# Web-based Supplementary Materials for "Analyzing multiple-probe microarray: estimation and application of gene expression indexes" by Mehdi Maadooliat, Jianhua Z. Huang, and Jianhua Hu

## 1 Web Appendix A: Simulation studies

We conducted simulation studies to evaluate the performance of the PSVD method in terms of parameter estimation and make comparisons to the entropy-based procedure in Hu et al. (2006).

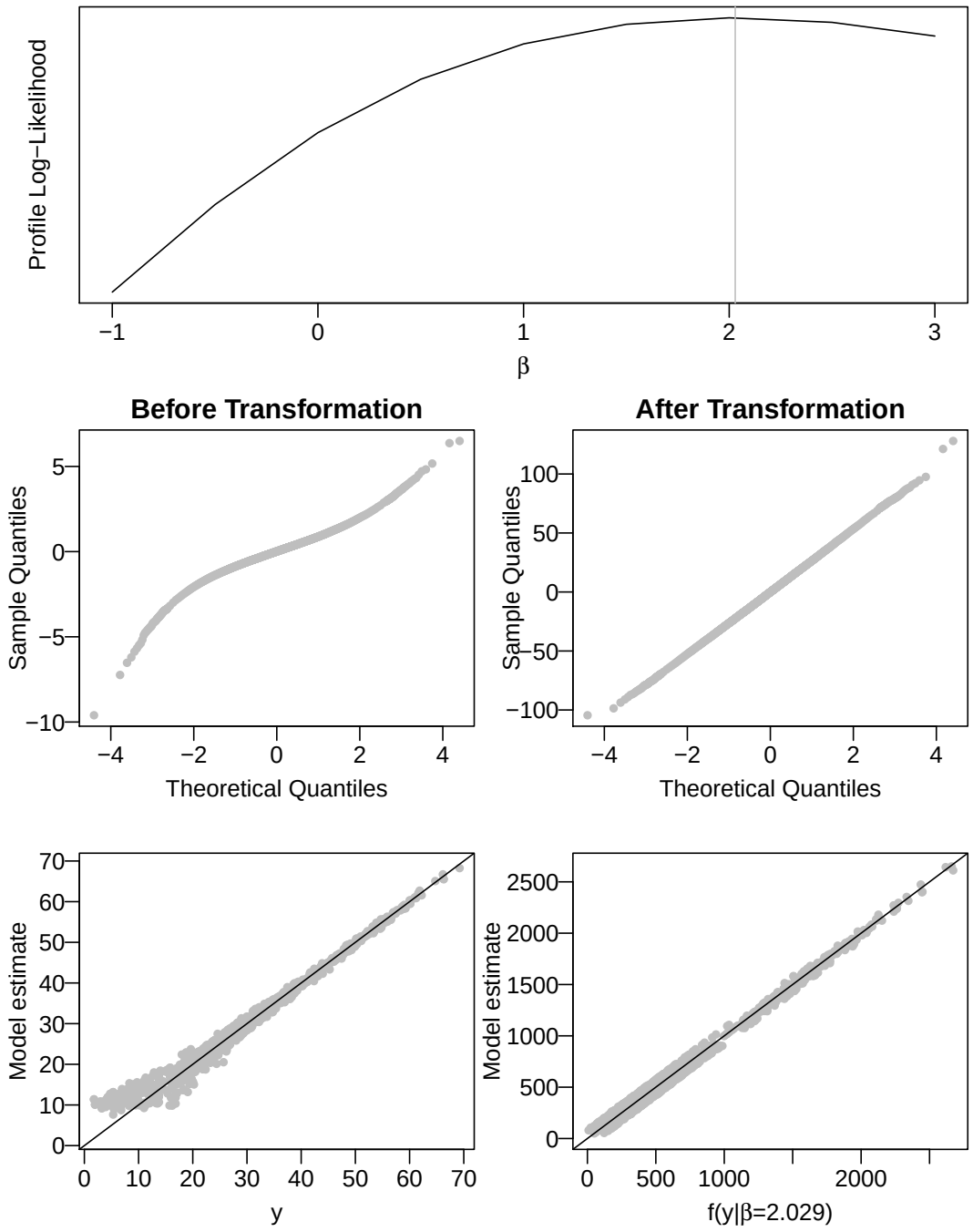
For the purpose of demonstration, we simulated the expression intensity data using model (2) for $H = 100$ genes with $I = 59$ (arrays) and $J = 16$ (probes). For each gene, we generated $\theta_i$'s from gamma distribution with shape parameter 3 and rate 1, multiplying by 150. We generated $\phi_j$'s from normal distribution with mean 1 and standard deviation 0.1 , and then scaled them to satisfy the constraint $\sum_j \phi_j^2 = J$. We considered $\beta = 2$ in the Box-Cox transformation throughout the simulation studies. We investigated four cases of the error distribution.

