\mu_{sg}\}, \{\tau_{sg}\}, \{w_{sg}\}) = f_l(\vec{y_s}; \{\mu_{sg}\}, \{\tau_{sg}\}, \{w_{sg}\}) \tag{5}$$

where f_l refers to the appropriate mixture of products over multivariate Normal densities with covariance structure $D + uu'$ for study s.

2.2 Combining information from multiple studies

Suppose that we have S studies, indexed by s. Furthermore assume that they contain expression information pertaining to the same m genes over K conditions. For a given gene pair, it is assumed that it follows one pattern with respect to the K conditions under consideration; that is to say, it belongs to a particular EC/DC class l.

Let the (K×S)-vector $\vec{y}$ refer to the concatenation of the K-vectors $\vec{y_s}$ in order. Given a class l, the $\vec{y_s}$ are assumed to arise from their study-specific distributions in a (conditionally) independent manner. Under this assumption, we may take products of (5) to obtain:

$$L(\vec{y}|l, \{\mu_{sg}\}, \{\tau_{sg}\}, \{w_{sg}\}) = \prod_{s=1}^{S} f_l(\vec{y_s}; \{\mu_{sg}\}, \{\tau_{sg}\}, \{w_{sg}\}) \tag{6}$$

Note that the product over studies of f_l is the same for all gene pairs and hence for all correlation vectors $\vec{y}$ within an EC/DC class. Assuming the $\vec{y}$ arise from a mixture over EC/DC classes, with mixing proportions $\pi_1, ..., \pi_L$, $\sum_{l=1}^{L} \pi_l = 1$ and defining $\vartheta = (\{\mu_{sg}\}, \{\tau_{sg}\}, \{w_{sg}\})$ and $\theta = (\pi_1, ..., \pi_L, \vartheta)$, the observed likelihood is derived:

$$L(\vec{y}; \theta) = \sum_{l=1}^{L} \left[\pi_l \times \prod_{s=1}^{S} f_l(\vec{y_s}; \vartheta) \right] \tag{7}$$

The vectors of transformed correlations, $\vec{y}$, are assumed to arise from Equation (7) in an independent manner, conditional on the system-wide hyperparameters contained in θ. While this is obviously not true in practice, as pairs that involve common genes are not independent, correlations among such pairs are less dependent than the genes contained in the pair, and are therefore only strongly dependent when the genes in question are strongly correlated (Langford et al., 2001). As a result, this violation is less severe than that made in many gene specific analyses (Smyth, 2005; Kendziorski et al., 2003; Broet et al., 2002); and empirical results suggest that the violation is not severely detrimental in practice (see

Section 3).

Adding this assumption to Equation (7) and liberating the suppressed gene pair subscript $i = 1, ..., p$, where p is the number of pairs, yields

$$L(\vec{y_1}, \ldots, \vec{y_p}; \theta) = \prod_{i=1}^{p} \sum_{l=1}^{L} \left[\pi_l \times \prod_{s=1}^{S} f_l(\vec{y_{i_s}}; \vartheta) \right] \tag{8}$$

Note that when $S = 1$ this meta-analysis framework simplifies into a framework for DC analysis within a single study.

2.3 Parameter estimation

Consider the complete data likelihood for the model described above:

$$L_C(\theta; \vec{y_1}, \ldots, \vec{y_p}, \{z_{il}\}) = exp \left[\sum_{i=1}^{p} \sum_{l=1}^{L} \mathcal{I}\{z_{il} = 1\} \left(\log \pi_l + \sum_{s=1}^{S} \log f_l(\vec{y_{i_s}}; \vartheta) \right) \right] \tag{9}$$

where each $z_{il} \in \{0, 1\}$ denotes whether or not the true EC/DC class of pair i is l.

An Expectation-Maximization (EM) algorithm (Dempster et al., 1977) can be used to obtain estimates of $\theta = (\pi_1, ..., \pi_L, \mu_1, \ldots, \mu_S, \tau_1, \ldots, \tau_S)$ from the data. Specifically, on the t^{th} iteration, the E step consists of calculating the complete-data sufficient statistics $\{\omega_{il}\} = \{E[z_{il}]|\vec{y_1}, \ldots, \vec{y_p}, \theta^{(t)}\}$ for all i and l via Equation (5) and Bayes' theorem, under the current value of θ, $\theta^{(t)}$, in order to obtain the Q-function:

$$
\begin{aligned}
Q(\theta|\vec{y_1},\ldots,\vec{y_p},\theta^{(t)}) &= E_{\{z_{il}\}|\vec{y_1},\ldots,\vec{y_p},\theta^{(t)}}[\log L_C(\theta;\vec{y_1},\ldots,\vec{y_p},\{z_{il}\})] \\
&= \sum_{i=1}^{p}\sum_{l=1}^{L}\omega_{il}\left(\log\pi_l + \sum_{s=1}^{S}\log f_l(\vec{y_{i_s}};\vartheta)\right) \\
&= \sum_{i=1}^{p}\sum_{l=1}^{L}\omega_{il}\log\pi_l \\
&\quad + \sum_{i=1}^{p}\sum_{l=1}^{L}\omega_{il}\sum_{s=1}^{S}\log f_l(\vec{y_{i_s}};\vartheta) \qquad (10)
\end{aligned}
$$

We will use Ω to refer to the p-by-L matrix $\{\omega_{il}\}$.

The M step then maximizes the Q-function for θ. In this context, maximization can be divided into two parts as Equation (10) is separable into two summands, one a function of the mixing proportions and the other a function of ϑ. Closed-form maximizers $\hat{\pi}_1, ..., \hat{\pi}_L$ are obtained by taking column-wise means of Ω. Closed-form solutions for the constituents of ϑ are not available, but can be obtained numerically, subject to the constraints that $\forall s$, G_s is known, $\tau_{gs} > 0$, and $\sum_g w_{sg} = 1$. In the current implementation, **bobyqa** in *R/minqa* is used (Powell, 2009) for this numerical optimization.

For each study s, knowledge of G_s is essential to proper execution of the framework; good initial estimates of ϑ are also helpful to speed convergence. Both are estimated from the data in this empirical Bayesian framework. While a number of methods may be used for this purpose, we prefer the M-clust algorithm (Fraley and Raftery, 2002, 2006) as implemented in *R/mclust*: When properly queried using the transformed correlations from the data as inputs, **M-clust** will return good values for G_s, the $\{\mu_{sg}\}$ and the $\{\tau_{sg}\}$; **M-clust** also returns a mixture component classification and a measure of classification uncertainty for each datum, which can be used to estimate the w_{sg} via averaging the uncertainty values.

While the use of the Fisher Z-transform along with assumptions of parametric forms and

conditional independence of pairs produce likelihoods that can be quickly evaluated, there is still a considerable computational burden for even modestly large m, as the number of pairs is quadratic in the number of genes; the M step is by far the most expensive in terms of runtime. Noting this, as well as the fact that the mixing proportions take much longer to converge than the estimates of the members of ϑ, a modification to the standard EM is used, as detailed below. Also provided is a heuristic which can be used to further cut runtime.

2.3.1 A special case of the two-cycle alternating ECM

This modified EM consists of running an E step to calculate the Q-function, then calling the numerical optimizer to update ϑ. Then, a cycle of calculating the E step and maximizing to obtain estimates of the mixing proportions, $\hat{\pi}_1, ..., \hat{\pi}_L$, keeping the current estimate of ϑ fixed, is performed until estimates of the mixing proportions converge. This process replaces a single iteration of the standard EM, and repeats until convergence of θ is achieved.

