CHALMERS | GÖTEBORG UNIVERSITY

PREPRINT 2007:32

Empirical Bayes Models for Multiple Probe Type Arrays at the Probe Level

MAGNUS ÅSTRAND
PETTER MOSTAD
MATS RUDEMO

Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
GÖTEBORG UNIVERSITY
Göteborg Sweden 2007

Preprint 2007:32

Empirical Bayes Models for Multiple Probe Type Arrays at the Probe Level

Magnus Åstrand, Petter Mostad, Mats Rudemo

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden
Göteborg, October 2007

Preprint 2007:32
ISSN 1652-9715

Matematiska vetenskaper
Göteborg 2007

Empirical Bayes models for multiple probe type arrays at the probe level

Magnus Åstrand, Petter Mostad, Mats Rudemo

October 1, 2007

Abstract

When analyzing microarray data a primary objective is often to find differentially expressed genes. With empirical Bayes and penalized t-tests the sample variances are adjusted towards a global estimate, producing more stable results compared to ordinary t-tests. However, for Affymetrix type data a clear dependency between variability and intensity-level generally exists, even for logged intensities, most clearly for data at the probe level but also for probe-set summarizes such as the MAS5 expression index. As a consequence, adjustment towards a global estimate results in an intensity-level dependent false discovery rate. We propose two new methods for finding differentially expressed genes, Probe level Locally moderated Weighted median-t (PLW) and Locally Moderated Weighted-t (LMW). Both methods use an empirical Bayes model taking the dependency between variability and intensity-level into account. A global covariance matrix is also used allowing for differing variances between arrays as well as array-to-array correlations. PLW is specially designed for Affymetrix type arrays (or other multiple-probe arrays). Instead of making inference on probe-set summaries, comparisons are made separately for each perfect-match probe and are then summarized into one score for the probe-set. The proposed methods are compared to 11 existing methods using five spike-in data sets. PLW has the most accurate ranking of regulated genes in four out of the five data sets, and LMW consistently performs better than all examined moderated t-tests when used on RMA expression indexes. Supplementary material together with the R package plw with functions for LMW and PLW is available at http://www.math.chalmers.se/∼astrandm.

1 Introduction

Microarrays are widely used for measuring gene expression in biomedical research. For the purpose of finding differentially expressed genes there exist numerous methods. In early studies genes where often ranked with respect to fold-change. Genes showing fold-change above 2 (or 3) were regarded as potentially regulated and were selected for further investigation. The obvious drawback with such an approach, as pointed out by many authors, is that genes with high fold-change may also be highly variable and thus with low significance of the regulation. On the other hand, since the number of replicates in many studies is small, variance estimators computed solely within genes are not reliable in that very small values can occur just by chance. As a consequence the ordinary t-test suffers from low power and is not a better option for filtering out regulated genes.

Many methods have been proposed to improve on the variance estimator in order to find more powerful statistical tests for differential expression. In empirical Bayes methods (Baldi and Long 2001, Lönnstedt and Speed 2002, Smyth 2004, Kristiansson et al. 2005, Sjögren et al. 2007) and the penalized t-test suggested by (Opgen-Rhein and Strimmer 2007), the gene-specific variance estimator is modified in order to produce more stable results. With proportions determined

by the accuracy of the gene-specific variance estimators, a mixture of the gene-specific variance estimator and a global variance estimate is used in place of the gene-specific variance estimator in the denominator of the t-test. Similarly, in Efron et al. 2001, and in the Significance Analysis of Microarrays (SAM) method (Tusher et al. 2001) a constant is added to the gene-specific sample standard deviation.

Another approach is to pool variance estimators for genes having similar expression level, thus modeling the variance as a function of intensity-level. For example Eaves et al. 2002, use a weighted average of the gene-specific variance estimator and a pooled estimate based on the 500 genes with most similar mean expression level, and Jain et al. 2003, suggest the local-pooled-error method (LPE) where a variance function fitted to estimated variances and mean intensities is used. Comander et al. 2004, pool genes with respect to minimum intensity rather than mean intensity, and Hu and Wright 2007, use a hierarchical model with a linear relationship between variance and intensity-level. Of these four methods, only the one suggested by Hu and Wright 2007, takes the accuracy of the gene-specific variance estimators into account when setting the weights for the gene-specific estimator and the pooled estimator, respectively. On the other hand Hu and Wright 2007, only deal with a linear relationship between variance and intensity-level.