### 1.1 Normal errors

We generated the errors ($\epsilon_{ij}$'s) from mean zero normal distribution with the standard deviation of 25. The simulated data was generated from

$$y_{ij} = f^{-1}(\theta_i \phi_j + \epsilon_{ij} | \beta = 2), \quad i = 1 \dots I \quad j = 1 \dots J \tag{A1}$$

The PSVD estimate of $\beta$ using all the data of 100 genes is 2.029, while the entropy method estimate of $\beta$ is 2.023. The plot of profile log-likelihood versus the value of $\beta$ is contained in the top panel of Web Figure 1. A smooth concave function is clearly seen and the maximum location is indicated by the vertical line. It is not surprising for $\beta$ estimates to be similar between the PSVD and entropy based methods in the normal case.




**Web Figure 1:** The top panel contains the plot of the profile log-likelihood for $\beta$; The second row contains the QQ-plot of residuals for the Li-Wong and PSVD models; The left bottom panel contains the plot of $(\hat{\theta}_i\hat{\phi}_j)$'s obtained from LWR (without transformation) versus $y_{ij}$; and the right bottom panel contains the plot of mean expression estimates obtained from the transformation model (PSVD) versus $\widehat{f(y_{ij})}$.

The normal quantile-quantile plots of the estimated residuals before and after the transformation are shown in the left and right middle panels of Web Figure 1, respectively. It is obvious that the residual distribution is closer to normal with the transformation. We also displayed the plot of mean expression estimates $(\hat{\theta}_i\hat{\phi}_j)$'s obtained from the Li-Wong model (i.e., without transformation) versus $y_{ij}$ in the left bottom panel and that of mean expression estimates obtained from the transformation model versus $\widehat{f(y_{ij})}$'s in the right bottom panel. It is clear that the transformation results in more or less homoscedastic residuals, indicating good model fit.

To assess the variability of parameter estimation, we implemented the PSVD and entropy based methods for each gene separately and obtained 100 estimates of $\beta$ using each method. We reported the average value of $\hat{\beta}$ and the corresponding standard error in Web Table 1. We observe that the two methods yielded very similar mean values (close to the true value of 2) but the standard error using the PSVD method is only 45.5% of that using the entropy based method.

**Web Table 1:** Comparisons between the PSVD and entropy-based methods in four cases of residual distribution.

| Error Distribution | PSVD | Entropy |
|---|---|---|
| Normal | 2.026(0.081) | 2.017(0.176) |
| $t$ with $df = 3$ | 1.977(0.226) | 1.961(0.609) |
| Double Exp(1.5) | 2.034(0.127) | 2.001(0.258) |
| Skew Normal(1) | 2.017(0.084) | 2.024(0.183) |

## 1.2 Non-normal errors

We also investigated the robustness of the two methods through three non-normal cases. We considered model (2) with errors generated from the following zero-mean distribution: (a) The $t$ distribution with 3 degrees of freedom, multiplied by 15; (b) The double exponential distribution with the scale parameter 1.5; (c) The skew normal distribution with location parameter being $\frac{1}{\sqrt{\pi}}$, scale parameter 1 and shape parameter 1. We again obtained the simulated data using (A1). The last three rows of Web Table 1 contain the average values of $\hat{\beta}$ and the corresponding

standard errors. The parameter estimation variability of the entropy method is always at least twice as large as that of the PSVD method, which is consistent with the finding in the normal distribution case.

