Contents

CHAPTER 1

Local Pooled Error Methods for Enhancing Statistical Power in Small Sample Microarray Data Analysis

Jae K. Lee[1], HyungJun Cho[1] & Michael O'Connell[2]
University of Virginia[1] & Insightful Inc.[2]

Abstract: In microarray data analysis, gene discovery based on fold-change value is often misleading because its error variability can be dramatically different among different genes under different biological conditions and intensity ranges. The standard approach is thus emerging as one based on statistical significance and hypothesis testing for each gene's differential expression, with careful attention paid to multiple comparison issues. However, when only a small number of replicated arrays are available, these approaches can often be underpowered and may result both in high false positive and false negative error rates due to their inaccurate *within-gene* error estimation. In this chapter, we introduce *local pooled error* (LPE) method for testing two comparing conditions in a microarray study and its extension to multiple condition microarray data analysis *empirical Bayes heterogenous error model* (HEM), which significantly improve statistical discovery power in microarray data analysis with limited replication by pooling and utilizing other similar genes' error information. The open source software packages of these LPE-based approaches are available as `LPE` and `HEM` at the Bioconductor web site (http://www.bioconductor.org); LPE is also available through the commercial software product S+ArrayAnalyzer^{TM}, based on S-PLUS(R).

1.1 Introduction

Each gene's differential expression pattern in a microarray experiment is usually assessed by (typically pairwise) contrasts of mean expression values among experimental conditions. Such comparisons have been routinely assessed as fold changes whereby genes with greater than two or three fold changes are selected for further investigation. It has been frequently found that a gene showing a high fold-change between experimental conditions might also exhibit high variability and hence its

differential expression may not be significant. Similarly, a modest change in gene expression may be significant if its differential expression pattern is highly reproducible. A number of authors have pointed out this fundamental flaw in the fold-change based approach (Jin et al., 2001). And, in order to assess differential expression in a way that controls both false positives and false negatives, the standard approach is emerging as one based on statistical significance and hypothesis testing, with careful attention paid to reliability of variance estimates and multiple comparison issues.

The classical two-sample t-statistic has been initially used for testing each gene's differential expression; the procedures such as the Westfall-Young step-down method have been suggested to control FWER (Dudoit et al., 2002). These t-test procedures, however, rely on reasonable estimates of reproducibility or within-gene error to be constructed, requiring a large number of replicated arrays. When a small number of replicates are available per condition, e.g. duplicate or triplicate, the use of naive, within-gene estimates of variability does not provide a reliable hypothesis testing framework. For example, a gene may have very similar differential expression values in duplicate experiments by chance alone. This can lead to inflated signal-to-noise ratios for genes with low but similar expression values. Furthermore, the comparison of means can be misled by outliers with dramatically smaller or bigger expression intensities than other replicates. As such, error estimates constructed solely within genes may result in underpowered tests for differential expression comparisons and also result in large numbers of false positives.

A number of approaches to improving estimates of variability and statistical tests of differential expression have thus recently emerged. Several variance function methods have been proposed, including a simple regression estimation of local variances (Kamb and Ramaswami, 2001) and a two-parameter variance function of mean expression intensity (Durbin et al., 2002). The variance function methods described above borrow strength across genes in order to improve reliability of variance estimates in differential expression tests. The local-pooled-error (LPE) estimation strategy has also been introduced for improving such within-gene expression error estimation (Lee and O'Connell, 2003; Jain et al., 2003). Especially, LPE variance estimates for genes are formed by pooling and smoothing the error variability of genes with similar expression intensities from replicated arrays. From this error pooling, the LPE approach effectively handles many statistical artifacts in large screening analysis, e.g. where a gene with low expression may have very low variance by chance and the resulting signal-to-noise ratio is unrealistically large, or vice versa. This, in turn, leads to a dramatically improved statistical testing framework for the discovery of biologically-relevant differential expression patterns in a microarray study: LPE test for comparing two contrasting conditions and empirical-Bayes heterogenous error model (HEM) for identifying differentially expressed genes under multiple experimental conditions.

We note that LPE methodology is conceptually similar to the probe information pooling technique described in the next Chapter that pools the error information of the number of changed probes in each local (differential-expression) intensity region.

The former directly pools the error information of the original expression intensities, whereas the latter utilizes the error information of the changed-call probe counts in each local intensity region.

The LPE-based approaches are available as open source R software packages LPE and HEM in the Bioconductor project (http://www.bioconductor.org); the LPE method is also available through the commercial software product S+ArrayAnalyzer^{TM}, based on S-PLUS(R) (O'Connell, 2003).

1.2 Local Pooled Error Test

1.2.1 Method

The Local Pooled Error (LPE) method constructs error variance estimates by pooling variance estimates for genes with similar expression intensities from replicated arrays within experimental conditions (Jain et al., 2003). LPE carefully leverages the observation that genes with similar expression intensity values often show similar variability within each experimental condition and that error variability of (log) expression is a decreasing (or non-increasing) function of intensity in practical microarray data. The latter is due to the fact that microarray instrumentation exhibits common background noise that is a bigger proportion of gene expression intensity in a low intensity region than that in a high intensity region. Figure 1.1, illustrates this phenomenon with LPE-estimated baseline error distributions under three different experimental conditions for a mouse immune-response microarray study described below.

To take into account heterogeneous error variability across different intensity ranges described above, the LPE method is applied as follows (refer to Jain et al. (2003) for a more detailed technical description). For oligo array data, let $y_{1_{ik}}$ and $y_{2_{ik}}$ be the observed expression intensities at gene i for replicate k under two conditions. For duplicate arrays, $k = 1, 2, \ldots$ plots of $A = \log_2(y_{1_{ik}} y_{2_{ik}})/2$ vs. $M = \log_2(y_{1_{ik}}/y_{2_{ik}})$ can facilitate the investigation of between-duplicate variability in terms of overall intensity. The M versus A plot (M v A or Bland-Altman plot), provides a very raw look at the data and is useful in detecting outliers and patterns of intensity variation as a function of mean intensity (Dudoit et al., 2002). At each of the local intensity regions of the M v A plot under a particular biological condition, the unbiased estimate of the local variance is obtained. A cubic spline is then fit to these local variance estimates to obtain a smooth variance function. The optimal choice of the effective degree of freedom df_λ of the fitted smoothing spline is obtained by minimizing the expected squared prediction error or generalized cross validation error (GCV). This two-stage error estimation approach—estimation of error of M within quantiles and then non-parametric smoothing on these estimates is used because direct non-parametric estimation often leads to unrealistic (small or large) estimates of error when only a small number of observations are available at a fixed-width intensity range.