In other words, one iteration of the standard EM is replaced by:

1. Calculate $\{\omega_{il}\}$ given $\vec{\pi}$ and ϑ.
2. Maximize Q w.r.t ϑ given the $\{\omega_{il}\}$ (note that due to the separable nature of Q, the derived argmax does not depend on $\vec{\pi}$). Update the members of ϑ with their respective components of the joint argmax.
3. Calculate $\{\omega_{il}\}$ given $\vec{\pi}$ and this updated ϑ.
4. Maximize Q w.r.t $\vec{\pi}$ given $\{\omega_{il}\}$, (note that this step likewise does not depend on ϑ). Update $\vec{\pi}$ with the argmax.
5. Repeat 3 and 4 until $\vec{\pi}$ converges

Technically speaking, this modification to the standard EM is not novel. It can be construed as an incarnation of the two-cycle alternating expectation-conditional maximization

(TCA-ECM) algorithm presented by Meng and van Dyk in 1997 (personal communication). However, it is a special case of that rather general framework and can dramatically reduce runtimes if the following conditions hold:

1. Good initial estimates for one subset of θ may be computed using the data (and hence estimates for this subset converge quickly during the EM), while only poor or uninformed estimates exist for the other, and

2. The M step is computationally cheap for the latter subset (e.g., closed-form maximizers exist) but expensive for the former (e.g., numerical optimization is required).

When these conditions hold, this modification to the standard EM (referred to hereinafter as the TCA-ECM) requires fewer iterations than the standard EM, is considerably faster, and provides the same estimates of θ, up to specified convergence and iteration tolerances (simulations not shown; see also Meng and van Dyk (1997)).

2.3.2 A helpful heuristic

To further reduce runtime, one can use a random subset of the data (e.g., 0.1% of all gene pairs) to perform computations related to ϑ, specifically while either calling **M-clust** or the pertinent portion of the TCA-ECM. This heuristic is exceedingly beneficent when the number of pairs is large, since the number of data points p (which number in the hundreds of thousands if not millions when $m > 500$) is far greater than the number of free parameters being estimated. The validity and efficacy of this heuristic will be illustrated in Section 3.

3 Simulations

The proposed methodology provides a way to identify DC gene pairs, but it relies on numerical approximation methods and it assumes conditions that are never fully satisfied in practice

(e.g., assumptions of conditional independence). To assess the methodology we performed a small set of simulation studies. These provide some, albeit limited, insight into the quality of parameter estimates from the TCA-ECM algorithm and associated heuristics, potential gains in computational time, and how violations of assumptions affect inference. Perhaps most importantly, the simulation results also provide information on error rates related to DC inference and facilitate a comparison to related approaches.

We consider three scenarios. The first simulation (SIM I) is designed to assess relative performance among the modified EM algorithms under ideal conditions (i.e., the transformed correlations are drawn and made noisy under the model described in Section 2.1). A simulated data set consists of 10,000 observations simulated from a Normal-Mixture-Normal model with varying ϑ. There are five of these and their descriptions are given below; see also Web Figures 1a-1e.

A1: Basic Normal A single Normal distribution: $N(0, 0.3^2)$

B1: Twin Peaks An even mixture of a $N(-0.1, 0.1^2)$ and a $N(0.2, 0.1^2)$

B2: SIM II-ish A 9-1 mixture of a $N(0, 0.2^2)$ and a $N(0.5, 0.1^2)$

C1: Tri Peaks An even mixture of a $N(-0.5, 0.1^2)$, a $N(0, 0.1^2)$ and a $N(0.5, 0.1^2)$

C2: Shrugging Normal A 1-2-1 mixture of a $N(-0.3, 0.1^2)$, a $N(0, 0.1^2)$ and a $N(0.3, 0.1^2)$

The next two sets of simulations (SIM II, SIM III) are designed to assess performance under more realistic conditions. A single dataset in SIM II contains three groups of 100 genes, simulated in each of two conditions. Within each group and condition, the genes are all correlated; genes in different groups are uncorrelated. Two covariance matrices, one for each of two conditions, are created such that the strengths of the correlations in the first group are not the same between conditions, but all other correlations are unchanged.

Under this setup all gene pairs involving two genes from the first group (i.e., 1-1, 1-2, ..., 99-100, 100-100) are DC, while all other pairs are EC. A single simulation consists of 200 chips, 100 in each condition; data are drawn from a multivariate Normal distribution with mean zero and covariance as dictated by condition. SIM III is identical except that the three groups now have 1000, 1000 and 2000 genes, respectively, rather than each having 100 genes as in SIM II. As in SIM I, twenty simulations are considered in each; further details for the setup of SIM II and SIM III are given in Web Appendix A. For each simulated data set, Fisher z-values were obtained from correlations calculated using the biweight midcorrelation which was used to minimize the effect of potential outliers (Wilcox, 1997).

In SIM II, the proposed model is fit using the one-step version of the TCA-ECM applied to the full data set. A gene pair is identified as DC under a soft thresholding mechanism if the posterior probability of DC exceeds a critical value which controls the posterior expected false discovery rate at 5% (Berger, 1980, p. 164). A gene pair is identified as DC under a hard thresholding mechanism if the posterior probability of DC exceeds 0.95. This threshold conservatively controls the posterior expected false discovery rate at 5%. Both results are presented.

Our approach is compared to an FDR-controlled pairwise application of Box's M-test (Mardia et al., 1979) and to the ECF approach of Lai et al. (2004). For the M-test, the p-values obtained from each pair of genes are converted into q-values (Storey, 2002) and thresholded to get a list of pairs with FDR of 5%. In the ECF approach, the distribution from which the null is drawn is data-independent once the number of subjects (microarrays) in each condition is known. Using the code from the ECF website and details provided in personal communications, we simulate from this null one million times to obtain ECF thresholds corresponding to multiple-comparison adjusted p-values ranging from $p = 10^{-1}$ to 10^{-4}; particular thresholds corresponding to 5% and 10% FDR control as derived in (Lai et al., 2004) are among those used (see Table 2).

3.1 Results

Table 1 provides timing, parameter and deviance estimates derived from data generated under SIM I. Deviance is defined here as $1000 \times ||f_t - f_e||_2$, where f_t and f_e are the true and estimated densities for the distribution from which the transformed correlations are generated; this is done in order to compare the estimated ϑ to the truth in situations where a model of different complexity is deemed best for the simulated data (e.g., a two-component f_e is chosen when f_t truly uses three components). Averages across 20 data sets are shown with standard deviations given in parentheses. The results indicate that the heuristic versions of the TCA-ECM provide performance and accuracy that are very close to those obtained from the full TCA-ECM in a fraction of the time.

Power, FDR, and runtime for the proposed approach, the ECF procedure of Lai et al. (2004) and Box's M-test (Mardia et al., 1979) evaluated using the data from SIM II are shown in Table 2. The full TCA-ECM was not used for this simulation as it was not shown to be substantially superior to the one-step TCA-ECM in SIM I. The results suggest that the proposed approach has well controlled FDR, with power that is increased over that obtained from ECF for each of the thresholds considered (including one corresponding to $p = 10^{-1}$).

Since computation in the ECF approach is prohibitive over many simulations when p (the number of pairs) is relatively large, as in SIM III, in Web Table 1 we provide results from only a single simulation under the SIM III framework. Although results from a single simulation are clearly limited, we see that the results are consistent with those from SIM II (shown in Table 2). Web Table 1 also shows performance of the TCA-ECM restricted to 0.1% subsets of the pairs for parameter estimation. Note that for 4000 genes, this restriction still leaves $\sim$ 8000 pairs from which the relatively few parameters (there are at most 8 hyperparameters and mixing proportions) are estimated. Web Table 2 reports power, FDR, and runtime for results derived from 20 runs of SIM III for this restricted version of the TCA-ECM. These results suggest that the use of the restricted one-step TCA-ECM reduces

runtime considerably, and does not detrimentally impact observed FDR or power.

4 Prostate cancer case study

As an application of this approach, we considered three studies of prostate cancer for which microarray expression data was available for normal and diseased subjects. These studies will be referred to as the Monzon, Taylor, and Roth studies, respectively. They are described in some detail in Web Appendix B. In short, each study utilized a different Affymetrix microarray platform, and each dataset is available at the Gene Expression Omnibus (with GEO accession ids GSE6919, GSE21034, and GSE7307, respectively). The Monzon study considers samples from 18 normal and 65 diseased prostates; Taylor considers 29 normal and 150 diseased; Roth considers 7 normal and 17 diseased.

