

# Supplement to "Degrees of differential gene expression: Detecting biologically significant expression differences and estimating their magnitudes"

David R. Bickel

*Medical College of Georgia  
Office of Biostatistics and Bioinformatics  
Augusta, GA 30912-4900*

## Applications to cancer data

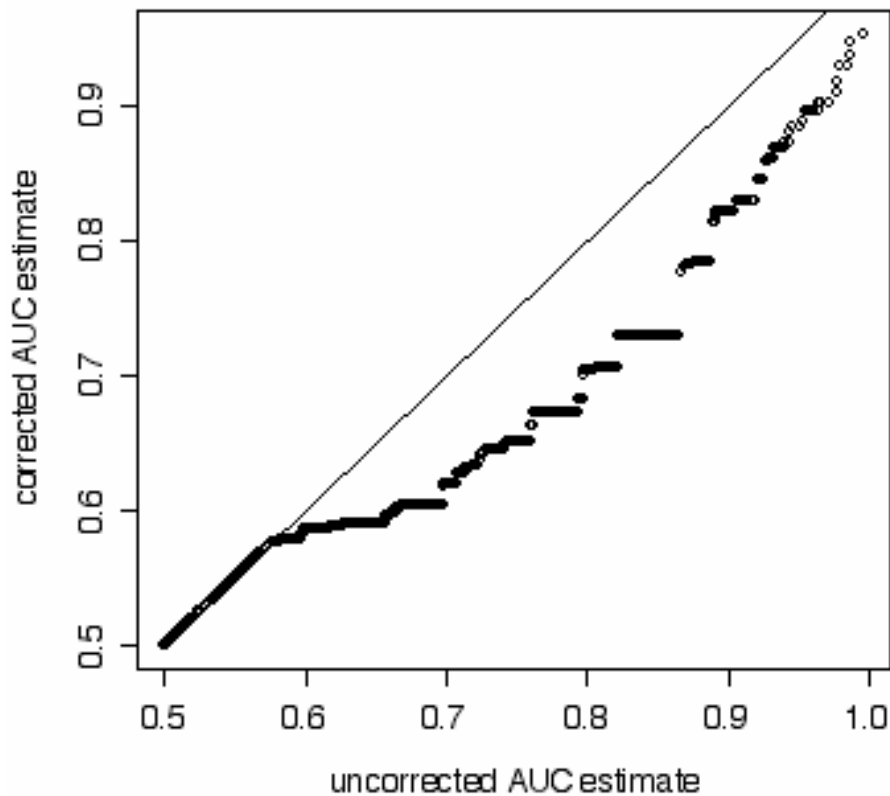
Microarrays are increasingly being used to identify genes as potential biomarkers for use in classifying tumors. When the level of gene expression is defined as the adjusted AUC (6) between microarrays of two different tumor types, higher expression differences indicate more confidence in a gene as a biomarker for use in a classification algorithm. Both the above method of estimating such differences and the above method of determining which genes have clinically relevant differential expression are illustrated with two publicly available leukemia data sets: that of Golub *et al.* (1999) and that of Yeoh *et al.* (2002). Before performing the analyses, each microarray was normalized by dividing all of its expression values by the median expression value of the array.

The new methods were applied to the oligonucleotide microarrays from the bone marrow cells of two patient groups of Golub *et al.* (1999): 38 patients with B-cell acute lymphoblastic leukemia (B-ALL) and 25 patients with acute myeloid leukemia (AML). A reasonable choice of a 5th or 50th percentile for genes discovered to be differentially expressed would be at least 0.60 since only a 60% probability of correctly distinguishing a B-ALL patient from an AML patient would not usually be clinically relevant. Table 2 shows the numbers of genes that would be discovered to be differentially expressed when the 5th quantile or 50th quantile (median) is chosen to be 0.50, 0.60, 0.70, or 0.80. ( $\alpha_0 = \alpha_{\min} = 0.50$  is included only because it corresponds in the 5th percentile case to standard false discovery rate control at the 5% level, but it is too small to be clinically meaningful.) The expression difference for each gene  $i$  was defined as the best possible AUC between the two groups (6) and was estimated by both  $t_i$  and  $\tilde{z}_i$ ; the two estimators are compared in Fig. 4. It can be seen that the estimates differ mostly for intermediate values of the AUC.