Based on the LPE estimation above, statistical significance of the LPE-based test is

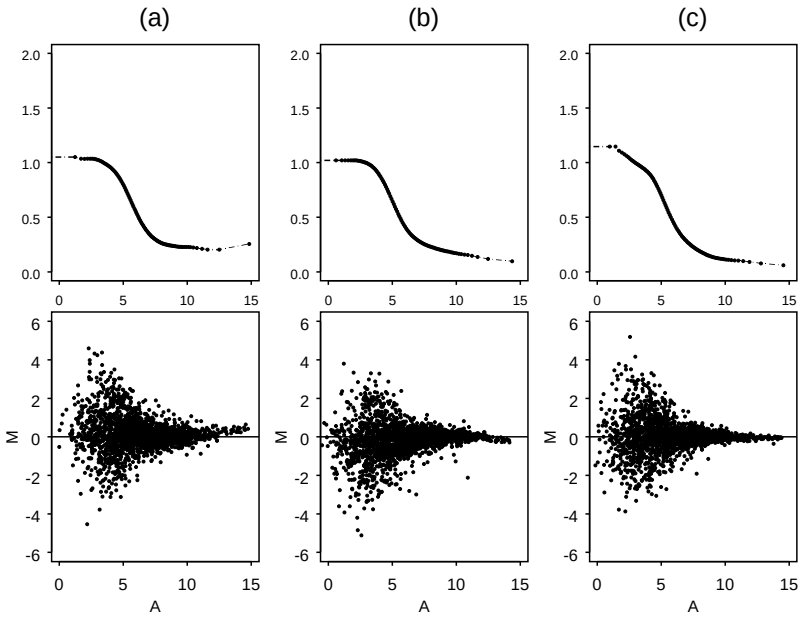

Figure 1.1 *Log intensity ratio* $\log_2(y_{1_{ik}}/y_{2_{ik}})$ *(M) as a function of average gene expression* $\log_2\sqrt{y_{1_{ik}}y_{2_{ik}}}$ *(A). Top row of panels (a), (b) and (c) represent local pooled error (LPE) for naïve, 48 hour activated, and T-cell clone D4 conditions respectively for the mouse immune response microarray study reported by Jain et al. (2003). Variance estimates in percentile intervals are shown as points, and smoothed curve superimposing these points is also shown. Bottom row of panels represents the corresponding M vs A graph. The horizontal line represents identical expression between replicates.*

evaluated as follows. First, each gene's medians under the two compared conditions are calculated to avoid artifacts from outliers. The approximate normality of medians can be assumed with a small number of replicates based on the fact that the individual log-intensity values within a local intensity range follow a normal distribution; see the supplementary data in Jain et al. (2003). The LPE statistic for the median (log-intensity) difference z is then calculated as:

$$z = (\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}_{pooled} \tag{1.1}$$

where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the median intensities in two comparing array experimental conditions Y_1 and Y_2, and $\hat{\sigma}_{pooled}$ is the pooled standard error, $[\hat{\sigma}_1^2(\hat{\mu}_1)/n_1 + \hat{\sigma}_2^2(\hat{\mu}_2)/n_2]^{1/2}$ from the LPE-estimated baseline variances of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$. The LPE approach shows a significantly better performance than two-sample t-test, SAM, and Westfall-Young's permutation tests, especially when the number of replicates is smaller than ten. In a simulation study from a Gaussian distribution without extreme outliers, the LPE method showed a significant improvement of statistical power with three and five replicates, as reported in Jain et al. (2003).

1.2.2 Example: Micorarray Study of T-cell Immune Responses

Cytotoxic T cells play a central role in the pathophysiology of many inflammatory lung diseases wherein they accumulate in the alveolar space and/or in the interstitium. A microarray study was performed to investigate the role of T cells among three different immune exposure populations: naïve (no exposure), 48-hour activated, and $CD8^+$ T cell clone D4 (long-term mild exposure), using triplicate microarrays of Affymetrix® murine chip, MG-U74A with 12,488 genes (Jain et al., 2003). Signal intensity values were obtained from the Affymetrix's MicroArray Suite software (MAS 5.0). Many genes exhibiting significant differential expression patterns were identified by the LPE test. The LPE method identified genes that are well-known in the literature for their mouse immune response function. Other hypothesis testing methods, such as the Westfall-Young procedure and the two-sample t-test, were not able to identify some of these genes.

In order to examine the relationship between the LPE p-values and fold change values more systematically, a scatter plot (or volcano plot) of each gene's LPE p-values vs. fold change comparing the naïve and the $CD8^+$ T-cell D4 clone conditions is shown in Figure 1.2. The two horizontal lines represent two fold changes in both directions and the vertical line the Bonferroni-adjusted LPE p-value 0.05. The numbers of genes in each sector of the left panel are also shown. Note that the two RNA samples—naïve and $CD8^+$ T-cell clone are biologically quite heterogeneous, and a large number of differentially expressed genes were identified both by LPE test and fold change. In this figure a weak correlation is found between significant differential expression and fold change. This suggests differential-expression discovery based on fold-change alone is misleading because a large number of insignificant genes are identified with high fold-changes in the low intensity region as displayed with the blue color in Figure 1.2(b).

1.2.3 Resampling-Based FDR Estimation for LPE Tests

An extremely large number of e.g. $>$40K genes can be represented on a microarray, and as such, comparisons between experimental conditions for all genes must be take false positive error rate and multiple comparison issues into account. In order to control the false-positive rate, traditional statistical methods often control the family-wise error rate (FWER), the probability of incorrectly accepting at least one false-positive hypothesis (or type-I error) among all hypotheses. For example, the commonly-used Bonferroni correction divides the type I error α by the total number of hypotheses for the test of each gene's differential expression, assuming the hypotheses under consideration are independent (Dudoit et al., 2002). However, this independence assumption is unlikely to be true in microarray data, as functions of many genes are interrelated in varying degrees. Moreover, the methods controlling FWER are frequently found to be too conservative to identify many important genes in biological applications. Several authors (e.g., Sidak, WestFall and Young) have developed step-down procedures that apply the severe Bonferroni correction only to