### 1.3 Sensitivity to model misspecification

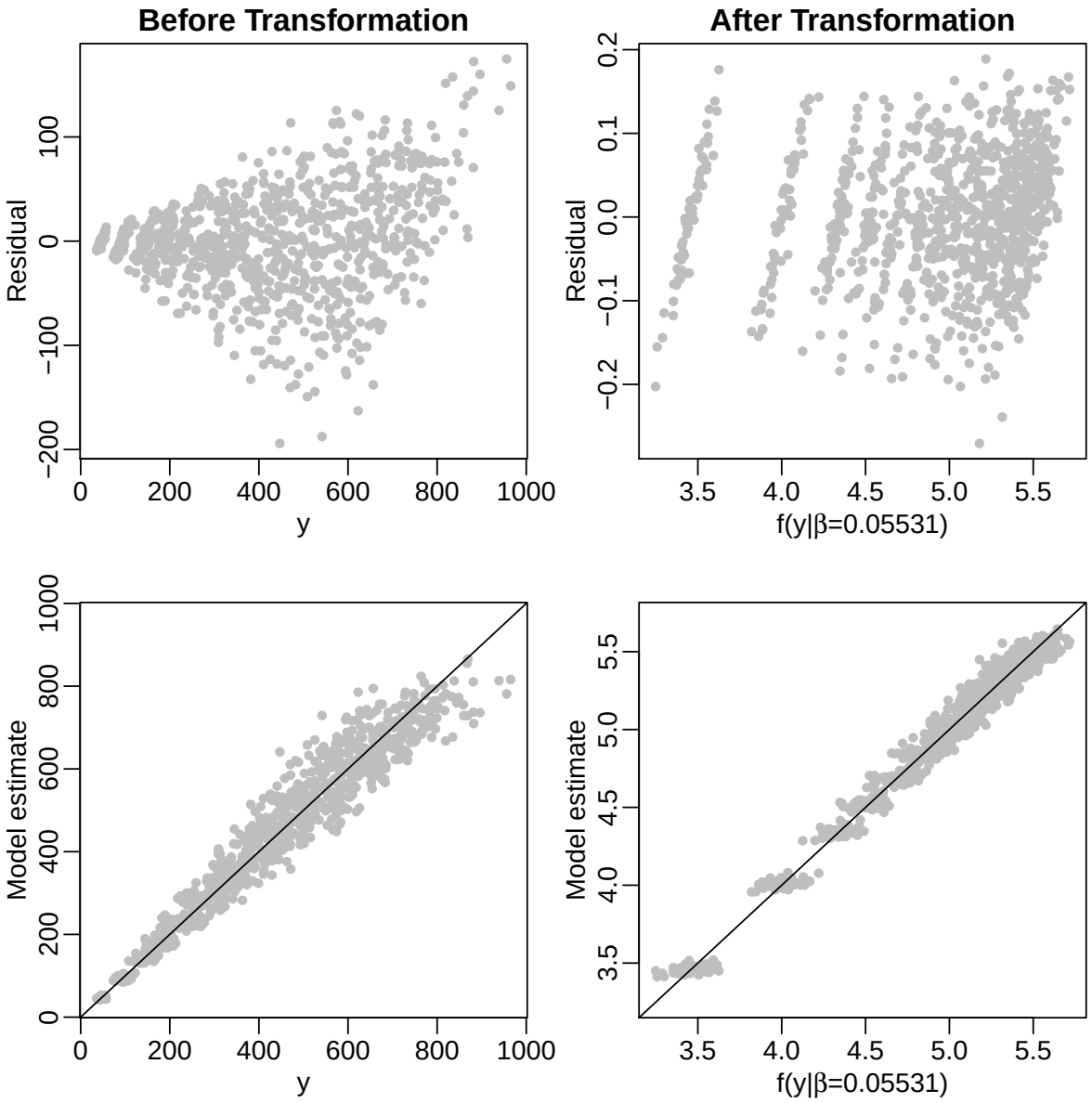
We also studied sensitivity of the proposed model to misspecification of the error distribution. We have generated the pseudo-observations $y_{ij}$'s with $\beta = 1$, specifically $y_{ij} = \theta_i\phi_j + \epsilon_{ij}$. We considered $\epsilon_{ij}$'s either from non-normal distributions or with unequal variances. We observed that fitting model (2) is essentially to find a transformation model ($f(y_{ij}|\hat{\beta}) = \hat{\theta}_i\hat{\phi}_j + \hat{\epsilon}_{ij}$) such that the distribution of the residuals ($\hat{\epsilon}_{ij}$) is as the closest as possible to homogeneous mean zero normal distributions posterior to the transformation. We investigated various heavily tailed or skewed distributions with constant variance for the error terms. A common observation is that the obtained $\hat{\beta}$ is very close to the true value of 1 which suggests no transformation under the proposed model. The result is consistent with that showed in Web Table 1, supporting the good performance of the proposed procedure. Since the Gaussian family only requires the first two moments to characterize the distribution, the obtained transformation results could be optimal in these cases of mean zero distributions with constant variance for the residuals.