5 th percentile, $\xi_0$	Number of discoveries	Rejection threshold, $\tau$	Median, $\xi_0$	Number of discoveries	Rejection threshold, $\tau$
0.50	1285	0.69	0.50	4028	0.58
0.60	209	0.83	0.60	676	0.75
0.70	0	$\infty$	0.70	50	0.92
0.80	0	$\infty$	0.80	0	$\infty$

**Quantile analyses for the B-ALL/ALL set of gene-gene comparisons.** The number of discoveries is the number of genes for which  $T_i \geq \tau$ , where  $\tau$  was estimated with the goal that the given quantile (5th percentile or median) of  $\{\xi_i : T_i \geq \tau\}$  equals 0.50, 0.60, 0.70, or 0.80. If it is impossible to achieve the desired quantile, then  $\tau = \infty$ , i.e., no discoveries are made, as in Benjamini and Hochberg (1995).

**Table 2**



**Comparing two estimators of expression difference:  $\tilde{z}_i$  versus  $T_i$  between the B-ALL and AML microarrays.** The corrected estimate  $\tilde{z}_i$  versus the uncorrected estimate  $T_i$ . The line of equality is included for comparison.

**Figure 4**

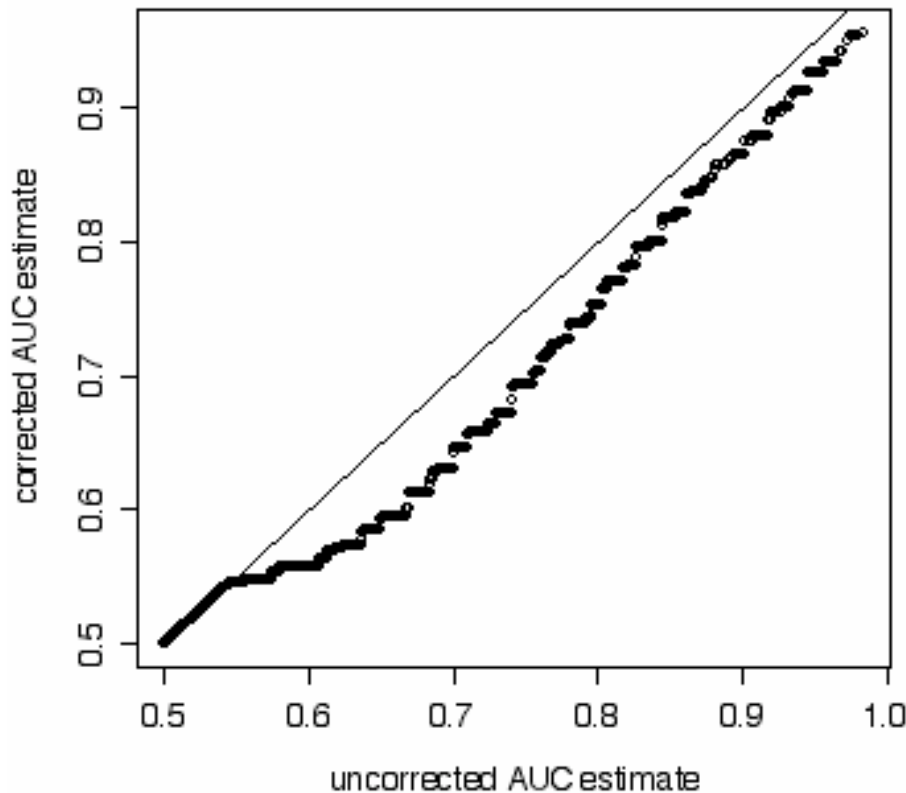
The same methods were applied to the largest two subgroups of the Yeoh *et al.* (2002) data set of oligonucleotide microarrays from children with various subtypes of leukemia: the hyperdiploid > 50 chromosomes subgroup (H50, 64 children) and the TEL-AML1 subgroup (TEL-AML1, 79 children). A comparison between Tables 2 and 3 shows that more genes are differentially expressed, with lower rejection thresholds, for the H50/TEL-AML1 set of comparisons than for the B-ALL/AML set of comparisons, at each value of  $\xi_0$  below 0.80. Fig. 5 indicates less bias in the uncorrected estimate for H50/TEL-AML1 than Fig. 4 does for B-ALL/AML, but even in the H50/TEL-AML1 case, the difference between the uncorrected and corrected estimates is only negligible for  $AUC \leq 0.54$ , which is too close to random selection ( $AUC=0.50$ ) to be of much interest.