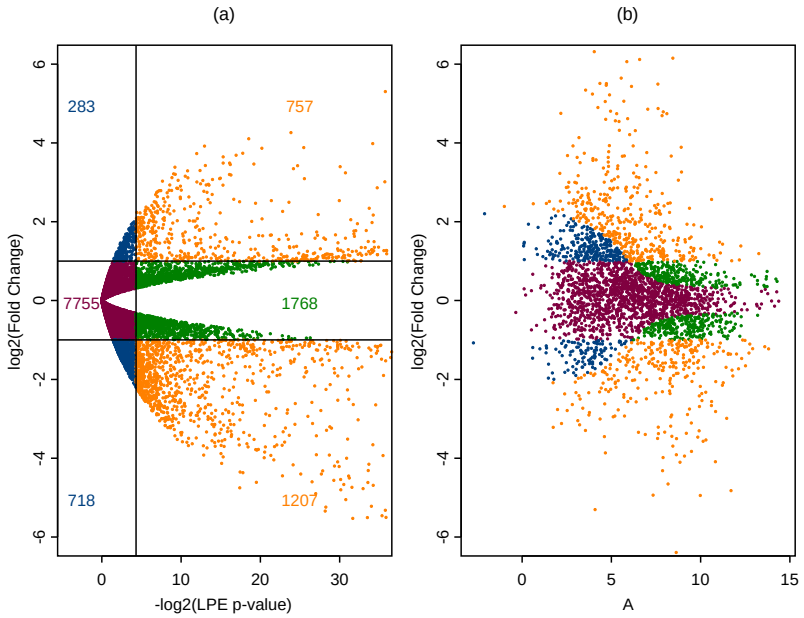

Figure 1.2 *Fold change* (log_2) *of gene expression and LPE p-values* $(-log_2)$ *for naïve mice and CD8*$^+$ *T-cell D4 clone (left panel) conditions. The two horizontal lines mark the two-fold change threshold and the vertical line marks the threshold of cutoff Bonferroni-adjusted p-value* $= 0.05$*. Genes shown in green color undergo low fold change but changes are significant—these genes are missed by fold change method alone. Genes shown in blue color have high differential expression but are not significant and would be detected as false positives by a fold change method. The right panel shows the distribution of genes in M v A format. There is no clear-cut relation between significant and high fold-change genes, and hence LPE is required for such differentiation. Numbers shown in each sector of left panel represent the number of genes in that sector.*

the most extreme value of the test statistic, and step down the correction with the value of the test statistic. However, these methods result in high false-negative error rates, likely missing many genes that are truly differentially expressed.

Benjamini and Hochberg (BH; 1995) suggested that controlling false discovery rate (FDR), or the expected proportion of false positives among all positive (or rejected) hypotheses, is more appropriate for large screening problems. Benjamini and Yekutieli (BY; 2001) proposed a new FDR procedure considering a certain dependency structure among the test statistics. However, both the BH and BY procedures may still be too conservative when applied to real microarray data analysis (Dudoit et al., 2002). This is mainly due to the fact that the independence or the artificial dependency assumptions made in these approaches may not be supported in real microarray data applications. Furthermore, microarray experiments are often conducted

with a small number of replicates due to limited availability of RNA samples and/or budgetary constraints as mentioned earlier.

One of the key issues in estimating FDR is the assumption regarding the underlying null distribution. The Significance Analysis of Microarrays (SAM) method (Tusher et al., 2001) uses a full permutation strategy, sampling across all genes and conditions to generate such a null distribution (mix-all). However, this strategy breaks many intrinsic correlation structures and does not generate a realistic or *biologically-relevant null* distribution for microarray data. Chip-by-chip permutation strategies, which randomly shuffle all the columns (chips) and preserve gene structure, are not applicable when the sample size is small because the number of independent permutations is too small to generate a null distribution with enough granularity to support desired significance calculations. To provide more stable estimation of such FDR values, a method based on the spacings LOESS histogram (SPLOSH) was also proposed based on a certain assumption about the p-value distribution (Pounds and Cheng, 2004). In order to further improve the FDR estimation in practical microarray data analysis, a *rank-invariant resampling* (RIR) approach can be applied to microarray data with a small number of replicates as follows.

Generation of biologically relevant null distribution: It is critical to generate an underlying null distribution as close as possible to real microarray data because a gene's statistical significance can be dramatically different under different underlying null distributions. Therefore, a resampling strategy needs to be designed to preserve the biological structure of each microarray data set as much as possible. Before describing this resampling strategy, an algorithm is defined for constructing intervals in the resampling strategy. A naive approach for construction of intervals is to partition intensity ranges so that each interval has an equal number of genes. This approach may yield overly large test statistics in high intensity levels because intensities are very sparse in high levels and condense in the middle levels. In order to obtain the local intervals of the genes with homogeneous variances, adaptive intervals are constructed by the following algorithm.

Adaptive Interval (AI) Algorithm

1. Estimate a baseline variance function for all data under consideration (within each experimental condition) by LPE
2. Obtain medians and variance estimates for each gene.
3. Order the medians and variances by the medians and denote the ordered medians and variances by $\xi_{(i)}$ and $\sigma_{(i)}$.
4. Obtain the first interval with threshold values $\xi_{(1)}$ and $\xi_{(1)} + \sigma_{(1)}$.
5. Obtain the next interval with $\xi_{(2)}$ and $\xi_{(2)} + \sigma_{(2)}$, where $\xi_{(2)}$ is the smallest median such that $\xi_{(2)} \geq \xi_{(1)} + \sigma_{(1)}$.
6. Repeat step 5 to obtain the next intervals with $\xi_{(i)}$ and $\xi_{(i)} + \sigma_{(i)}$, where i is the index of the smallest median such that $\xi_{(i)} \geq \xi_{(i-1)} + \sigma_{(i-1)}$ until all the data are assigned to certain intervals.

Note that the number of genes in each interval is forced to be between given minimum and maximum numbers, typically between 10 and (1/100 of the total number of genes) for the minimum and maximum numbers, respectively. Note also that this AI algorithm is applied to the replicated array data under each experimental condition separately.

The RIR procedure for generating null data is then as follows.

1. Calculate medians for each gene and obtain the ranks of these medians within each experimental condition.
2. Calculate rank differences between two conditions for each gene.
3. Construct the first intensity intervals using the AI algorithm above and retain rank-invariant genes by eliminating a certain percentage of genes with largest rank differences within each interval.
4. Construct the final intensity intervals of rank-invariant genes using the AI algorithm.
5. Obtain a set of null data by resampling intensities of rank-invariant genes within each interval.
6. Repeat the above step B times, e.g., 1,000, to obtain B independent sets of resampled null data.

In step 5 of the above procedure, a certain percentage of genes are eliminated to retain only rank-invariant expressed genes. In this current application, 50% of all genes with largest rank differences are eliminated in this step. Note that the AI algorithm is used twice in this RIR procedure; the first time to remove rank-variant genes evenly throughout the whole intensity range. Without this step, many genes in low intensity ranges would be unproportionately removed due to the larger variability in those ranges. This is a particularly important issue for Affymetrix data that have been summarized using the MAS5 procedure.