To evaluate the performance of the model with unequal variance errors, we generated the errors ($\epsilon_{ij}$'s) from the normal distribution with mean zero and the standard deviation of the magnitude of $\phi_j$'s multiplied by 25. That is, we let the variances of the error terms vary cross the probe. PSVD obtained $\beta = 0.06$ which is similar to a log-transformation. Web Figure 2 shows the advantage of such a transformation model for a randomly selected gene. Both the residual plots in the upper panels and the plot of mean expression estimates obtained from the PSVD model versus $\widehat{f(y_{ij})}$ in the lower panels manifest that the proposed transformation corrects the pattern of increasing variability along the expression intensity of $y$.

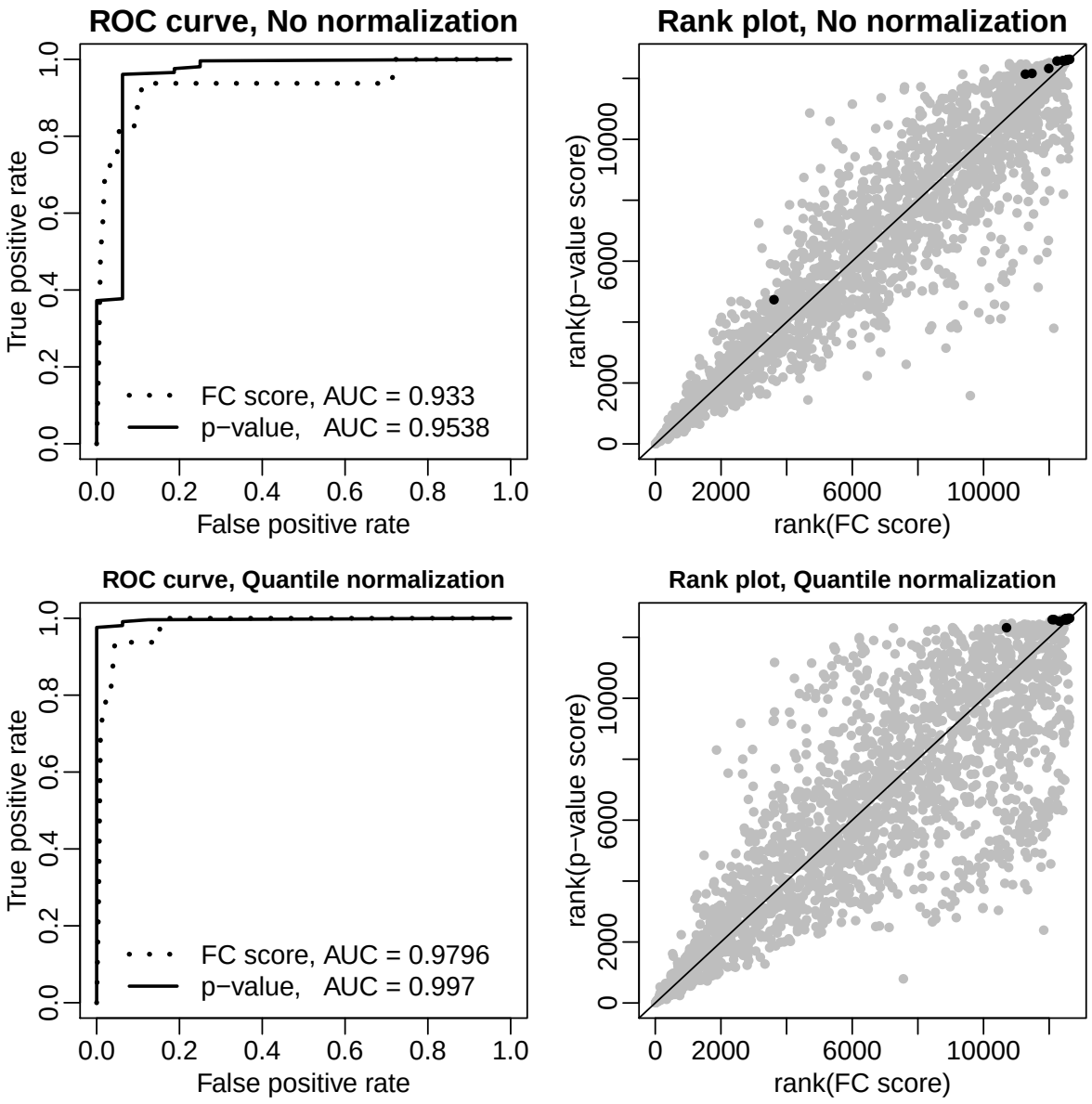
## 2 Web Appendix B: Additional results for discussion of the two practical issues