5 th percentile, $\xi_0$	Number of discoveries	Rejection threshold, $\tau$	Median, $\xi_0$	Number of discoveries	Rejection threshold, $\tau$
0.50	2413	0.63	0.50	6080	0.56
0.60	745	0.74	0.60	1210	0.70
0.70	180	0.86	0.70	335	0.82
0.80	0	$\infty$	0.80	0	$\infty$

**Quantile analyses for the H50/TEL-AML1 set of gene-gene comparisons.**

This table is constructed as Table 2.

**Table 3**

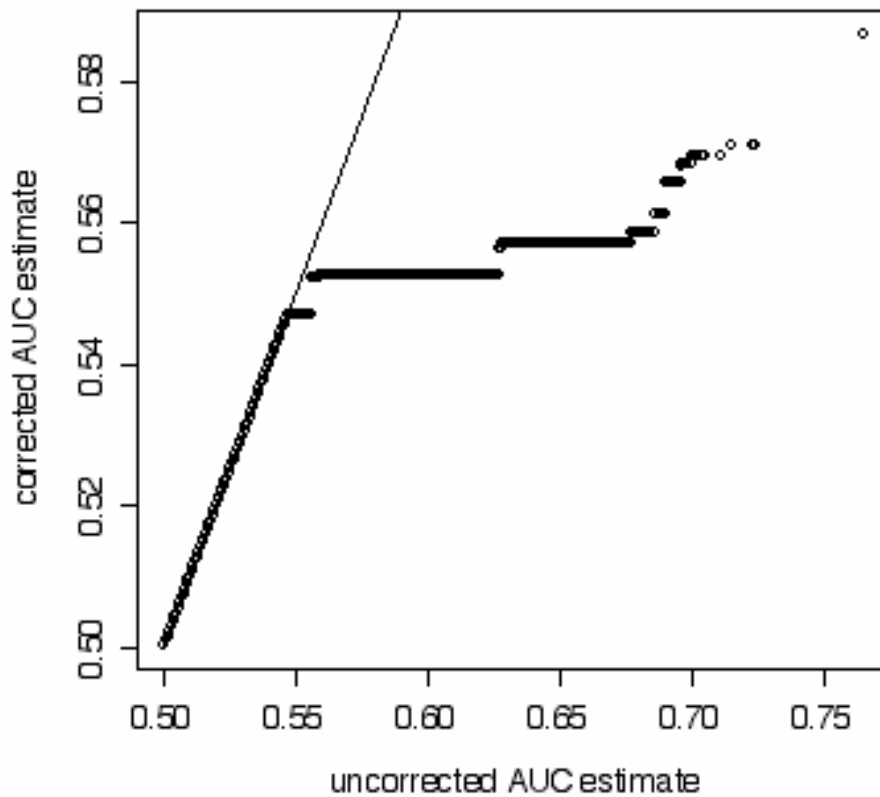


**Comparing two estimators of expression difference:  $\tilde{z}_i$  versus  $T_i$  between the H50 and TEL-AML1 microarrays.** The corrected estimate  $\tilde{z}_i$  versus the uncorrected estimate  $T_i$ , as in Fig. 4. The line of equality is included for comparison.

**Figure 5**

In order to assess the performance of the proposed methods in the absence of differential expression, they were also used to compare a random subgroup of the TEL-AML1 microarrays to another random subgroup. The first group consisted of 39 microarrays selected randomly without replacement from the 79 TEL-AML1 microarrays, and the second group consisted of another 39 microarrays selected randomly from the remaining 40 microarrays. The two groups were compared, as if to detect differential expression and to estimate the extent of any differential expression. The results reflect the fact that the expression values for the two groups were drawn from the same distribution. No genes were found to have a fifth quantile of discovered differential expression greater than expected by chance ( $z_0 = z_{\min} = 0.50$ ), but one gene was found to have a median of discovered differential expression greater than expected by chance. However, even that gene was not found to be differentially expressed using the more clinically reasonable null hypotheses ( $z_0 \geq 0.60 > z_{\min}$ ). The population AUC is exactly 0.5 for each gene, but estimates of the AUC vary for any two groups of microarrays randomly

selected. As expected, the uncorrected estimates are farther from the population value than the corrected estimates (Fig. 6). It is noteworthy that the correction adjusts the estimates more in this case than in the cases in which there is real differential expression (Figs. 4-5). However, the example is somewhat artificial since an investigator would not normally be interested in the degree of differential expression of genes not found to be differentially expressed by the hypothesis testing procedure.