The three studies have 8,631 genes in common, many of which are homologues within gene families. From the 8,631, we selected 5,765 genes representing unique genes from the families. For the Monzon and Roth studies, any gene for which two or more probes existed was represented by a single probe, which was chosen by calculating the average intensity across all arrays within that study for all probes corresponding to that gene and then taking the probe with the median such average. Background correction of the intensities contained in the raw (.CEL) files was performed using RMA (Bolstad et al., 2003). Quantile normalization was not done due to issues concerning destruction of correlations across samples (see Discussion). Rather, chip-specific intensities were normalized to have the same median across studies.

With condition defined by disease status (K=2), the proposed approach was used to identify DC gene pairs separately for all three studies and together in a meta-analysis. Over 16.6 million gene pairs were considered. As in the simulations, the biweight midcorrelation (Wilcox, 1997) was used prior to Fisher's Z-transformation in order to minimize the effects of outliers.

4.1 Results

When analyzing the studies individually, 14,954 and 115,279 gene pairs were declared DC by the approach for the Monzon and Taylor studies, respectively; 408 of these pairs were in common across the two studies. No pairs were flagged in the much smaller Roth study. In contrast, the meta-analysis yields 141,678 DC gene pairs. This set contains the 408 gene pairs identified by both Monzon and Taylor separately; in addition, 8,055 of 14,954 and 97,064 of 115,279 pairs carried over their DC identification to the meta-analysis.

Figure 1 shows posterior probabilities of DC obtained from the Monzon and Taylor individual analyses for all 16.6 million pairs plotted against a white background. The color of each pair's point corresponds to the posterior probability of DC generated by the meta-analysis, ranging from blue (nil evidence of DC) to red (high evidence of DC). Dashed lines indicate the 0.95 cutoffs used in the individual analyses; those points boxed into the upper right-hand corner reflect the 408 pairs taken by both studies.

Lest the reader think that Figure 1 indicates that Roth's study did not contribute to the meta-analysis results, consider Web Figure 2. The structure is similar to Figure 1, except that now only those 141,678 pairs taken by the meta-analysis are plotted (in red). While the curve largely describes the results of Monzon and Taylor, note the handful of points that lie far beyond the curve. These pairs are taken by the meta-analysis due to the Roth study. As a point of emphasis, pairs for which the posterior probability of DC obtained from the Roth study is greater than 0.5 are circled in black.

Figure 2 shows two gene pairs identified by the meta-analysis as well as the separate Monzon and Tayor analyses. The plotted points are colored by condition, with non-cancerous subjects in purple and cancerous subjects in orange. A "robust" regression line (i.e., one based on only those points used internally by the biweight midcorrelation calculation) is overlaid for each condition as a visual aid. When viewing these regression lines as proxies for correlation, it is important to note tightness around the line as well as slope; however,

since these lines were fit using least squares, their trajectories are driven by vertical (y-axis) deviations. Web Figure 3 similarly highlights two other pairs chosen as DC by the meta-analysis, but neither of the individual studies. Although there appears to be a clear DC relationship, the individual studies are underpowered (relative to the meta-analysis) to identify these pairs as DC at an FDR of 5%.

5 Discussion

Understanding the genetic basis of disease requires identifying genes that are differentially regulated between healthy and affected conditions. For over a decade, thousands of investigations utilizing high-throughput expression data have focused on identifying DE genes. Although tremendously powerful in many settings, it is becoming increasingly clear that overlooking other types of differential regulation, such as DC, can be critically limiting and in some cases can lead to incorrect inference (Mentzen et al., 2009).

The empirical Bayesian approach presented here provides a much needed method for identifying DC pairs while controlling a specified FDR. The pair specific posterior probability distributions facilitate classification of each pair into its most likely EC/DC class. The approach does not restrict DC gene pairs to those that are highly correlated in at least one condition, it allows for the identification of DC pairs that change in magnitude but not sign across conditions, and it does not involve tests for DE. This last point is important since many of the best methods for normalization prior to a DE analysis change the correlation structure between genes in a significant way and are therefore not optimal when DC identification is of interest (Qiu et al., 2005). In other words, the most appropriate method for normalization depends in part on whether subsequent analysis involves test for DE or DC, and it is not yet clear how best to normalize measurements prior to applying methods that aim to do both simultaneously.

Although our approach does not involve identification of DE genes, the hierarchical framework presented here is conceptually similar to our previous work which proposed a log-normal normal hierarchical model for identifying DE genes (Kendziorski et al., 2003). A main difference is that here we introduce a more flexible prior which accommodates the structure of transformed correlations observed in practice. More importantly, the Normal observation component is used here to describe Fisher Z-transformed correlations calculated from gene pairs rather than log-transformed expression from individual genes as in a DE analysis. In addition to providing a more flexible model that can accommodate the distribution of co-expressions observed in practice, these differences have important implications computationally. In particular, estimation of hyperparameters via the EM algorithm as previously described becomes arduous in studies of co-expression even when the number of genes is modest (since all pairs are considered). To address this, we have proposed a modification to the EM algorithm referred to as the TCA-ECM along with a heuristic which provides reliable parameter estimates in substantially reduced computing time. As the conditions specified in Section 2.3.1 are not specific to this application, the TCA-ECM as implemented here should prove advantageous in other more general mixture model settings where the conditions hold.

As with any modeling framework, the proposed approach makes a number of assumptions that should be checked in practice. A main one is that transformed correlations are approximately Normal. The biweight midcorrelation or Spearman's correlation provide estimates that are largely robust to outliers; and the biweight midcorrelation was used here as it has been shown to be superior to Spearman's correlation in many regards (Wilcox, 1997). A second assumption concerns conditional independence of the correlations (Equation (8)), which is clearly violated in most cases. The simulation study suggests that this violation results in very slight increases in FDR with little change in power. Further work is required to more completely assess the impact on inference when the model assumptions are violated.

While the proposed methodology does not make use of (and is hence not constrained by)

previously defined sets of genes, the pair specific posterior probabilities of DC provided by the approach can augment an analysis that has identified a group of genes as interesting *a priori*. The upper panel of Figure 3 shows six genes identified in Gorlov et al. (2009) as being individually significant in the transition from normal prostate to localized prostate cancer. No information is provided in Gorlov et al. (2009) on the relationships among the genes, but as one can see in Figure 3, there are a number of interesting features among the pairs. Notably, SRM appears to be a DC hub within this gene set as it is DC with a number of other genes.

It may also be of interest to construct groups of genes strongly DC with a gene of interest. Such a gene could be identified *a priori* or could be chosen from a list rank ordered by overall DC. Take, for instance, PARM1. PARM1 is believed to play a role in prostate cancer progression, as it is thought to enable certain cells in the prostate to resist apoptosis (The Human Gene Compendium, 2011). Additionally, it is in the top 20 genes when one rank orders the genes by their upper 0.00001 quantile of meta-analysis posterior probabilities of DC, across all $m-1$ pairs involving themselves. The lower panel of Figure 3 shows twelve genes greedily chosen to form a DC subnetwork, using PARM1 as a seed. Specifically, each gene was added sequentially into the subnetwork by virtue of having the highest average meta-analysis posterior probability of DC with respect to pairs involving genes already in the subnetwork. The result is a novel subnetwork of genes that exhibits strong DC patterns among its members. The network shows strong correlation among members in the non-cancerous condition that is lost when cancer is present, suggesting a de-regulation among members. A number of prostate and/or cancer related genes are identified. Perhaps most interesting is TP53TG1, which has been shown to play an important role in signaling of TP53, the well known tumor suppressor gene (Takei et al., 1998).

