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-A contains three groups of 100 genes, simulated in each of two conditions. Two covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ 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). SIM II-B is similar except that the groups are of sizes 1000, 1000 and 2000.

$$\mathbf{\Sigma}_k = \left(\begin{array}{ccccccccccccccc}
d_1 & \gamma_k & \gamma_k & \dots & \gamma_k & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\
\gamma_k & d_2 & \gamma_k & \dots & \gamma_k & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\
\gamma_k & \gamma_k & d_3 & \dots & \gamma_k & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\gamma_k & \gamma_k & \gamma_k & \dots & d_{100} & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & \dots & 0 & d_{101} & \gamma & \gamma & \dots & \gamma & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & \dots & 0 & \gamma & d_{102} & \gamma & \dots & \gamma & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & \dots & 0 & \gamma & \gamma & d_{103} & \dots & \gamma & 0 & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \dots & 0 & \gamma & \gamma & \gamma & \dots & d_{200} & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & d_{201} & \gamma & \gamma & \dots & \gamma \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & \gamma & d_{202} & \gamma & \dots & \gamma \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & \gamma & \gamma & d_{203} & \dots & \gamma \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & \gamma & \gamma & \gamma & \dots & d_{300}
\end{array}\right)$$

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: A single dataset in SIM III contains two groups of 1500 genes, simulated in each of two conditions. Two covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are created, one for each condition, such that the average strengths of the correlations in each group are not the same between the first thirty genes and the rest across conditions (-0.4 vs 0.4), but all other correlations, both inter- and intra-group, are unchanged (0.4, 0.015 or 0).

$$\mathbf{\Sigma}_k = \begin{pmatrix}
d_1 & \gamma & \cdots & \gamma & \gamma_k & \gamma_k & \cdots & \gamma_k & 0 & 0 & \cdots & 0 & \delta & \delta & \cdots & \delta \\
\gamma & d_2 & \cdots & \gamma & \gamma_k & \gamma_k & \cdots & \gamma_k & 0 & 0 & \cdots & 0 & \delta & \delta & \cdots & \delta \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\gamma & \gamma & \cdots & d_{30} & \gamma_k & \gamma_k & \cdots & \gamma_k & 0 & 0 & \cdots & 0 & \delta & \delta & \cdots & \delta \\
\gamma_k & \gamma_k & \cdots & \gamma_k & d_{31} & \gamma & \cdots & \gamma & 0 & 0 & \cdots & 0 & \delta & \delta & \cdots & \delta \\
\gamma_k & \gamma_k & \cdots & \gamma_k & \gamma & d_{32} & \cdots & \gamma & 0 & 0 & \cdots & 0 & \delta & \delta & \cdots & \delta \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\gamma_k & \gamma_k & \cdots & \gamma_k & \gamma & \gamma & \cdots & d_{1500} & 0 & 0 & \cdots & 0 & \delta & \delta & \cdots & \delta \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & d_{1501} & \gamma & \cdots & \gamma & \gamma_k & \gamma_k & \cdots & \gamma_k \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \gamma & d_{1502} & \cdots & \gamma & \gamma_k & \gamma_k & \cdots & \gamma_k \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \gamma & \gamma & \cdots & d_{1530} & \gamma_k & \gamma_k & \cdots & \gamma_k \\
\delta & \delta & \cdots & \delta & \delta & \delta & \cdots & \delta & \gamma_k & \gamma_k & \cdots & \gamma_k & d_{1531} & \gamma & \cdots & \gamma \\
\delta & \delta & \cdots & \delta & \delta & \delta & \cdots & \delta & \gamma_k & \gamma_k & \cdots & \gamma_k & \gamma & d_{1532} & \cdots & \gamma \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\delta & \delta & \cdots & \delta & \delta & \delta & \cdots & \delta & \gamma_k & \gamma_k & \cdots & \gamma_k & \gamma & \gamma & \cdots & d_{3000}
\end{pmatrix}$$

where $\gamma_1 = \gamma = 0.4$, $\gamma_2 = -0.4$, $\delta = 0.015$ and the d_i are all 1.

Web Appendix C: Descriptions of the Prostate Cancer Data Sets

Monzon

A study of prostate cancer gene expression profiles containing 18 normal and 65 diseased samples. The subjects in this study were assayed using Affymetrix Human Genome U95A 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). We note that expression data also exists for this study as assayed on the B through E counterparts of the aforementioned microarray platform, but these were not included in our analysis as these expression scores correspond to expressed sequence tags and not full-length, fully annotated genes.