**Comparing two estimators of expression difference:  $\tilde{z}_i$  versus  $T_i$  between two random subgroups of TEL-AML1 microarrays.** The corrected estimate  $\tilde{z}_i$  versus the uncorrected estimate  $T_i$ , as in Fig. 4. The line of equality is included for comparison.

**Figure 6**

## Appendix A: Algorithm for estimating expression differences

The motivation of this algorithm is described in the section of the manuscript entitled, "Method 2: Correcting the bias in estimates of expression differences."

**Step 1.** For  $i = 1, 2, \dots, m$ , let  $T_i$  be the observed test statistic (estimated expression difference) of the  $i$ th of  $m$  genes.

Sort the test statistics of Step 1 from smallest to largest and assign a rank to each the test statistic, from 1 for the smallest test statistic to  $m$  for the largest test statistic, with ties broken arbitrarily or randomly. For  $i = 1, 2, \dots, m$ , call  $\rho_i$  the rank of  $T_i$ , so that, for example,  $\rho_i = 5$  if  $T_i$  is the fifth smallest test statistic. (If there are a large number of ties, random tie-breaking should be used and this step should be placed immediately after Step 3, letting  $\rho_{j,i}$  be the rank of the  $i$ th gene for the  $j$ th iteration, and appropriately adjusting Steps 8, 10, 13, 14, and 15.)

**Step 2.** For  $i = 1, 2, \dots, m$ , call  $\rho_i$  the rank of  $T_i$ , so that, for example,  $\rho_i = 5$  if  $T_i$  is the fifth smallest test statistic. (If there are a large number of ties, random tie-breaking should be used and this step should be placed immediately after Step 3, letting  $\rho_{j,i}$  be the rank of the  $i$ th gene for the  $j$ th iteration, and appropriately adjusting Steps 8, 10, 13, 14, and 15.)

**Step 3.** Let  $j = 1$  and let  $J$  be a large integer representing the number of resampling iterations.  $J$  should be large enough to average out fluctuations from one random sample to another, but small enough to compute all the iterations in a reasonable amount of time.

**Step 4.** For each of the two groups of microarrays being compared, randomly choose half of the microarrays of each group, without replacement, to determine the ranks, leaving the other half of the microarrays of each group to estimate the gene expression differences at given ranks. Thus, each of the two comparison groups consists of *microarrays for ranks* and an equal number of *microarrays for expression*.

**Step 5.** Considering only the microarrays for ranks, for  $i = 1, 2, \dots, m$ , let  $t_{j,i}^{**}$  be the test statistic (estimated expression difference) for the  $i$ th gene.

**Step 6.** Considering only the microarrays for expression, for  $i = 1, 2, \dots, m$ , let  $t_{j,i}^*$  be the test statistic (estimated expression difference) for the  $i$ th gene.

**Step 7.** For  $r = 1, 2, \dots, m$ , let  $t_{j,(r)}^{***}$  be the test statistic of those of Step 6 that corresponds in the gene to the  $r$ th lowest value of the test statistics of Step 5. ( $t_{j,(r)}^{***}$  is the  $j$ th estimate of the expression difference for the gene with the  $r$ th smallest test statistic among the test statistics of Step 1.) That is, define  $k(j, r)$  such that  $t_{j,k(j,r)}^{**}$  is less than or equal to exactly  $r$  values of the  $m$  test statistics of Step 5 (with ties broken arbitrarily or randomly), and let  $t_{j,(r)}^{***} = t_{j,k(j,r)}^*$ . Thus, for example,  $t_{j,(3)}^{***}$  is set equal to  $t_{j,k(j,3)}^*$ , where  $k(j, 3)$  is the index of the gene of the third lowest test statistic of the microarrays for ranks in the  $j$ th iteration:  $|\{t_{j,k}^{**} \mid t_{j,k}^{**} \leq t_{j,k(j,3)}^{**}\}| = 3$ , which could also be expressed as  $t_{j,k(j,3)}^{**} = t_{j,(3)}^{**}$ . (The set  $\{t_{j,k}^{**}\}_{k=1}^m$  is identical to the set  $\{t_{j,(r)}^{**}\}_{r=1}^m$ , but the values of  $t_{j,(r)}^{**}$  are ordered such that  $t_{j,(1)}^{**} \leq t_{j,(2)}^{**} \leq t_{j,(3)}^{**} \leq \dots \leq t_{j,(m)}^{**}$ .) For the  $j$ th iteration,  $t_{j,(r)}^{***}$  is the corrected estimate of the expression difference of the gene with the  $r$ th lowest estimated expression difference.