In summary, it is becoming increasingly clear that important types of differential regulation are missed by traditional tests for DE genes; DC measures are one such type. The pro-

posed approach is computationally efficient, providing an FDR controlled list from roughly 16.6 million gene pairs in just under an hour on a standard DELL Xeon 5670 with 1600 Mhz and 48GB of RAM. It should prove to be a useful complement to a traditional DE analysis. However, unlike most DE methods, the approach utilizes correlations as opposed to gene specific expression measures, and as a result it is directly applicable to other types of high-throughput studies where correlations are of some interest. Integrating multiple types of high-throughput studies at once requires extending the framework. Current efforts in this direction are underway.

6 Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 3 and 4 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgements

The authors wish to thank Douglas Bates, Peter Qian, Michael Newton, Xiao-Li Meng, David van Dyk, Yinglei Lai, Hongyu Zhao, and Kevin Eng for conversations, correspondences and comments that helped to improve the manuscript.

References

W. T. Barry, A. B. Nobel, and F. A. Wright. A statistical framework for testing functional categories in microarray data. *Annals of Applied Statistics*, 2(1):286–315, 2008.

M. S. Bartlett. An inverse matrix adjustment arising in discriminant analysis. *Annals of Mathematical Statistics*, 22(1):107–111, 1951.

R. F. Bartlett. Linear modelling of pearson's product moment correlation coefficient: an application of fisher's z-transformation. *Journal of the Royal Statistical Society (Series D)*, 42(1):45–53, 1993.

E. T. Bell. Exponential numbers. *American Mathematics Monthly*, 41:411–419, 1934.

J. O. Berger. *Statistical Decision Theory and Bayesian analysis*. Springer, New York, 1980.

B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, 19(2):185–193, 2003.

P. Broet, S. Richardson, and F. Radvanyi. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology*, 9(4):671–683, 2002.

J. K. Choi, U. Yu, O. J. Yoo, and S. Kim. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21(24):4348–4355, 2005.

A. de la Fuente. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26(7):326–333, 2010.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society (Series B)*, 39(1):1–38, 1977.

R. A. Fisher. The general sampling distribution of the multiple correlation coefficient. *Journal of the Royal Statistical Society (Series A)*, 121(788):654–673, 1928.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

C. Fraley and A. E. Raftery. MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics, 2006. (revised 2009).

I. P. Gorlov, J. Byun, O. Y. Gorlova, A. M. Aparicio, E. Efstathiou, and C. J. Logothetis. Candidate pathways and genes for prostate cancer: a meta-analysis of gene-expression data. *BMC Medical Genomics*, 2(48), 2009.

D. A. Harville. *Matrix algebra from a statistician's perspective*. Springer-Verlag, New York, 1997.

K. Kato, T. Toki, M. Shimizu, T. Shiozawa, S. Fujii, T. Nakaido, and I. Konishi. Expression of replication-licensing factors MCM2 and MCM3 in normal, hyperplastic, and carcinomatous endometrium: correlation with expression of Ki-67 and estrogen and progesterone receptors. *Internation Journal of Gynecological Pathology*, 22(4):334–340, 2003.

M. P. Keller, YJ Choi, P. Wang, D. B. Davis, M. E. Rabaglia, A. T. Oler, D. S. Stapleton, C. Argmann, K. L. Schueler, S. Edwards, H. A. Steinberg, E. C. Neto, R. Kleinhanz, S. Turner, M. K. Hellerstein, E. E. Shadt, B. S. Yandell, C. Kendziorski, and A. D. Attie. A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Research*, 18(5):706–716, 2008.

C. M. Kendziorski, M. A. Newton, H. Lan, and M. N. Gould. On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22:3899–3914, 2003.

Y. Lai, B. Wu, L. Chen, and H. Zhao. A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics*, 20(17):3146–3155, 2004.

E. Langford, N. Schwertman, and M. Owens. Is the property of being positively correlated transitive? *The American Statistician*, 55(4):322–325, November 2001.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

X-L Meng and D. van Dyk. The EM algorithm – an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society (Series B)*, 59(3):511–567, 1997.

W. I. Mentzen, M. Floris, and A. de la Fuente. Dissecting the dynamics of dysregulation of cellular processes in mouse mammary glad tumor. *BMC Genomics*, 10(601), 2009.

M. A. Newton, F. A. Quintana, J. A. den Boon, S. Sengupta, and P. Ahlquist. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics*, 1(1):85–106, 2007.

J. R. Pollack. A perspective on dna microarrays in pathology research and practice. *American Journal of Pathology*, 171:375–385, 2007.

M. J. D. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report NA06, DAMTP, University of Cambridge, 2009.

X. Qiu, A. I. Brooks, L. Klebanov, and A. Yakovlev. The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics*, 6(120), 2005.

G. K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society (Series B)*, 62:479–498, 2002.

Y. Takei, S. Ishikawa, T. Tokino, T. Muto, and Y. Nakamura. Isolation of a novel tp53 target gene from a colon cancer cell line carrying a highly regulated wild-type tp53. *Genes Chromosomes Cancer*, 23(1), 1998.

The Human Gene Compendium. prostate androgen-regulated mucin-like protein 1. `http://genecards.org/cgi-bin/carddisp.pl?gene=PARM1`, 2011.

M. Watson. Coxpress: differential co-expression in gene expression data. *BMC Bioinformatics*, 7(509), 2006.

R. R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, San Diego, California, 1997.

A. Y. Yakovlev, L. Klebanov, and D. Gaile (Eds.). *Statistical Methods for Microarray Data Analysis*. Springer, New York, 2010.

M. J. Zilliox and R. A. Irizarry. A gene expression bar code for microarray data. *Nature Methods*, 4(11):911–913, 2007.

Table 1: SIM I

	A1	B1	B2	C1	C2
Time					
TCA-ECM	68 (3)	529 (84)	3600 (6366)	735 (108)	2933 (402)
1-step TCA-ECM	26 (3)	153 (2)	184 (22)	216 (1)	218 (3)
$\hat{\pi}_1$ ($\pi_1 = 0.95$)					
TCA-ECM	0.949 (0.003)	0.949 (0.003)	0.949 (0.004)	0.950 (0.002)	0.949 (0.003)
1-step TCA-ECM	0.949 (0.003)	0.948 (0.003)	0.948 (0.004)	0.949 (0.002)	0.948 (0.003)
Deviance †					
TCA-ECM	0.1 (0.1)	1.2 (0.8)	2.3 (0.7)	0.7 (0.4)	2.7 (2.0)
1-step TCA-ECM	0.1 (0.1)	1.1 (0.7)	0.5 (0.4)	0.8 (0.4)	10.5 (6.6)

Hyperparameters estimated using the full and one-step versions of the TCA-ECM under five different true distributions of transformed correlations (see Web Figure 1). Values shown are means calculated over 20 simulated data sets; standard deviations are shown in parentheses. Computational time is given in seconds.

† Defined as $1000 \times ||f_t - f_e||_2$ where f_t and f_e are the true and estimated densities for the distribution from which the transformed correlations are generated.

Table 2: SIM II

Approach	Obs. FDR	Obs. Power	Time *
1-step TCA-ECM (soft threshold)	0.054 (0.021)	0.952 (0.099)	549 (195)
1-step TCA-ECM (hard threshold)	0.0004 (0.0003)	0.869 (0.162)	549 (195)
ECF w/ $p = 10^{-1}$	0.277 (0.028)	0.932 (0.064)	134+ (1)
ECF w/ $p = 10^{-2}$	0.037 (0.012)	0.718 (0.141)	134+ (1)
ECF w/ $p = 10^{-3}$	0.006 (0.004)	0.452 (0.154)	134+ (1)
ECF w/ $p = 5 \times 10^{-3}$ †	0.004 (0.003)	0.381 (0.145)	134+ (1)
ECF w/ $p = 10^{-4}$ ‡	0.001 (0.001)	0.240 (0.116)	134+ (1)
Box's M-test	0.084 (0.035)	0.856 (0.067)	27 (1)

Average FDR and power from the proposed approach with hyperparameters estimated using the one-step versions of the TCA-ECM under soft and hard thresholding. Values are means calculated over 20 simulated data sets; standard deviations are shown in parentheses. Results from the ECF approach of Lai *et al.* (2004) and Box's M-test Mardia *et al.* (1979) are also shown. Computational time is given in seconds.