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 there.

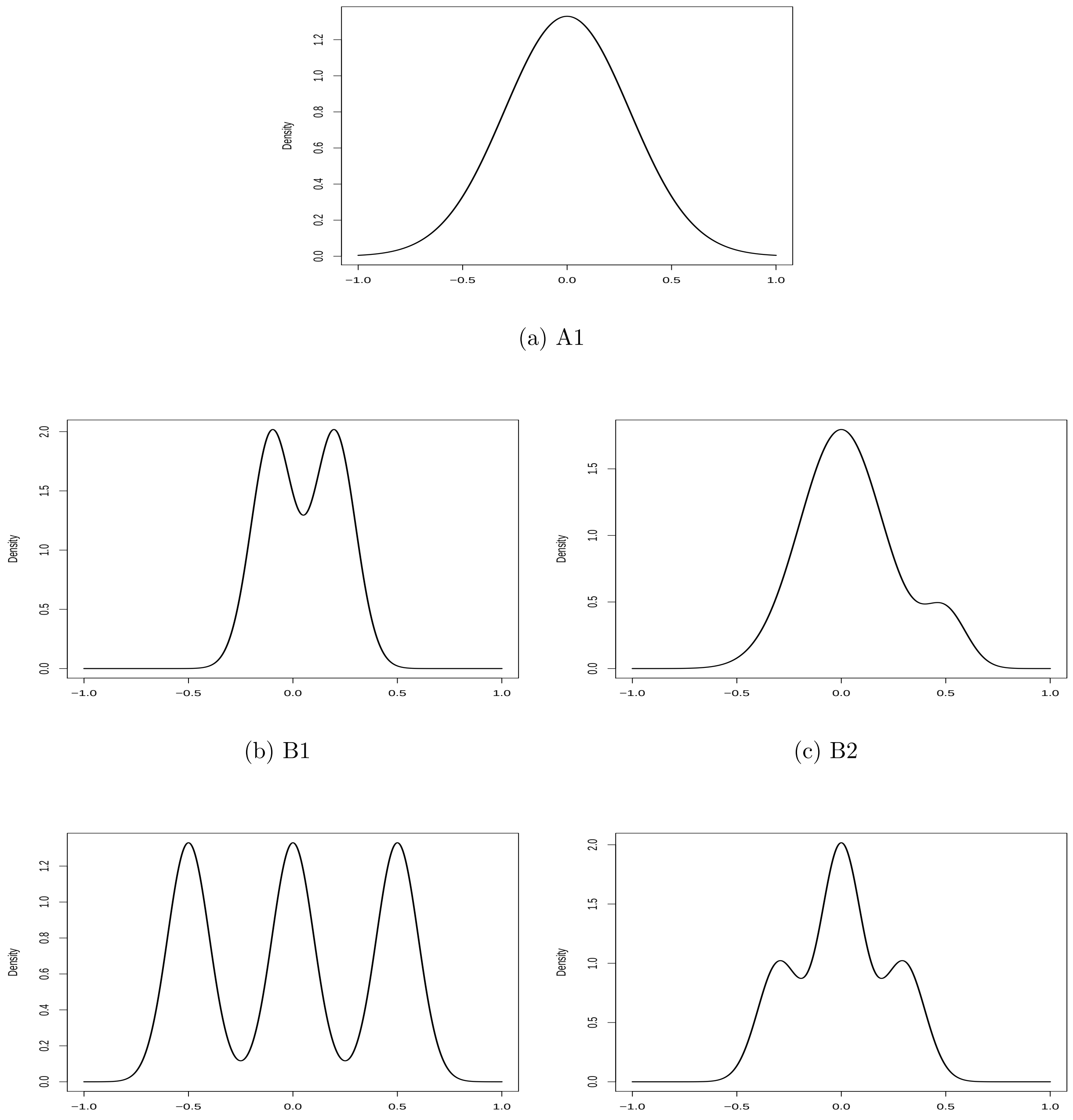
Taylor

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 II-A (True and False Positives)

Approach	Obs. FP	Obs. TP
1-step TCA-ECM (soft threshold)	280.1 (117.3)	4710.1 (490.5)
1-step TCA-ECM (hard threshold)	1.7 (1.2)	4302.6 (800.4)
ECF w/ $p = 10^{-1}$	1764.4 (196.3)	4613.6 (316.3)
ECF w/ $p = 10^{-2}$	129.8 (27.5)	3552.6 (696.6)
ECF w/ $p = 10^{-3}$	11.6 (4.5)	2238.2 (762.6)
ECF w/ $p = 5 \times 10^{-3}$	5.5 (3.0)	1886.4 (717.6)
ECF w/ $p = 10^{-4}$	0.9 (0.8)	1190.2 (571.8)
Box's M-test	395.9 (181.9)	4237.8 (332.5)

Average observed numbers of true and false positives 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.