RIR-based FDR estimation: Suppose Z^0 is a LPE Z-statistic calculated from null data as described above. Generation of the null data is repeated many times independently. Let Z be a LPE Z-statistic computed from the real data. FDR at a threshold value δ can be estimated as

$$\widehat{FDR}(\Delta) = \frac{\hat{\pi}_0(\lambda)\bar{R}^0(\Delta)}{R(\Delta)}, \tag{1.2}$$

where $\bar{R}^0(\Delta)$ is the average number of Z^0 equal to or greater than Δ and $R(\Delta)$ is the number of Z equal to or greater than Δ. The proportion $\pi_0(\Delta)$ of true null genes in real data can be estimated by the number of $\{Z \leq \lambda_q\}$ divided by the average number of $\{Z^0 \leq \lambda_q\}$, where λ_q is the q-th quantile of Z^0 as suggested by Storey and Tibshirani (2003), e.g. 0.9 for q. A gene's FDR value might be estimated as zero when no gene in the resampled null data exceeds its Z; in these cases the minimum estimate of FDR is forced to be the reciprocal of the product between the numbers of genes and resampled null data sets, which is the finest resolution of this RIR FDR

Table 1.1 *Numbers of differentially expressed genes discovered by five methods*

FDR cutoff	BY	BH	SPLOSH	Mix-all	RIR
0.0001	1397	1730	2876	2542	2074
0.001	1730	2162	3134	2958	2485
0.01	2160	2849	3467	3694	3382
0.05	2670	3661	5654	4594	4548

estimation. Note that the confidence bounds for $\widehat{FDR}(\Delta)$ at each threshold value c can also be obtained from the B resampled null data sets.

Comparison with other FDR estimation methods: SAM's full permutation (or *mix-all*) strategy randomly samples all intensity values across genes and conditions to generate null data, of which FDR estimation can be similarly performed as described above for our RIR approach. Benjamini and Hochberg (BH; 1995) proposed the step-up procedure to control FDR. These approaches can be compared with our RIR approach based on the LPE statistics in the following manner. Let $z_{(1)} \geq z_{(2)} \geq \ldots \geq z_{(G)}$ be LPE z-statistics for discovery of differential expression of G genes. Denote the corresponding ordered raw p-values as $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(G)}$. BH adjusted p-values are defined as $\tilde{p}_{(i)} = \min_{k=1,\ldots,G}\{\min(p_{(k)}G/k, 1))\}$. For control of FDR at level α, a gene i is claimed as significant if $\tilde{p}_{(i)} \leq \alpha$. Thus, the BH estimate of FDR at a given critical value c can conservatively be defined as $\tilde{p}_{(i^*)}$, where i^* is $\min\{i : z_{(i)} \geq c\}$. The adjusted p-values of Benjamini and Yekutieli (BY; 2001) are defined as $\tilde{p}_{(i)} = \min_{k=1,\ldots,G}\{\min(p_{(k)}G\sum_{j=1}^{G}(1/j)/k, 1))\}$. Utilizing the full information in the p-value distribution, the SPLOSH FDR estimate is derived as $h_{(i)} = \min_{k\geq i}(r_{(k)})$, where $r_{(k)}$ is cFDR estimate of gene k. A comparison between these five FDR estimation methods and the RIR method is shown in 1.1 for the mouse immune-response microarray data.

Table 1.1 displays the numbers of the selected differentially expressed genes at FDR 0.0001, 0.001, 0.01, or 0.05. The results show that BH and BY are more conservative than others, whereas the SPLOSH and mix-all methods are more liberal than the others. Table 1.2 shows the FDR estimates for the five well-known genes that were reported and confirmed in the original study (Jain et al., 2003). The FDR estimates of several genes among them were greater than 0.01 by conservative BH and BY. One or more genes' FDR estimates were greater than 0.01 by SPLOSH and mix-all, whereas RIR identified all of these genes with FDR < 0.01.

Table 1.2 *FDR estimates of well-known genes found to be differentially regulated genes*

Gene Title	BY	BH	SPLOSH	Mix-all	RIR
CD97 antigen	.0230	.0023	.0489	$< .0001$	.0006
GATA-binding protein-3	.0208	.0021	.0489	$< .0001$	.0006
CD40 ligand transcript	.1005	.0103	$< .0001$	.0007	.0034
Granzyme K	.2768	.0277	.0524	.0037	.0091
Fas-associated factor-1	1	.1100	$< .0001$	.0335	.0038

1.3 Empirical Bayes Heterogeneous Error Model (HEM)

1.3.1 Background

Microarray experiments are often performed under multiple experimental conditions. The statistical testing methods discussed above are inefficient and restrictive for analyzing such data sets because they have to be applied to each pairwise comparison among many different combinations of the multiple conditions. Analysis of variance (ANOVA) approaches have been suggested to examine and evaluate the statistical significance of differential expression one gene at a time, controlling for the random chance of false positives among all candidate genes in microarray data (Kerr and Churchill, 2001; Woolfinger et al., 2001).

Under a Bayesian testing framework, several approaches have been developed for analyzing microarray data: Bayesian parametric modeling (Newton et al., 2001), Bayesian regularized t-test (Baldi and Long, 2001), Bayesian hierarchical modeling with a multivariate normal prior (Ibrahim et al., 2002), and Bayesian hierarchical error model with two error components (Cho and Lee, 2004). In order to improve the error estimation accuracy of large-screening microarray data, empirical Bayes (EB) techniques have also been applied (Efron et al., 2001; Newton and Kendziorski, 2003). In these cases, the Empirical Bayes priors are considered for mixture distributions of equivalently and differentially expressed genes. Certain Bayesian approaches have considered heterogeneous error variability in microarray data including the Bayesian model presented in Ibrahim et al. (2002).

However, the the error estimation in these classical and Bayesian approaches is not accurate when the number of replicated arrays is small. Furthermore, these modeling approaches are limited in that they are not able to capture heterogeneous error components accurately in microarray data due to the unidentifiability and computational restrictions of numerous error components. Consequently, these approaches do not provide a reliable statistical inference framework when the number of array replicates is very small, as is typically the case in investigating complex biological and biomedical mechanisms (Lee, 2002).

To remedy these restrictions, a heterogenous error model (HEM) approach is suggested to estimate heterogeneous technical and biological errors in microarray data

separately and accurately (Cho and Lee, 2004). In particular, using LPE-estimated empirical Bayes prior specifications, HEM takes into account the fact that these two heterogeneous error components can often be observed separately, the former at different intensity ranges and the latter for different genes and conditions. Similar to the RIR-based FDR evaluation for LPE, a resampling-based evaluation of the *false discovery rate* (FDR) is also used for HEM, fully utilizing the distributional information of the original raw data.