* Times for the ECF results do not include the runtime required for the simulation of the ECF null, which depends linearly on the number of null simulations. When using one million null simulations (as was the case here) this adds an additional 795 seconds to the ECF's runtime.

† This threshold is approximately 0.1/300, the value Lai et al. would use for 10% FDR control.

‡ This threshold is approximately 0.05/300, the value Lai et al. would use for 5% FDR control.

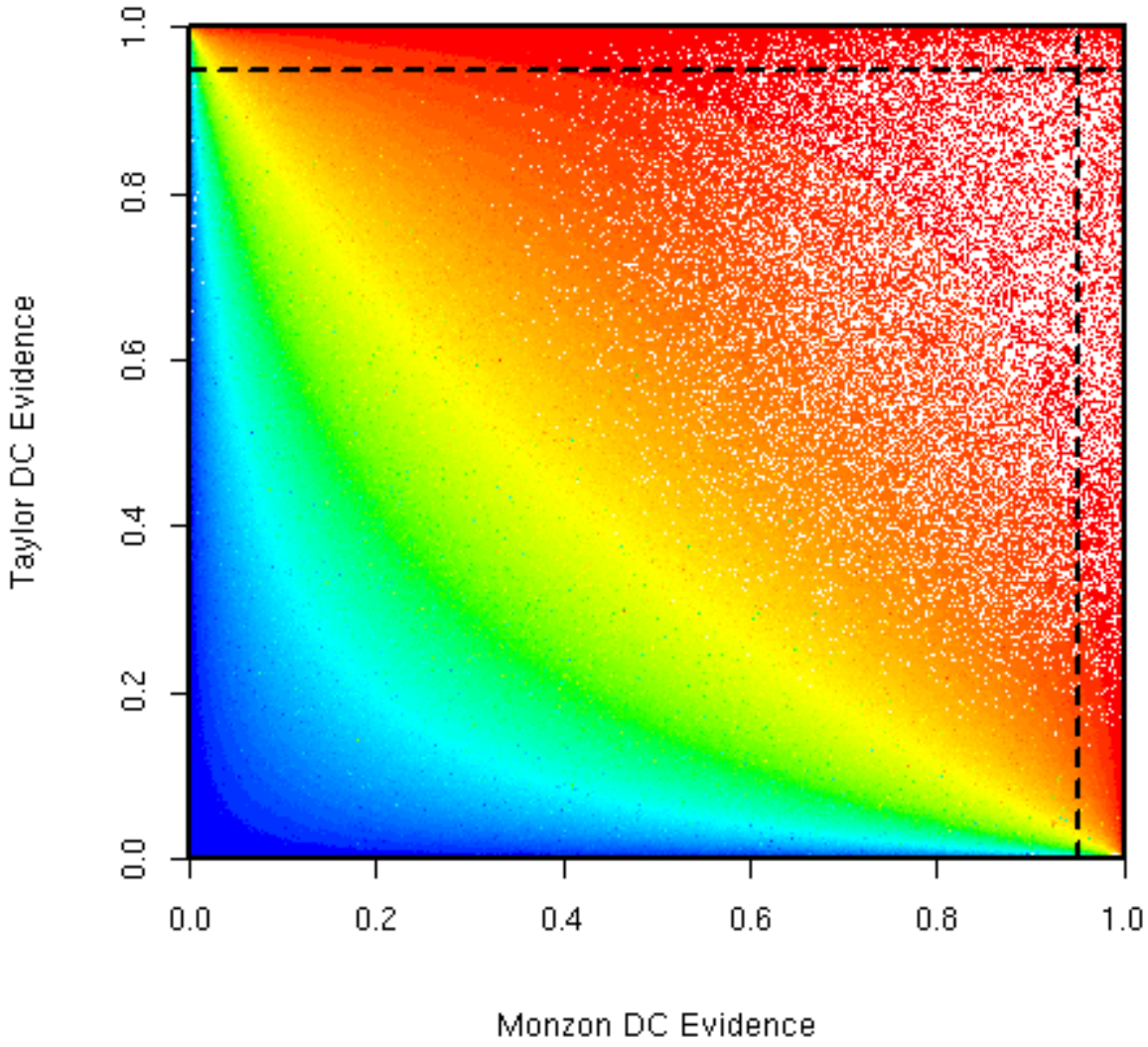
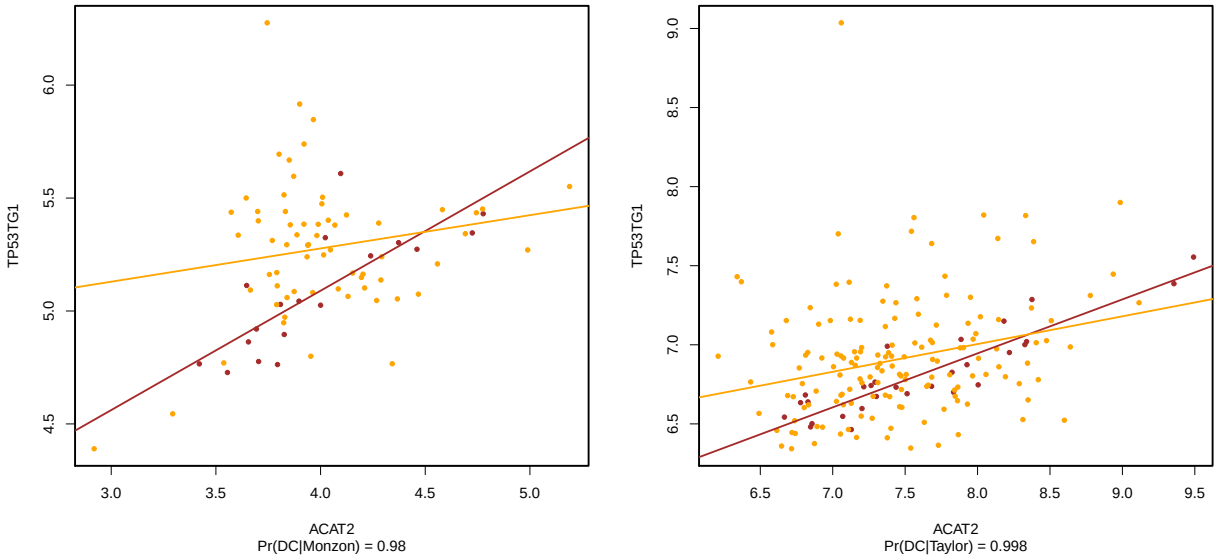
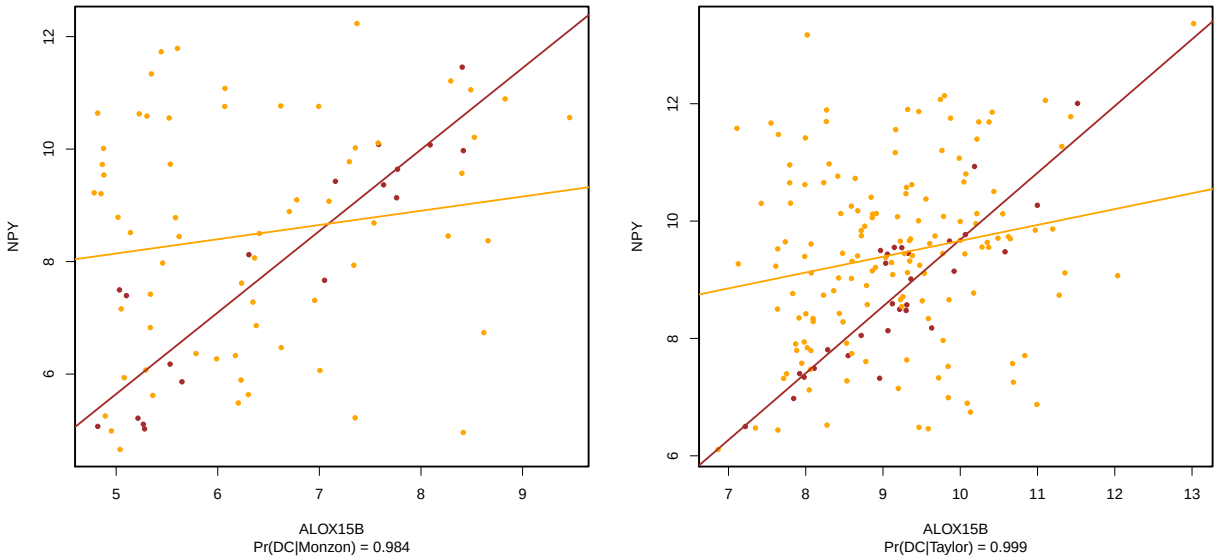