(a) A1

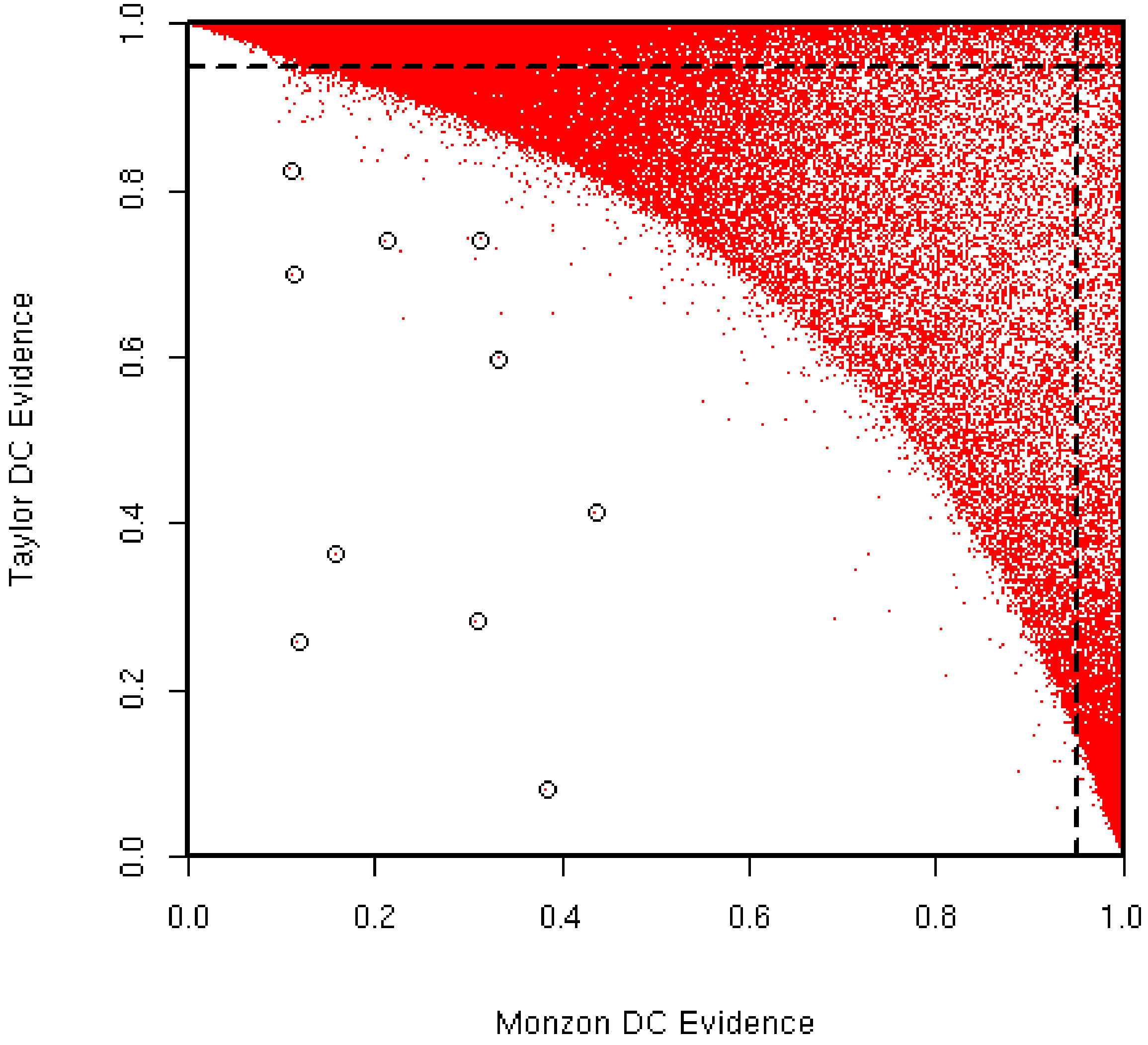
(b) B1

(c) B2

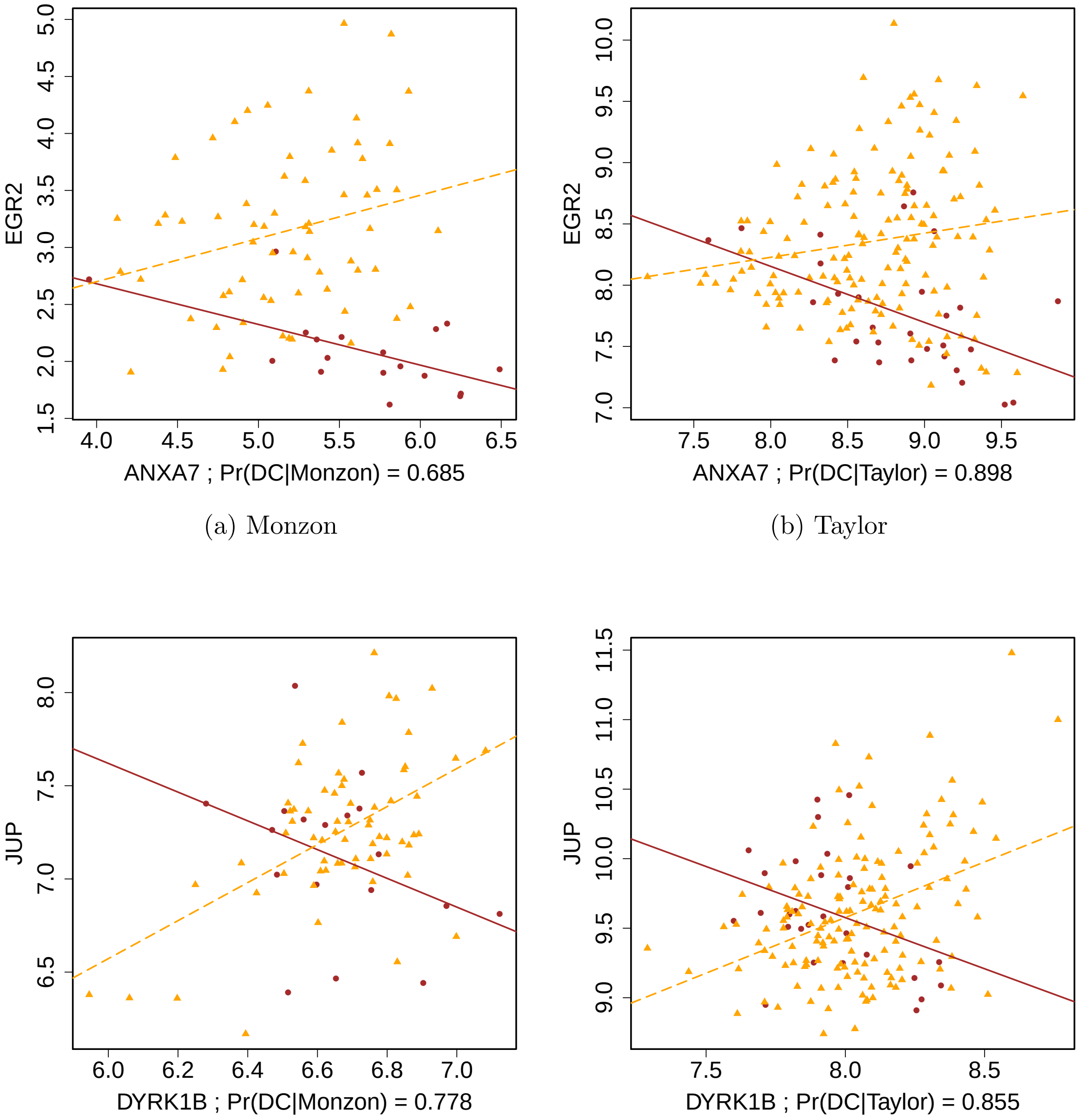
(d) C1

(e) C2

Web Figure 1: The densities for the distributions from which transformed correlations are drawn under the ideal framework of SIM I: (A1) A single Normal distribution: $N(0, 0.3^2)$; (B1) an even mixture of a $N(-0.1, 0.1^2)$ and a $N(0.2, 0.1^2)$; (B2) a 9-1 mixture of a $N(0, 0.2^2)$ and a $N(0.5, 0.1^2)$; (C1) an even mixture of a $N(-0.5, 0.1^2)$, a $N(0, 0.1^2)$ and a $N(0.5, 0.1^2)$; and (C2) a 1-2-1 mixture of a $N(-0.3, 0.1^2)$, a $N(0, 0.1^2)$ and a $N(0.3, 0.1^2)$.