- For  $i = 1, 2, \dots, m$ , let  $t_{j,i}^{***} = t_{j,(\rho_i)}^{***}$  or, if Step 2 follows Step 3, let  $t_{j,i}^{***} = t_{j,(\rho_{ji})}^{***}$ . This
- Step 8.** ensures that the ranks of the test statistics of Step 1 are the same as those of Step 5, while using the estimates of Step 6.
- Step 9.** Increment  $j$  by one ( $j \leftarrow j + 1$ ).
- Step 10.** If  $j \leq J$ , repeat all steps from Step 4 (using different random numbers) through Step 10. When  $j > J$ , proceed to the next step.
- Step 11.** For  $i = 1, 2, \dots, m$ , let  $t_i^{***} = \frac{1}{J} \sum_{j=1}^J t_{j,i}^{***}$ . This gives the mean estimate of the expression difference over all  $J$  iterations for each of the  $m$  genes.
- Step 12.** For  $i = 1, 2, \dots, m$ , set  $\hat{z}_i$  equal to  $t_i^{***}$  or  $T_i$ , whichever is lower.  $\hat{z}_i$  is the estimated expression difference of the  $i$ th gene, after correction for the fact that it will be regarded as the  $\rho_i$ th lowest test statistic.
- Step 13.** For  $r = 1, 2, \dots, m$ , set  $l(r)$  equal to the value for which  $\rho_{l(r)} = r$ . For example,  $l(3)$  will be the gene index at which  $T_{l(3)}$  is the third lowest test statistic among those of Step 1.
- Step 14.** Let  $\tilde{z}_{l(m)} = \hat{z}_{l(m)}$ . This will be the highest possible estimate of the expression difference since it corresponds to the gene with the highest uncorrected test statistic of Step 1.
- Step 15.** For  $r = m - 1, m - 2, \dots, 2, 1$ , set  $\tilde{z}_{l(r)}$  equal to  $\hat{z}_{l(r)}$  or  $\tilde{z}_{l(r+1)}$ , whichever is lower. This ensures that no estimate of the expression difference is greater than an estimate considered higher in rank.

The software mentioned in the abstract implements all steps of this algorithm with R functions. R is similar to S-PLUS (Insightful Corporation), but it has more convenient scoping rules, runs on more operating systems, and is freely available from [www.r-project.org](http://www.r-project.org).

## Appendix B: Aid to parameter selection

	$\Delta_{\text{goal}}$	Expression difference ( $\mathcal{Z}_i$ ) definition	Value of no diff. ( $\mathcal{Z}_{\text{min}}$ )	Threshold for relevant expression difference ( $\mathcal{Z}_0 \geq \mathcal{Z}_{\text{min}}$ )
<b>Options</b>	1 % 5 % 10 % 20 % 50 %  other (several may be compared)	AUC' = $  \text{AUC} - 1/2   + 1/2$ (for clinical research and other settings when sensitivity and specificity are important)  Absolute value of t–statistic or other difference–scale ratio (for biological studies in which the amount of RNA is important)	If $\mathcal{Z}_i$ is AUC', $\mathcal{Z}_{\text{min}} = 1/2$ .  If $\mathcal{Z}_i$ is the absolute value of a difference–scale ratio, $\mathcal{Z}_{\text{min}} = 0$ .	If $\mathcal{Z}_i$ is AUC', values of $\mathcal{Z}_0$ may be between 0.5 and 1.0.  If $\mathcal{Z}_i$ is the absolute value of a difference–scale ratio, values of $\mathcal{Z}_0$ may be above 0.
<b>Notes</b>	$\Delta_{\text{goal}}$ is the dFDR or the probability of a lack of differential expression for genes considered differentially expressed. 5 % is conventional, but values up to 50 % may be informative when $\mathcal{Z}_0 > \mathcal{Z}_{\text{min}}$ .	Studies concerned with classification or prediction benefit from the AUC or a related definition of expression difference, whereas studies concerned with amounts of RNA relative to the variability benefit from t–like statistics as definitions of expression difference.	Lowest $\mathcal{Z}_i$ allowed	The smallest expression difference that is relevant clinically or biologically depends on the definition of expression difference and on the goals of the researcher. The exact value chosen is not as important as the fact that the value has some scientific plausibility, since any reasonable value will have better properties than the conventional selection, $\mathcal{Z}_0 = \mathcal{Z}_{\text{min}}$ .