The Web Figures 3 and 4 contain the results based on the expression index estimates of the Li-Wong model. They are similar to Figures 4 and 5 of the results based on the PSVD method

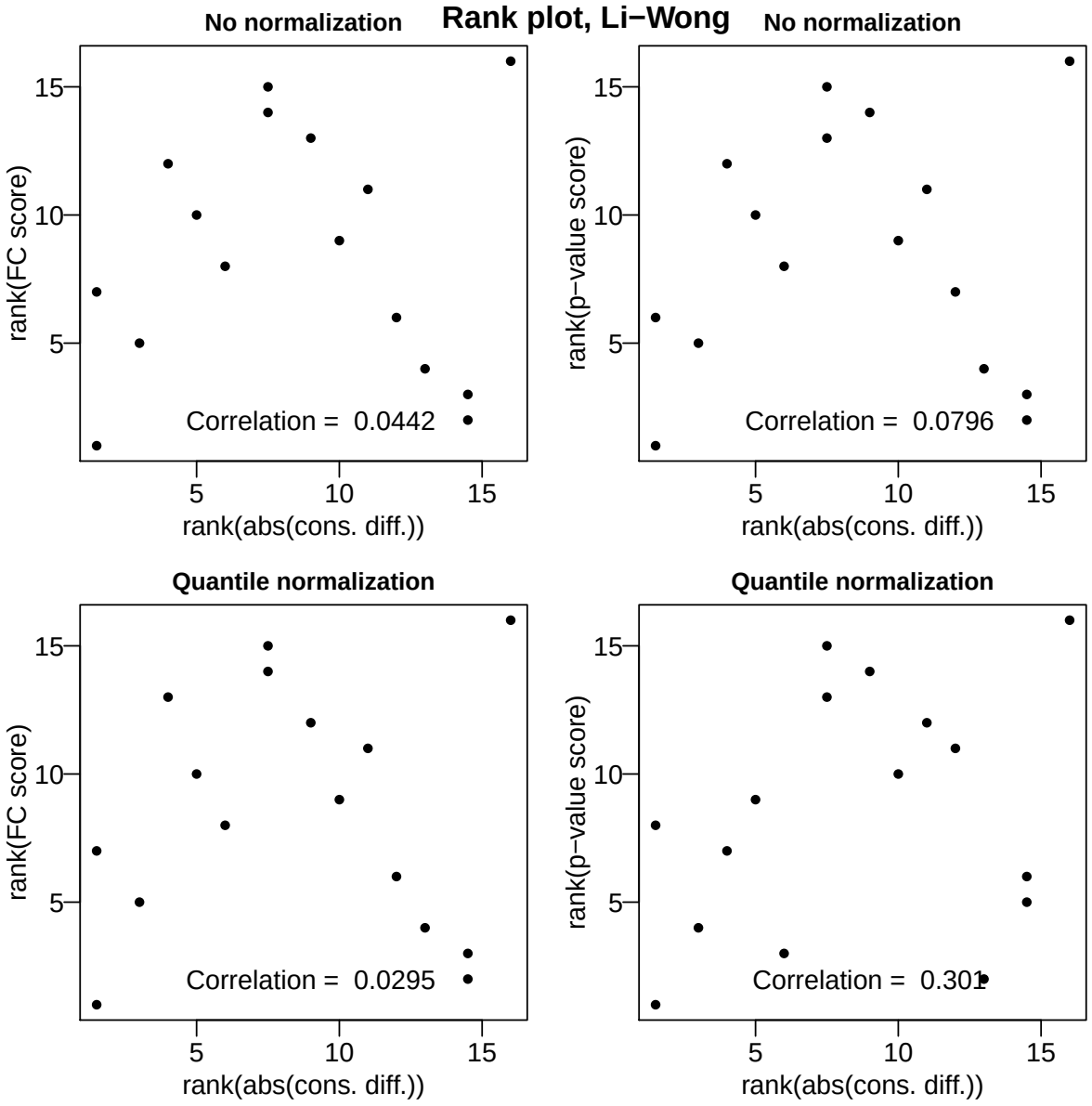
in the manuscript, respectively. In summary, we obtain the similar results as described in the manuscript.