The type of arrays considered in this article is the Affymetrix GeneChip arrays. These arrays are one color arrays and each gene is represented by a set of probes, the probe-set, consisting of 10-16 probe-pairs. Each probe-pair consists of one perfect match (PM) probe and one mismatch (MM) probe. The probes are 25 bases long and the PM and MM probes have identical sequences of bases except for the middle probe which in the MM probe is set to the complementary base of that in the PM probe. The MM probes are thus designed to measure the background intensity for the corresponding PM probe. The standard way of dealing with the multiple-probes is to derive a summary measurement, an expression index, for each probe-set (gene) and array (sample), for example using the RMA method (Irizarry et al. 2003) or the Affymetrix MAS5 algorithm. The expression indexes are then used in downstream analysis by only considering the expression index itself, the precision of the expression index is ignored. However, in the fully Bayesian probe-level BGX model (Hein et al. 2005) information about the accuracy of the expression index is obtained as a complete distribution which is subsequently used when computing the posterior distribution of differential expression. Also, the probe-level measurement error from the probabilistic probe-level model multi-mgMOS (Liu et al. 2005) is used when computing the probability of positive log-ratio in the PPLR method (Liu et al. 2006).

For Affymetrix type arrays a dependency between variability and intensity-level generally exists, even for log-transformed data. Figure 1 shows scatter plots of sample variance versus sample mean calculated on logged PM intensities (background corrected and normalized) and three different expression indexes: RMA, GCRMA and MAS5. Except for the RMA expression index a clear dependency between variability and intensity-level exists, with a unique signature for each type of pre-processing of the raw CEL-file data. The GCRMA expression index shows increasing variability with intensity-level while MAS5 shows the opposite relationship. As a consequence, methods assuming constant variance as well as methods adjusting the gene-specific variance (or standard deviation) estimators towards a global estimate suffer from intensity-level dependent false discovery rates. Figure 2 shows an example where the moderated t-test in the R-package LIMMA (Smyth 2004) was used on MAS5 expression indexes computed on a set of replicated arrays. The false discovery rate obtained with LIMMA follows the same pattern as in the right lower panel in Figure 1 where the same data set is used. In short, Figures 1 and 2 may be seen as a motivation for the proposed methods.

In this paper the hierarchical Bayesian model proposed by Kristiansson et al. 2005, is extended to incorporate the variability to intensity-level dependency. The Probe level Locally moderated Weighted median-t method (PLW) applies the extended model to logged PM inten-

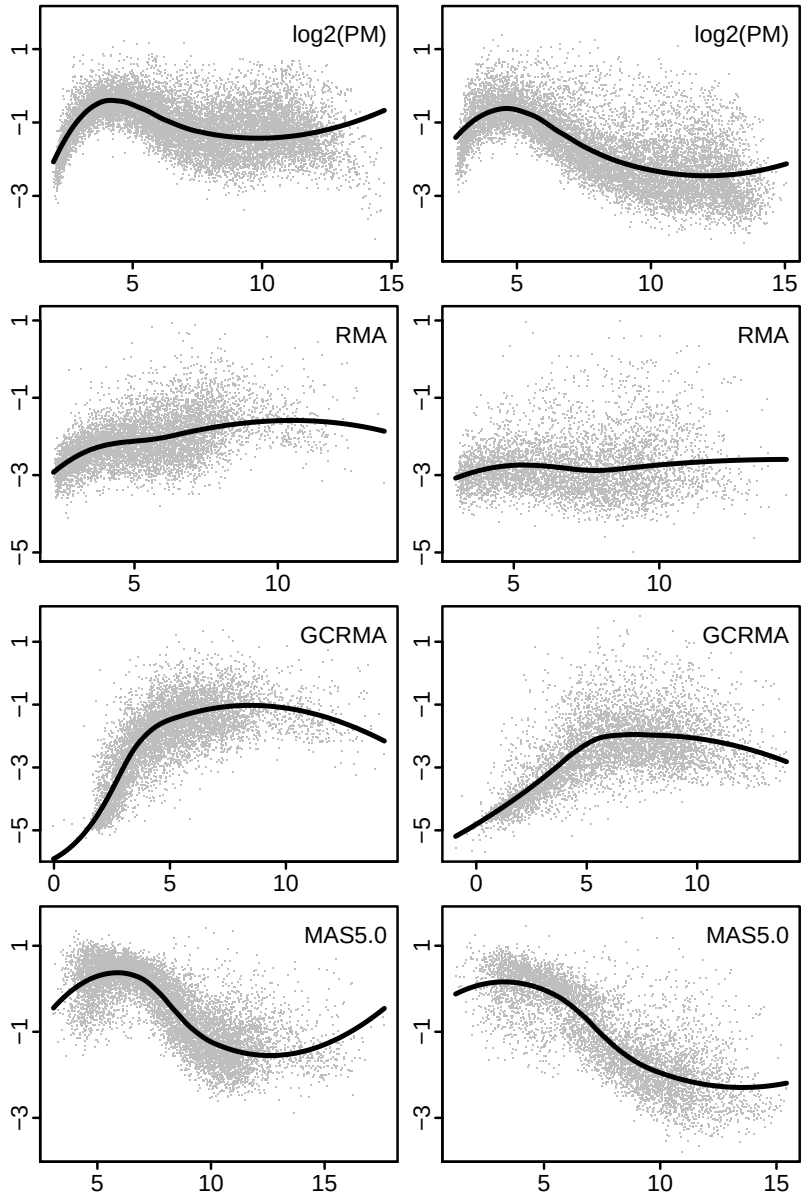

Figure 1: Scatter plots of sample variance (logged with base 2) against mean intensity for logged PM intensities and three expression indexes. Left and right panels show data set A and B, respectively (see Section 4.1).