Figure 1: A pictorial overview of the relationship between the posterior probabilities of differential correlation obtained from the Monzon and Taylor individual analyses and that obtained from the meta-analysis. All 16.6 million gene pairs are plotted; their color corresponds to the meta-analysis evidence, ranging from blue (nil evidence of DC) to red (high evidence of DC). Dashed lines indicate the 0.95 cutoffs used in the individual analyses; those points boxed into the upper right-hand corner reflect the 408 pairs taken by both studies.

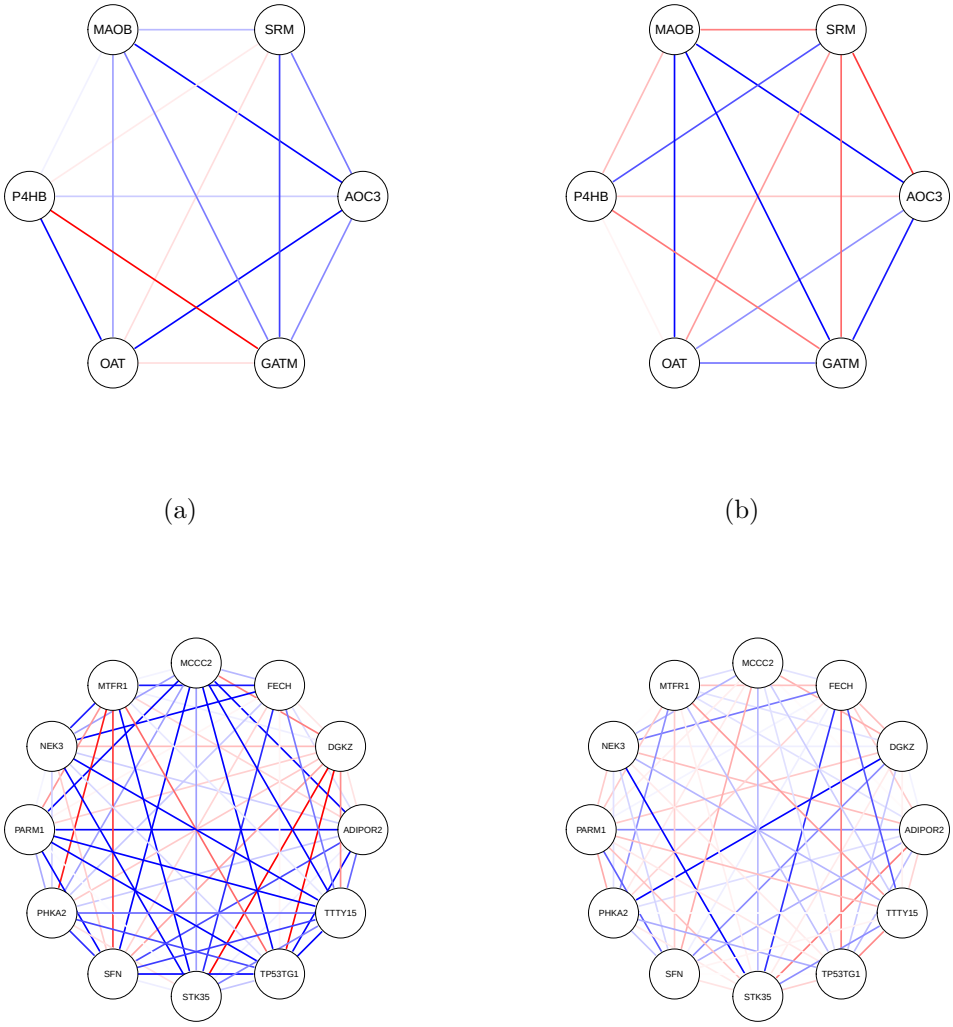


(a) Monzon (b) Taylor



(c) Monzon (d) Taylor

Figure 2: Two gene pairs deemed DC by Monzon, Taylor and the meta-analysis. The processed expression values for ACAT2∼TP53TG1 are plotted using data from (a) Monzon and (b) Taylor in the first two plots; ALOX15B∼NPY is similarly depicted in (c) and (d). Colors indicate condition, with non-cancerous subjects in purple and cancerous subjects in orange. A "robust" regression line is superimposed for each condition (see Methods).



(a) (b)

(c) (d)

Figure 3: In the upper panels, a graphical depiction of relationships within and across conditions for a small pathway related to prostate cancer identified by a recent paper (Gorlov et al., 2009) is shown. The two networks in (a) and (b) depict biweight midcorrelations observed among these genes within this pathway in non-cancerous and cancerous subjects, respectively, using the Monzon data. In the lower panels, relationships within and across conditions for a network of 12 genes greedily built up from a seed of PARM1 are shown, based on meta-analysis posterior probabilities of DC (but again illustrated using the Monzon data for (c) non-cancerous and (d) cancerous subjects). In both panels, deepness of color indicates strength of evidence of DC, where correlations of magnitude 0.5 or greater are given the deepest hue of red (−) or blue (+).

Web-based Supplementary Materials for "An Empirical Bayesian Approach for Identifying Differential Co-expression in High-throughput Experiments" by John A. Dawson and Christina Kendziorski

Web Appendix A: A single dataset in SIM II contains three groups of 100 genes, simulated in each of two conditions. Two covariance matrices Σ_1 and Σ_2 are created, one for each condition, such that the average strengths of the correlations in the first group are not the same between conditions (0.1 vs 0.6), but all others are unchanged (0.1 or 0):

$$\Sigma_k = \begin{pmatrix}
d_1 & \gamma_k & \gamma_k & \dots & \gamma_k & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\
\gamma_k & d_2 & \gamma_k & \dots & \gamma_k & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\
\gamma_k & \gamma_k & d_3 & \dots & \gamma_k & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\gamma_k & \gamma_k & \gamma_k & \dots & d_{100} & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & \dots & 0 & d_{101} & \gamma & \gamma & \dots & \gamma & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & \dots & 0 & \gamma & d_{102} & \gamma & \dots & \gamma & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & \dots & 0 & \gamma & \gamma & d_{103} & \dots & \gamma & 0 & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \dots & 0 & \gamma & \gamma & \gamma & \dots & d_{200} & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & d_{201} & \gamma & \gamma & \dots & \gamma \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & \gamma & d_{202} & \gamma_k & \dots & \gamma \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & \gamma & \gamma & d_{203} & \dots & \gamma \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & \gamma & \gamma & \gamma & \dots & d_{300}
\end{pmatrix}$$

where $\gamma_1 = \gamma = \frac{1}{9}$, $\gamma_2 = \frac{2}{3}$ and the d_i are $i.i.d.$ $N(\frac{10}{9}, 0.05^2)$.