**Summary of options available for customizing the detection of differential expression, if desired.** The software of the author has default values for those who do not wish to change the settings. This table is provided for convenience; fuller definitions and explanations are given in the text. For the estimation of expression differences, there are also choices of  $\mathcal{Z}_i$  available, but  $\Delta_{\text{goal}}$  and  $\mathcal{Z}_0$  do not apply.

**Table 4**

## References

- Benjamini, Y. and Hochberg, Y. (1995) "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B* **57**, 289-300
- Bickel, D. R. (2003a) "Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically," *Bioinformatics* **19**, 818-824
- Bickel, D. R. (2003b) "Selecting an optimal rejection region for multiple testing: A decision-theoretic alternative to FDR control, with an application to microarrays," submitted; available from [www.arxiv.org](http://www.arxiv.org) and [www.mathpreprints.com](http://www.mathpreprints.com)
- Dudoit, S., Fridlyand, J., and Speed, T. P. (2002) "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association* **97**, 77-87
- Efron, B. and Tibshirani, R. (2002) "Empirical Bayes methods and false discovery rates for microarrays," *Genetic Epidemiology* **23**, 70-86
- Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001) "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association* **96**, 1151-1160
- Fernando, R. L., Nettleton, D., Southey, B. R., Dekkers, J. C. M., Rothschild, M. F., and Soller, M. (2003) "Controlling the proportion of false positives (PFP) in multiple dependent tests," submitted
- Genovese, C. and Wasserman, L. (2002) "Bayesian and frequentist multiple testing," unpublished draft of 4/12/02
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999) "Molecular classification of cancer: Class discovery and class prediction by gene expression modeling," *Science* **286**, 531-537
- Green, D. M. and Swets, J. A. (1966) *Signal Detection Theory and Psychophysics*, John Wiley and Sons, Inc.: New York
- Hanley, J. A. and McNeil, B. J. (1982) "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* **143**, 29-36
- Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2002) "Optimal sample size for multiple testing: the case of gene expression microarrays," Technical Report, M. D. Anderson Cancer Center; available from <http://www.ceremade.dauphine.fr/~xian/publications.html> or [http://www.ceremade.dauphine.fr/CMD\\_2002.php](http://www.ceremade.dauphine.fr/CMD_2002.php)
- Pepe, M. S., Longton, G., Anderson, G. L., and Schummer, M. (2003) "Selecting differentially expressed genes from microarray experiments," *Biometrics* **59**, 133-142
- Storey, J. D. (2002) "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B* **64**, 479-498

- 
- Storey, J. D. (2003) "The positive false discovery rate: A Bayesian interpretation and the Q-value," *Annals of Statistics* **31** (6)
- Westfall, P. H. and Young, S. S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, John Wiley & Sons: New York
- Yeoh, E.-J., Mary E. Ross, Sheila A. Shurtleff, W. Kent Williams, Divyen Patel, Rami Mahfouz, Fred G. Behm, Susana C. Raimondi, Mary V. Relling, Anami Patel, Cheng Cheng, Dario Campana, Dawn Wilkins, Xiaodong Zhou, Jinyan Li, Huiqing Liu, Ching-Hon Pui, William E. Evans, Clayton Naeve, Limsoon Wong, and James R. Downing (2002) "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell* **1**, 133-143