**Web Figure 2:** The left and right columns contain the results without transformation and with transformation, respectively. The upper row contains the residual plots. The left bottom panel contains the plot of $(\hat{\theta}_i\hat{\phi}_j)$'s obtained from LWR (without transformation) versus $y_{ij}$; and the right bottom panel contains the plot of mean expression estimates obtained from PSVD versus $\widehat{f(y_{ij})}$.

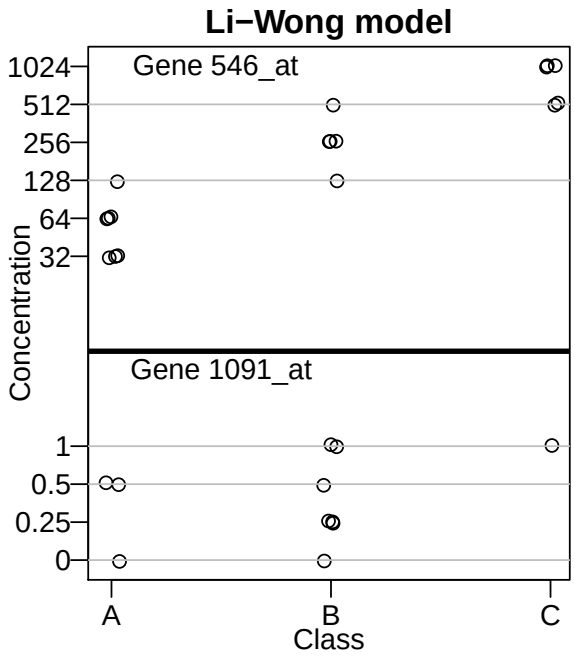



**Web Figure 3:** ROC curve and the rank plot for the Li-Wong Model. ROC curve for comparing the p-value and FC scores is given in the left column, and the rank plot for the rank(p-value score) vs. rank(FC score) is given in the right column. Upper row represents no normalization and the lower row represents the quantile normalization technique.




**Web Figure 4:** Rank plot for the 16 spiked-in genes based on the Li-Wong Model. Rank plot for the rank(FC score) vs. rank(abs(cons. diff.)) is given in the left column, and the rank plot for the rank(p-value score) vs. rank(abs(cons. diff.)) is given in the right column. Upper row represents no normalization and the lower row represents the quantile normalization technique.

Finally for genes *546_at* and *1091_at*, we repeated the K-means procedure to cluster the arrays into three classes using the expression indexes obtained from the Li-Wong model. The result is displayed in Web Figure 5.




**Web Figure 5:** K-means clustering using Li-Wong model. The upper region corresponds to gene *546_at* at the concentration levels of $32, 64, 128, 256, 512$, and 1024 picomolars; The lower region corresponds to gene *1091_at* at the concentration levels of $0, 0.25, 0.5$, and 1 picomolars. Each panel contains the plot of the concentration levels of the arrays versus their group memberships produced by K-means clustering.

## References

Hu, J., Wright, F. A., and Zou, F. (2006). Estimation of expression indexes for oligonucleotide arrays using the singular value decomposition. *Journal of the American Statistical Association* **101,** 41–50.