1.3.2 Heterogeneous Error Modeling

Suppose that y_{ijkl} is the l-th technically replicated gene expression value of the i-th gene for a particular k-th individual sample under the j-th biological condition, where $i = 1, \ldots, G; j = 1, \ldots, C; k = 1, \ldots, m_{ij}; l = 1, \ldots, n_{ijk}$. Assume that data are properly normalized and log-transformed (typically base 2). The heterogeneous error model (HEM) with two layers of error is considered as follows. HEM first separates the technical error e_{ijkl} from the observed expression value y_{ijkl} to obtain the expression value x_{ijk} free of the technical error. The first layer, thus, is defined as

$$y_{ijkl}|\{x_{ijk}, \sigma^2_{e_{ijk}}\} = x_{ijk} + e_{ijkl}. \tag{1.3}$$

The technical error e_{ijkl} is assumed to be $i.i.d.\ N(0, \sigma^2_{e_{ijk}})$, where its heterogeneous variance is defined to be a function of x_{ijk}, *i.e.*, $\sigma^2_{e_{ijk}} = \sigma^2_e(x_{ijk})$. This assumption is based on the fact that such technical error variances vary on different intensity levels in microarray data.

In the subsequent layer, expression intensity x_{ijk} is decomposed into additive effects of gene, condition, and interaction:

$$x_{ijk}|\{\mu_{ij}, \sigma^2_{b_{ij}}\} = \mu_{ij} + b_{ijk} = \mu + g_i + c_j + r_{ij} + b_{ijk}, \tag{1.4}$$

where μ is the parameter for the grand mean; g_i and c_j are the parameters for the gene and condition effects, respectively; r_{ij} is the parameter for the interaction effect of each gene-condition combination; and b_{ijk} is the error term for the biological variation, assuming $i.i.d.\ N(0, \sigma^2_{b_{ij}})$. The biological variance parameter $\sigma^2_{b_{ij}}$ is allowed to be heterogeneous for each combination of gene i and condition j because each gene can have its inherent, distinctive biological variation under a specific biological condition. Note that this two-layer HEM is conceptually similar to the two-consecutive regression fitting suggested by Woolfinger et al. (2001); HEM inference is based on the complete likelihood of the two layers, whereas the two-stage ANOVA models are separately fit in the latter. Note also that the above two-layer HEM is suitable for analyzing microarray data with both biological and experimental replicates. This HEM method is slightly modified in the later section when only one of the biological and technical replicates is available. The following section describes the LPE-derived empirical Bayes prior specifications for HEM (non-mathematical readers may skip this section).

1.3.3 LPE-based Empirical Bayes Prior Specifications

The two-layer HEM contains unobserved data as well as a large number of parameters. Most of these parameters can be efficiently estimated in a Bayesian framework, using conjugate priors such as a uniform distribution for μ and normal distributions for g_i, c_j, and r_{ij} with mean zero and variances σ_g^2, σ_c^2, and σ_r^2, respectively. Prior information is negligible or posterior distributions consistently converge to their target distribution when there are a large number of replicates. However, estimation of variance parameters, $\sigma_{b_{ij}}^{-2}$ and $\sigma_{e_{ijk}}^{-2}$, with a small number of replicates heavily depends upon the choice of priors; hence, constant gamma priors such as in Newton and kendziorski (2003) are not enough to precisely estimate heterogeneous variance parameters in this case. In order to correctly estimate heterogeneous variances in microarray data with limited replication, strong informative priors are needed. Thus, informative LPE-estimated EB priors are used with a non-constant gamma prior Gamma$(\alpha_b, \beta_{b_{ij}})$ for $\sigma_{b_{ij}}^{-2}$ with varying hyper-parameters and a non-parametric prior for $\sigma_{e_{ijk}}^{-2}$. These LPE-based EB priors are constructed as follows.

Suppose that there are two biological replicates (1 and 2) and two technical replicates (a and b) in a condition, *i.e.*, Y_{1a}, Y_{1b}, Y_{2a}, and Y_{2b}. For the technical error distribution, the variances of $Y_{1a} - Y_{1b}$, $Y_{1b} - Y_{1a}$, $Y_{2a} - Y_{2b}$, and $Y_{2b} - Y_{2a}$ are pooled to derive its baseline distribution because no biological variability between replicates a and b is involved. Similarly, for the biological error distribution, the variances of $Y_{1a} - Y_{2a}$, $Y_{2a} - Y_{1a}$, $Y_{2a} - Y_{1b}$, and $Y_{1b} - Y_{2a}$ are pooled to obtain its error distribution. Note, however, that the latter error distribution is for the *total* variance containing both technical and biological variances since both technical and biological errors are involved between replicates 1 and 2. Therefore, the biological variance estimate is obtained by LPE, subtracting the technical variance estimate from the corresponding total variance estimate at each local intensity region under their (orthogonal) independence assumption.

Based on the above modified-LPE estimates of the two error variances, the hyper-parameters of the EB priors can be defined. Specifically, the inverse of the biological variance, $\sigma_{b_{ij}}^{-2}$, is assumed to have a Gamma$(\alpha_b, \beta_{b_{ij}})$ prior satisfying $E(\sigma_{b_{ij}}^{-2}) = \alpha_b/\beta_{b_{ij}}$. It follows that $\mathrm{E}(\sigma_{b_{ij}}^2) = \beta_{b_{ij}}/(\alpha_b - 1)$ and $\mathrm{Var}(\sigma_{b_{ij}}^2) = \beta_{b_{ij}}^2/(\alpha_b - 1)(\alpha_b - 2)$. For positive expectations and variances, a value α_b is chosen such that $\alpha_b > 2$ (*e.g.* $\alpha_b = 3$). Given α_b and modified-LPE estimates $\tilde{\sigma}_{b_{ij}}^2$, $\beta_{b_{ij}}$ is obtained by the method of moment, *i.e.*, $\beta_{b_{ij}} = (\alpha_b - 1)\tilde{\sigma}_{b_{ij}}^2$. This provides Gamma priors that are dependent on each combination of gene and condition, so that a gene under a different condition has its specific error distribution for biological error. In contrast, technical error varies on different intensity levels, so that its baseline distribution can be estimated precisely. Therefore, the inverse of the technical error variance, $\sigma_{e_{ijk}}^{-2}$, is assumed to have a nonparametric prior rather than a Gamma prior. In order to fully utilize the LPE-estimated baseline distribution of technical error, a nonparametric EB prior specification is used for $\sigma_{e_{ijk}}^{-2}$ based on the following resampling algorithm:

1. Given a probability p, find sample quantiles $0 = \xi_0 < \xi_1 < \xi_2 < \ldots < \xi_Q$ of

median intensities corresponding to probabilities $0 < p < 2p < \ldots < (Q-1)p < Qp = 1$.