Web Appendix B:

Monzon A study of prostate cancer gene expression profiles containing 18 normal and 65 diseased samples. The subjects were assayed using Affymetrix Human Genome U95 Version 2 Arrays, for which there are 11,724 probes corresponding to annotated genes. The data are available at the Gene Expression Omnibus (GEO) website (GSE 6919); see also Chandran *et al.* (2007) and Yu *et al.* (2004).

Roth A project aimed at creating a human body index of gene expression. Normal and diseased subjects were assayed for a multitude of tissues. For prostate, there are 7 normal and 17 diseased samples, but eight of the latter were excluded, as they were not prostate cancer subjects but rather BPH (enlarged prostate). The subjects were arrayed using Affymetrix Human Genome U133 Plus 2 Arrays, for which there are 40,686 probes corresponding to annotated genes. The data are available at GEO (GSE 7307); no citation is given.

Taylor A study profiling the genomics of prostate cancer. It involved 29 normal and 150 diseased samples, measured using Affymetrix Human Exon 1.0 ST Arrays, for which there are 22,466 annotated genes. The data are available at GEO (GSE 21034); see also Taylor *et al.* (2010).

Web Table 1: SIM III (Single Run Results)

Approach	FDR	Power	Time *
1-step (soft threshold; all pairs)	0.046	0.902	64066
1-step TCA-ECM (soft threshold; 0.1% pairs)	0.038	0.969	3453
1-step TCA-ECM (hard threshold; all pairs)	0.0010	0.714	64066
1-step TCA-ECM (hard threshold; 0.1% pairs)	0.0010	0.839	3453
ECF w/ $p = 10^{-1}$	0.415	0.871	26800+
ECF w/ $p = 10^{-2}$	0.074	0.555	31855+
ECF w/ $p = 10^{-3}$	0.014	0.269	31857+
ECF w/ $p = 5 \times 10^{-3}$	0.009	0.209	31857+
ECF w/ $p = 10^{-4}$ †	0.003	0.106	31853+
Box's M-test	0.096	0.740	4980

Average FDR and power from the proposed approach with hyperparameters estimated using the one-step version of the TCA-ECM under soft and hard thresholding. Results from the ECF approach of Lai *et al.* (2004) and Box's M-test (Mardia *et al.* 1979) are also shown. Values are given for the same, single simulation across all methods. Computational time is given in seconds.

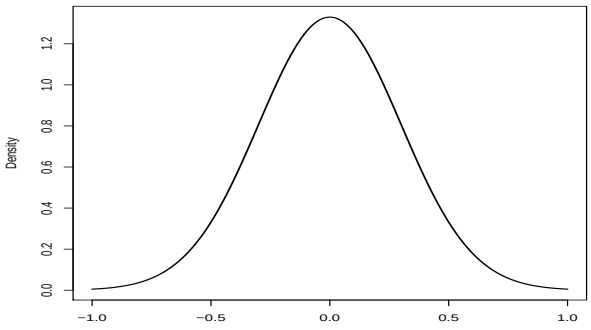
* Times do not include an additional 795 seconds used for ECF null simulation.

† This threshold is larger than 0.05/4000 and 0.1/4000, the values Lai *et al.* (2004) would use for 5% and 10% FDR control, respectively.

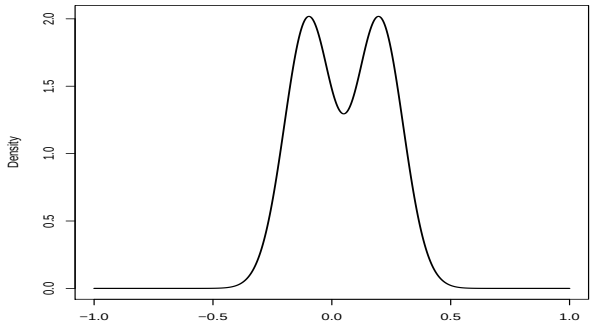
Web Table 2: SIM III

Approach	FDR	Power	Time
1-step TCA-ECM (soft threshold; 0.1% pairs)	0.058 (0.015)	0.985 (0.011)	3734 (362)
1-step TCA-ECM (hard threshold; 0.1% pairs)	0.0006 (0.0003)	0.913 (0.049)	3734 (362)

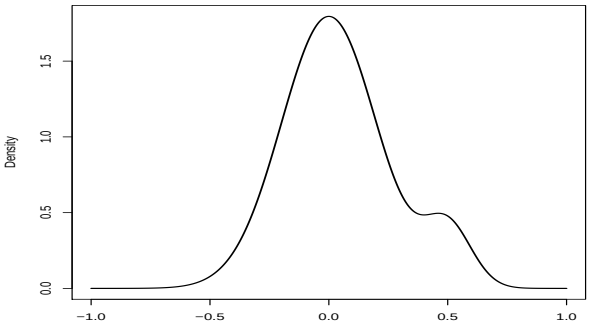
Average FDR and power from the proposed approach with hyperparameters estimated using the one-step version of the TCA-ECM under soft and hard thresholding and the subset heuristic with 0.1% of pairs. Values are means calculated over 20 simulated data sets; standard deviations are shown in parentheses. Computational time is given in seconds.