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.



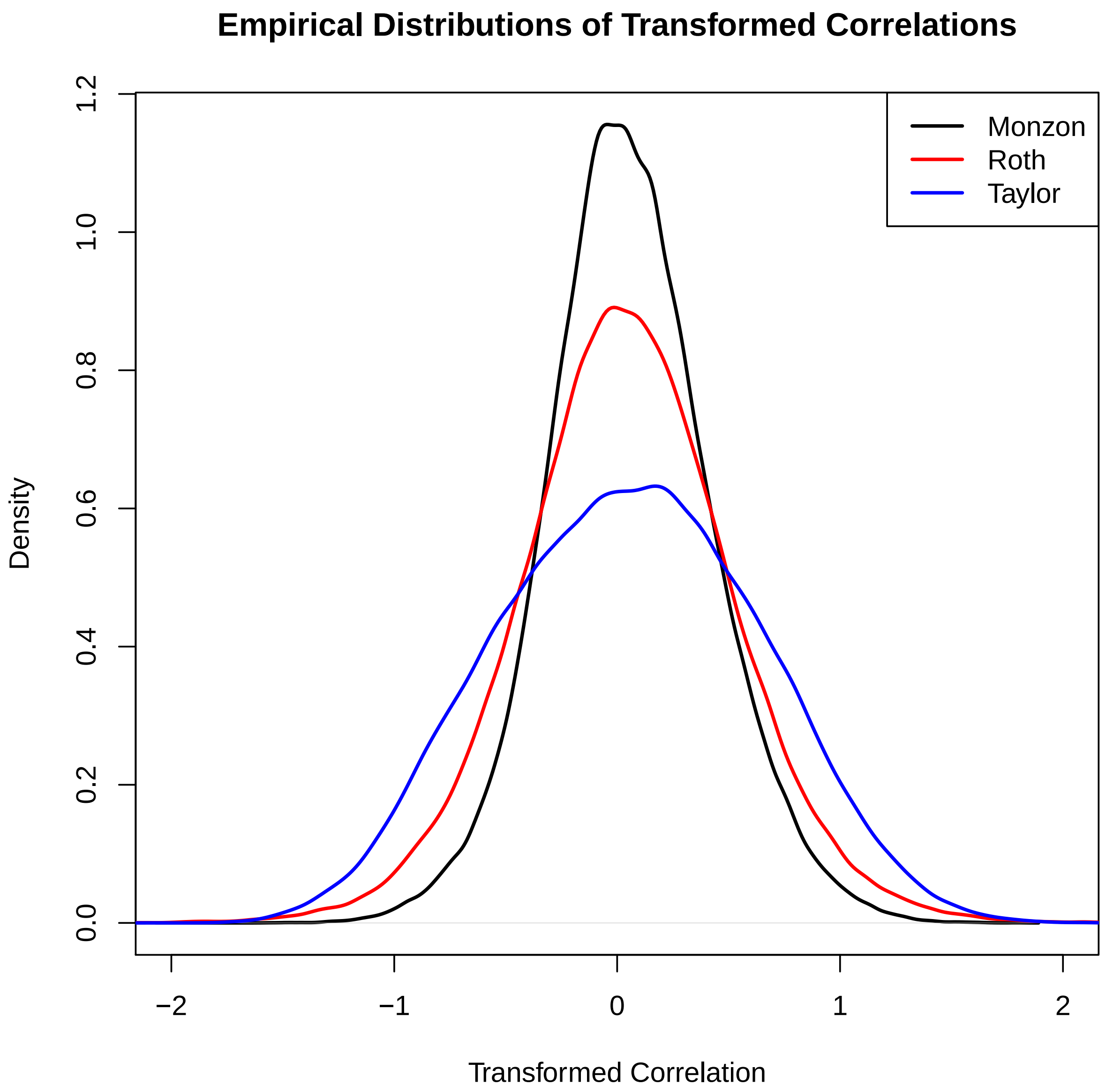
(a) Monzon

(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).



Web Figure 4: The empirical distributions of Fisher Z-transformed correlations over all gene pairs for all three prostate cancer data sets. Monzon is in black, Roth is in red and Taylor is in blue. Note that under Fisher's Z-transform, a correlation of 0.5 is approximately 0.55, a correlation of 0.7 is approximately 0.867 and a correlation of 0.9 is approximately 1.472.