2. Randomly sample gene vectors with replacement at each quantile range proportionally, and obtain a dataset $D^{(b)}$ with size G.
3. Apply LPE to $D^{(b)}$ with the above quantiles, and so obtain quantile experimental variance estimates $\tilde{\sigma}^{2(b)}_{e_{qj}}$, where $q = 1, \ldots, Q$.
4. Repeat steps 2 and 3 with B times, and obtain $\tilde{\sigma}^{2(b)}_{e_{qj}}$ where $q = 1, \ldots, Q; b = 1, \ldots, B$

The above algorithm is performed for each condition for our MCMC update describe below.

1.3.4 HEM Inference

The joint probability of the observed and unobserved variables for the two-layer HEM is $\mathbb{P}r(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) = \prod_{ijkl} \phi\left(\frac{y_{ijkl} - x_{ijk}}{\sigma_{e_{ijk}}}\right) \times \prod_{ijk} \phi\left(\frac{x_{ijk} - \mu - g_i - c_j - r_{ij}}{\sigma_{b_{ij}}}\right)$, where $\boldsymbol{\theta} = (\mu, \boldsymbol{g}, \boldsymbol{c}, \boldsymbol{r}, \boldsymbol{\sigma}^2_{\boldsymbol{b}}, \boldsymbol{\sigma}^2_{\boldsymbol{e}})$ and ϕ is the density function of the standard normal distribution. With the above prior specification, the posterior distribution $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ of the unobserved data $\boldsymbol{x}$ and the parameters $\boldsymbol{\theta} = (\mu, \boldsymbol{g}, \boldsymbol{c}, \boldsymbol{r}, \boldsymbol{\sigma}^2_{\boldsymbol{b}}, \sigma^2_{\boldsymbol{e}})$, given the observed data $\boldsymbol{y}$, is proportional to

$$\mathbb{P}r(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) \prod_i \phi(g_i/\sigma_g) \prod_j \phi(c_j/\sigma_c) \prod_{ij} \phi(r_{ij}/\sigma_r) \prod_{ij} \Gamma(\sigma^{-2}_{b_{ij}}; \alpha_b, \beta_{b_{ij}}) \prod_{ijk} h(\sigma^{-2}_{e_{ijk}}),$$

where $\mathbb{P}r(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})$ is the joint probability, $\Gamma(*; \alpha, \beta)$ is the density function of a Gamma distribution with mean α/β and variance α/β^2, and h is a unknown distribution.

In order to estimate such a large number of parameters and unobserved data, the MCMC technique is used to sample the parameters or unobserved data from their posterior conditional distributions. Unobserved data and parameters except for technical error $\sigma^{-2}_{e_{ijk}}$ can be estimated by Gibbs sampling. The conditional posterior distribution of technical error $\sigma^{-2}_{e_{ijk}}$ cannot be obtained explicitly, so the Metropolis-Hastings algorithm is applied.

HEM summary statistic: Suppose that posterior estimates of parameters are obtained, as described in Section 2.2. Denote posterior estimates by $\bar{\mu}$, $\bar{g}_i$, $\bar{c}_j$, $\bar{r}_{ij}$, $\bar{\sigma}^2_{b_{ij}}$ and $\bar{\sigma}^2_{e_{ijk}}$, and let $\bar{\mu}_{ij} = \bar{\mu} + \bar{g}_i + \bar{c}_j + \bar{r}_{ij}$. Based on these posterior estimates, one still needs a guiding statistic to evaluate the significance of overall differential expression patterns. Therefore, the HEM summary statistic, *H-score* is defined by utilizing the posterior estimates:

$$H_i = \sum_{j=1}^{C} \frac{w_{ij}(\bar{\mu}_{ij} - \bar{\bar{\mu}}_i)^2}{(\bar{\sigma}^2_{b_{ij}} + \sum_{k=1}^{m_{ij}} \bar{\sigma}^2_{e_{ijk}}/m_{ij})}, \tag{1.5}$$

where $w_{ij} = m_{ij} / \sum_{j=1}^{C} m_{ij}$ and $\bar{\bar{\mu}}_i = \sum_{j=1}^{C} \bar{\mu}_{ij}/C$. The form of H-score is similar to ANOVA F-statistic; however, H-score utilizes variance estimates that are non-constant over conditions as well as genes, and separately account for change of each condition divided by its own variance. HEM H-statistic does not follow any parametric distribution. Differentially expressed genes have large H-scores, so gene selection can be performed by the magnitude of its score; A rigorous selection criterion of differentially expressed genes is detailed by Bayesian FDR evaluation below.

1.3.5 Resampling-based Bayesian FDR Evaluation for HEM

Similar to the RIR-based FDR evaluation for LPE, it is important to generate ***biologically-relevant*** null distributions of small-sample microarray data for the HEM application. In order to obtain null data simulating biological microarray data, gene and condition identities are preserved in our resampling. That is, all of $y_{ij,1,1}, \ldots, y_{ij,m_{ij},n_{ijk}}$ for gene i and condition j are sampled simultaneously for a simulated gene under each condition. For example, consider a microarray study with two conditions. Suppose gene i is selected for condition 1. For condition 2, gene i' is then selected with the normal probability for $(\mu_{i,1} - \mu_{i',2})$, so that genes with means closer to the mean of gene i are sampled more likely. Gene vectors i and i' are then combined as a gene vector in our (simulated) null data. Similarly, gene vectors are selected for all conditions simultaneously if there are multiple conditions. This strategy maintains gene and chip identities so that their corresponding variance structure can be retained.

Suppose H-statistics and and H^0-statistics are computed from raw data and generated null data, respectively. Generation of the null data is repeated B times independently. Given a critical value Δ, the estimate of Bayesian FDR is calculated by equation (1.2). In the equation, $\bar{R}^0(\Delta) = \#\{H^0_{ib} | H^0_{ib} \geq \Delta,\ i = 1, \ldots, G, b = 1, \ldots, B\}/B$ is the average number of significant genes in the null data, and $R(\Delta) = \#\{H_i | H_i \geq \Delta,\ i = 1, \ldots, G\}$ is the number of significant genes in the raw data. The estimate of a correction factor with the λ-quantile m_λ of H^0_{ib} is $\hat{\pi}_0(\lambda) = \#\{H_i | H_i \leq m_\lambda\}/\#\{H^0_{ib} | H^0_{ib} \leq m_\lambda\}$, which is required because of the different numbers of true null genes in the raw data and the null data.