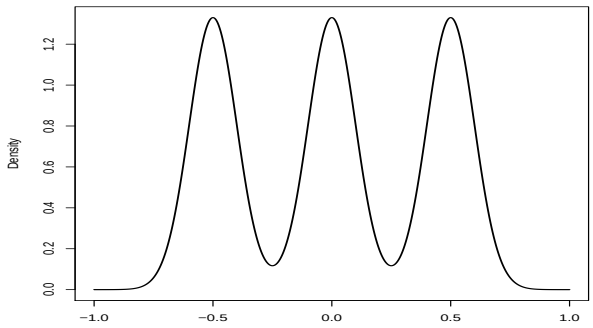
(a) A1



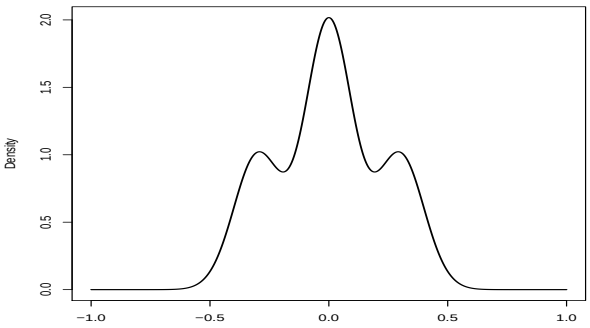
(b) B1



(c) B2

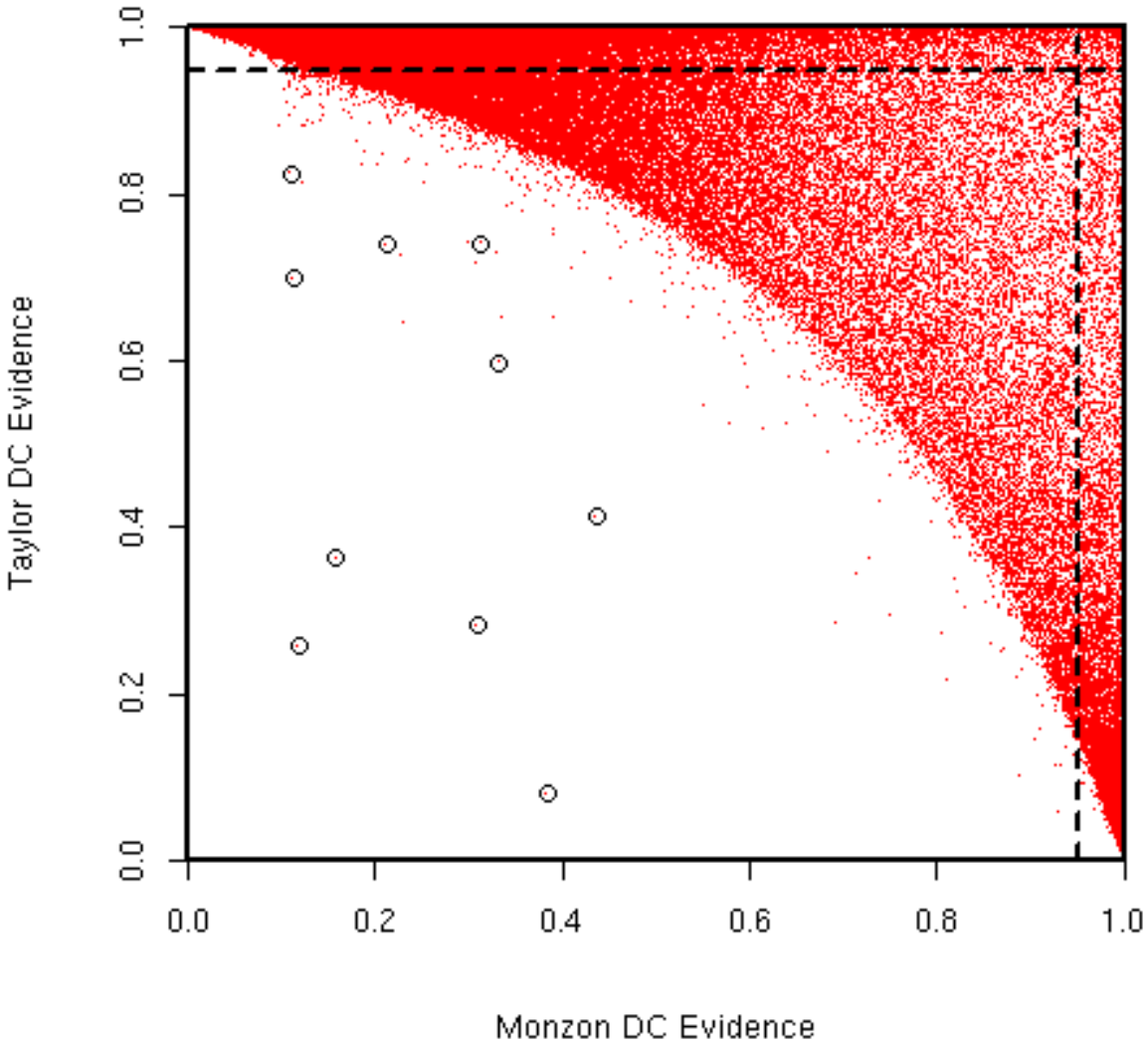


(d) C1

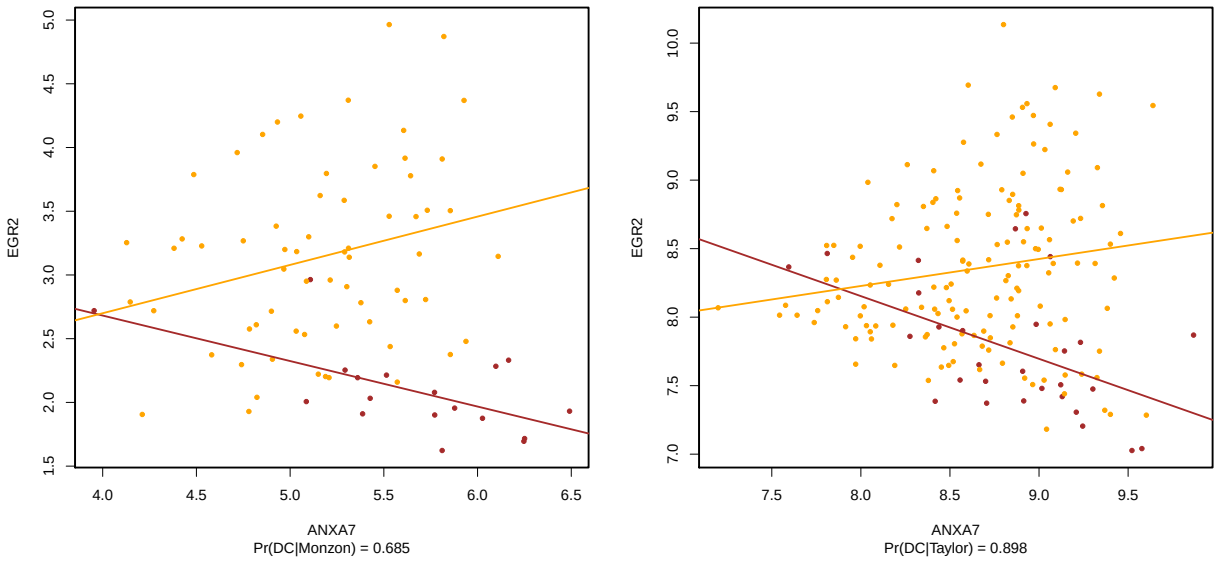


(e) C2

Web Figure 1: The densities for the distributions from which transformed correlations are drawn under the ideal framework of SIM I.

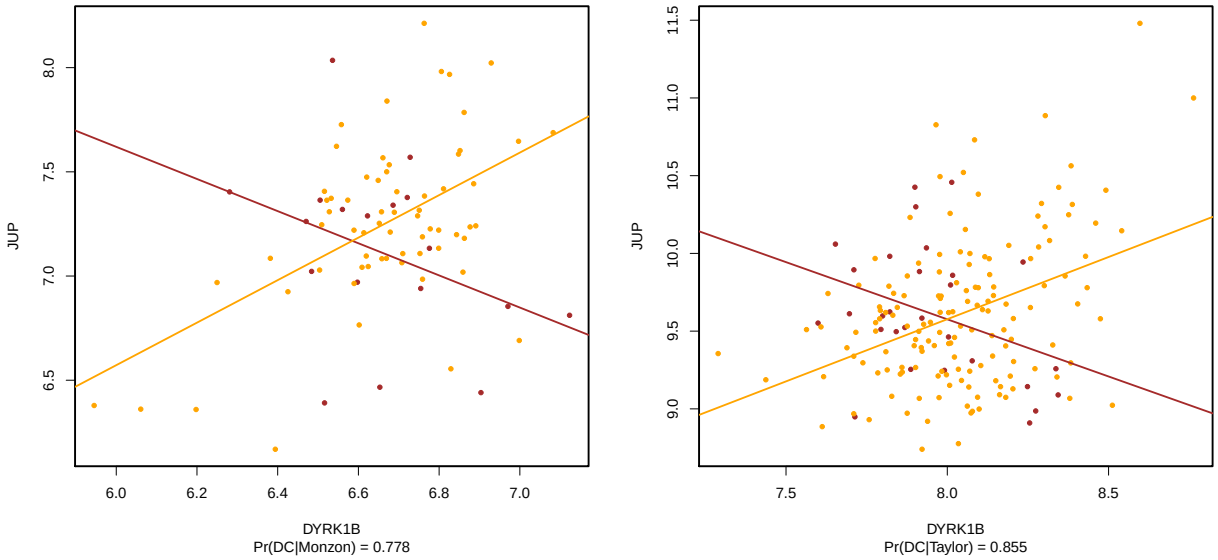


Web Figure 2: An illustration of how the Roth study contributes to the meta-analysis, despite its small sample size. The plot is similar to that in Figure 2 but only the 141,678 pairs taken by the meta-analysis are shown in red. The presence of pairs in regions with low posterior probabilities of DC for each individual study indicates that the Roth study has some, albeit limited, effect in the meta-analysis. To emphasize this point, pairs with posterior probability of DC greater than 0.5 are circled in black.



(a) Monzon

(b) Taylor



(c) Monzon

(d) Taylor

Web Figure 3: Two gene pairs deemed DC by the meta-analysis but not Monzon or Taylor individually. The processed expression values for ANXA7∼EGR2 are plotted using data from (a) Monzon and (b) Taylor in the first two plots; DYRK1B∼JUP is similarly depicted in (c) and (d). Colors indicate condition, with non-cancerous subjects in purple and cancerous subjects in orange. A "robust" regression line is superimposed for each condition (see Methods).