1.3.6 HEM Only with One Type of Replication

The two-layer HEM above was described for microarray data when both technical and biological replicates are available. If a microarray study does not have technical replicates but has some biological replicates, the technical error distribution cannot be separately observed and two error distributions are therefore confounded. In contrast, if only technical replicates are available, the biological variability cannot be observed from the data; this kind of array experiments may be performed for examining the effects of specific biological treatments and conditions on a single subject. In these cases, HEM is reduced into a model with one layer as follows:

$$y_{ijk} | \{\mu, g_i, c_j, r_{ij}, \sigma^2_{\epsilon_{ij}}\} = \mu + g_i + c_j + r_{ij} + \epsilon_{ijk}, \tag{1.6}$$

where $i = 1, \ldots, G, j = 1, \ldots, C, k = 1, \ldots, m_{ij}$ and ϵ_{ijk} is the error term for the biological and experimental error variation, assuming $i.i.d.\ N(0, \sigma^2_{\epsilon_{ij}})$. Note that m_{ij} is the number of technical (or biological) replicates in this model. The other parameters are the same as those in the two-layer model and the $l-$subscript is suppressed in this model.

For the one-layer HEM, the joint probability of the observed variables is $\mathbb{P}r(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{ijk} \phi(y_{ijk} - \mu - g_i - c_j - r_{ij})/\sigma_{\epsilon_{ij}}$, where $\boldsymbol{\theta} = (\mu, \boldsymbol{g}, \boldsymbol{c}, \boldsymbol{r}, \boldsymbol{\sigma^2_b})$. In this case the H-score summary statistic is defined slightly differently as $H_i = \sum_{j=1}^{C} w_{ij}(\bar{\mu}_{ij} - \bar{\bar{\mu}}_i)^2/\bar{\sigma}^2_{\epsilon_{ij}}$. The LPE-based nonparametric prior distribution is used for variance parameter $\sigma^2_{\epsilon_{ij}}$ similarly as before.

1.3.7 Examples

Ionizing radiation response data: The two-layer HEM with microarray data is applied to the transcriptional response microarray data of lymphoblastoid cells to ionizing radiation of which details can be found in Tusher et al. (2001). In brief, two wild-type human lymphoblastoid cell lines (1, 2) were grown in an unirradiated state (U) or in an irradiated state (I) 4 hours after exposure to a modest dose of 5 Gy of ionizing radiation. RNA samples from each combination of the two cell lines and two states were labeled and divided into two identical aliquots (A, B) that were hybridized independently to the Affymetrix HUGENEFL GeneChip, generating eight hybridized microarrays (U1A, U1B, U2A, U2B, I1A, I1B, I2A, I2B). Signal intensity values were obtained using the Affymetrix's Microarray Suite software (MAS5) algorithm and normalized/log2-transformed data.

The scatter plots of log-expression values between the two aliquots, two cell lines, and two conditions demonstrate that larger variability exists between the two cell lines than between the two aliquots (data not shown). This implies that the biological variability is distinguishable from technical variability. Accordingly, these data are fit to the two-layer HEM with $m_{ij} = 2$ biological replicates and $n_{ijk} = 2$ technical replicates for $G = 7129$ genes under $C = 2$ biological conditions. For the prior distributions $g_i \sim N(0, \sigma^2_g)$, $c_j \sim N(0, \sigma^2_c)$, $r_{ij} \sim N(0, \sigma^2_r)$, we use $\sigma^2_g = \sigma^2_c = \sigma^2_r = 1$. The LPE-based EB prior specifications is used for a Gamma prior on biological error and a nonparametric prior for technical error. For MCMC runs, 2000 burn-ins and 10000 main iterations are executed and their updated parameters values were collected.

Figure 1.3 displays the top 100, 500, and 1000 genes on the MvA plots of $(U - I)$ against $(U + I)/2$. Using HEM, the selected genes are well distinguished from the distribution of random genes (first column, Figure 1.3) demonstrating that HEM successfully identifies differentially expressed genes by capturing heterogeneous error variability in microarray data. In contrast, using SAM (Tusher et al., 2001), the boundary between the selected and unselected genes becomes obscure as the number

of the selected genes increases and many random genes seem to be identified significantly (second column, Figure 1.3). The results of HEM and SAM differ because HEM estimates heterogeneous technical and biological variances separately, whereas SAM includes a variance stabilizing factor that is common to all genes.

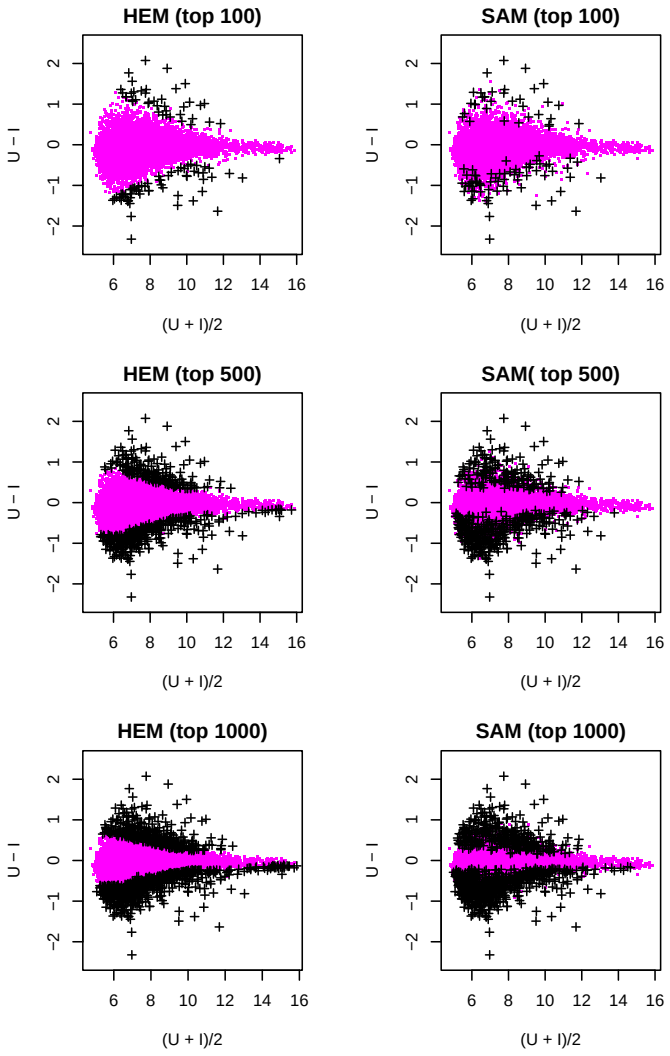

Figure 1.3 *Identification of differentially expressed genes by HEM or SAM for the ionizing radiation response data; dark (+) points represent top genes with large HEM H or SAM d-scores.*

In these data, the variability between two conditions is not as large as it is within conditions; this implies that it is difficult to identify differential expression or that

Table 1.3 *Top 10 genes selected by HEM for the T-cell immune response data*

Gene name	HEM		SAM	
	H-score	Rank	d-score	Rank
granzyme A	1362.7	1	2.02	3
ubiquitin specific protease 18	1163.3	2	1.84	8
chemokine (C-C) receptor 2	1148.7	3	1.95	7
S100 calcium binding protein A6 (calcyclin)	1142.4	4	1.56	22
disintegrin and metalloprotease domain 8	1033.8	5	1.96	5
granzyme K	984.0	6	2.12	2
cytotoxic T lymphocyte-assoc. protein2 α	969.7	7	1.46	34
chemokine (C-C motif) receptor 5	831.9	8	1.42	38
annexin A1	817.2	9	1.70	14
chemokine (C-X-C motif) receptor 3	686.6	10	1.22	72

there are a small number of differentially expressed genes. When $\widehat{FDR} = 0.01$ and 0.05, the thresholds of HEM H-scores are 2.98 and 1.69, respectively. In each of the cases, 11 and 17 genes are claimed as significant while no genes are claimed as significant under these levels of FDR by the corresponding SAM analysis. This confirms that the between-condition variation is not large in this case (the opposite case is described next).

T-cell immune response data: A one-layer HEM is applied to the mouse immune-response microarray data (Jain et al., 2003) since this experiment comprises three experimental conditions and only technical replicates. Many genes are found to be differentially expressed with a small FDR value, *e.g.*, 2464 genes are claimed as significant with an FDR less than 0.001 because variations between conditions are much larger than those within each condition. Many important genes also have large HEM H-scores. Table 1.3 displays the top ten genes selected by HEM, including their scores from HEM and SAM. The genes with the largest HEM H-scores have large SAM d-scores as well; five of which are in the top ten of SAM. Note that unlike the above ionizing radiation data example, which has both technical and biological replicates, this T-cell immune response microarray study has only technical replicates. Therefore, both HEM and SAM provide error estimates based on a single error term, so that the difference between their error estimates is relatively small in this study. However, when there exists unusual (very low or high) variability by chance among a small number of replicates for a gene (compared to other genes in the same local intensity range), its HEM error estimate can be significantly different from its SAM error estimate. Recall that the latter still is heavily weighted by the within-gene error estimate from the small number of replicates.

1.4 Conclusion

In this chapter we introduced LPE and LPE-based empirical Bayes HEM methods to accurately capture varying technical and/or biological error variances among different genes, experimental conditions, and intensity ranges in microarray data. It is difficult to estimate such numerous, heterogenous error components using the classical statistical estimation and standard Bayesian approaches in small-sample microarray data analysis. To overcome this limitation, we used advanced error-pooling techniques, such as Local Pooled Error estimation and LPE-based empirical Bayes specifications. These information-pooling approaches not only enabled us to precisely capture heterogeneous error components with limited replication, but also dramatically improved the statistical power for identification of differential expression genes, compared to widely-used SAM and ANOVA approaches in small-sample microarray data analysis.

Acknowledgement This study was supported by the American Cancer Society grant RSG-02-182-01-MGO of J.K.L.

1.5 References

P. Baldi, A.D. Long: A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes, *Bioinformatics*, **17,** (2001) 509-519.

Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, Methodological 57 (1995) 289-300.

Cho H and Lee JK: Bayesian hierarchical error model for analysis of gene expression data, Bioinformatics, 20 (2004) 2016-2025.

Dudoit, S., Yang, Y. H., Callow, M. J., Speed, T. P.: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, **12, 1,** (2002) 111-139.

Durbin, B. Hardin, J., Hawkins, D. and Rocke, D.: A Variance-Stabilizing Transformation for Gene-Expression Microarray Data, *Bioinformatics*, **18** Number Supplemental 1 (2002) S105-S110.

Efron, B and Tibshirani, R. and Storey, J.D. and Tusher, V.: Empirical Bayes Analysis of a Microarray Experiment, Journal of American Statisticl Association, 96 (2001) 1151-1160.

Ibrahim, J.G. and Chen, M.-H. and Gray, R.J.: Bayesian Models for Gene Expression with DNA Microarray Data, Journal of American Statistical Association, 97 (2002) 88-99.

Jain, N., Ley, K., Thatte, J., O'Connell, M., and Lee, J.K.: Local pooled error test for identifying differentially expressed genes with a small number of replicated microarrays, Bioinformatics, 19(15) (2003) 1945-51.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G. and Gibson, G.: The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster, *Nature Genetics*, **29,** (2001) 389-395.

Kamb, A., Ramaswami, A.: A simple method for statistical analysis of intensity differences in microarray-derived gene expression data. *BMC Biotechnology*, (2001) 1-8.

Kerr, M.K., Churchill, G.A.: Statistical design and the analysis of gene expression microarray data. Genetic Research 77 (2001) 123-128.

Lee, J. K.: Discovery and validation of microarray gene expression patterns, *LabMedica International*, **19, 2** (2002) 8-10.

Lee, J.K., O'Connell, M.: An S-PLUS library for the analysis of differential expression in *The Analysis of Gene Expression Data: Methods and Software*, edited by G. Parmigiani, ES Garrett, RA Irizarry, SL Zeger. Springer-Verlag, ch. 7. Springer-Verlag 2003.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., Tsui, K.W.: On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data . Journal of Computational Biology (2001) 8 37-52.

Newton, M.A. and Kendziorski, C.K.: Parametric Empirical Bayes Methods for Microarrays, edited by Parmigiani, Garrett, Irizarry, and Zeger. Springer: New York 2003

OConnell, M.: Differential Expression, Class Discovery and Class Prediction using S-PLUS and S+ArrayAnalyzer, SIGKDD Explorations, December 2003, vol 5, number 2.

Pinkel D: Cancer cells, chemotherapy and gene clusters. Nat. Genet. 24(3) (2000) 208-9.

Pounds, S. and Cheng, C.: Improving false discovery rate estimation. *Bioinformatics*, **20** (2004) 1737-1745.

Storey J. and Tibshirani, R.: SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays, in *The Analysis of Gene Expression Data: Methods and Software*, edited by G Parmigiani, ES Garrett, RA Irizarry, SL Zeger. Springer-Verlag, ch. 12., Springer-Verlag 2003

Tusher V., Tibshirani R. and Chu C.: Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, Proceedings of the National Academy of Sciences 98 (2001) 5116-21

Wolfinger, R.D. and Gibson, G. and Wolfinger, E.D. and Bennett, L. and Hamadeh, H. and Bushel, P. and Afshari, C. and Pales, R.S.: Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models, Journal of Computational Biology, 8 (2001) 37-52.