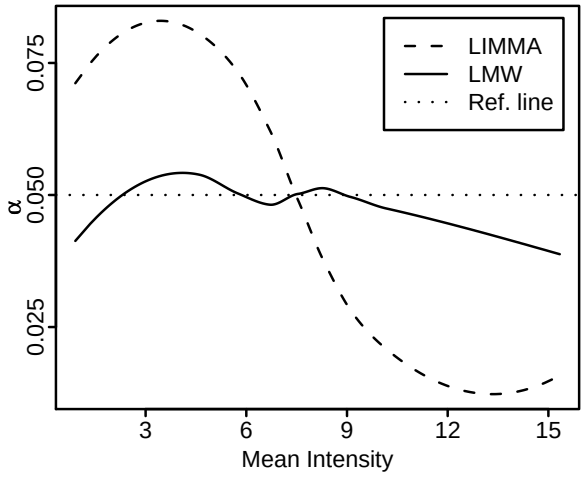

Figure 2: False discovery rate (α) calculated on re-sampled data and plotted against mean intensity. Data sets of size 6 were sampled from the complete data set B (see Section 4.1) of 18 replicated arrays and then analyzed using the Affymetrix MAS5 algorithm followed by a two group analysis of 3+3 arrays using the moderated t-test in the R-package LIMMA (Smyth 2004) and the proposed method LMW. False discovery rate were obtained by averaging over the sampled data sets using loess-curves fitted to mean intensity and indicator of significance (1 if the probe-set is among the 5% probe-sets with highest absolute statistic, 0 otherwise.)

sities resulting in moderated and weighted t-statistics for all PM probes. In the final step of PLW the median t-statistic of all PM probes building up each probe-set is computed, and this median is the value used for ranking the probe-sets with respect to differential expression.

The Locally Moderated Weighted-t method (LMW) is a more general method intended for single probe type of arrays or summary measures of multiple probe type arrays, such as RMA and MAS5. LMW use the same model as PLW but since only one t-statistic is obtained for each probe-set no median is calculated. The proposed methods are compared with existing methods on five publicly available spike-in data sets.

2 Methods

Given a set of n arrays let y_{ip} be the background corrected and normalized log-intensity on array i for PM probe p and put $y_p = (y_{1p}, \ldots, y_{np})^T$. The PM probes are divided into G (disjoint) probe-sets $\mathcal{G}_1, \ldots, \mathcal{G}_G$ and thus there are a total of $P = |\mathcal{G}_1| + \cdots + |\mathcal{G}_G|$ probes. For $p = 1, \ldots, P$ assume

$$\begin{aligned} y_p | c_p &\sim \mathrm{N}_n(\mu_p, c_p \Sigma) \\ c_p &\sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m \cdot \nu(\bar{\mu}_p)) \end{aligned} \tag{1}$$

where μ_p is the log-intensity profile for probe p across the n arrays with mean log-intensity level $\bar{\mu}_p$, Σ is an $n \times n$ covariance matrix, m is a real-valued parameter, and $\nu(\cdot)$ is a smooth real-valued function. N_n denotes an n-dimensional normal distribution, and $\Gamma^{-1}(a, b)$ denotes the inverse-gamma distribution with shape parameter a and scale parameter b. A cubic spline is used to parameterize the function $\nu(\cdot)$. Given set of K interior spline-knots

$$\nu(x) = \exp\{H(x)^T \beta\}$$

where β is a parameter vector of length $2K - 1$ and $H : \mathbb{R} \to \mathbb{R}^{2K-1}$ is a set of B-spline basis functions, see chapter 5 of Hastie et al. 2001.

As in the model suggested by Kristiansson et al. 2005, model (1) makes use of a global covariance matrix, thus allowing differing variances as well as correlations between arrays. To account for the dependency between variability and intensity-level the scale-parameter of the Γ^{-1}-distribution depends on the mean log-intensity level $\bar{\mu}_p$ for the probe through the smooth function ν.

We assume that the vector μ_p is determined by a full rank $n \times k$ design matrix D and a parameter vector γ_p of length k. The aim is to estimate and test hypothesis for δ_p, a linear combination of γ_p specified by a $1 \times k$ matrix C. In summary,

$$\mu_p = D\gamma_p \quad \text{and} \quad \delta_p = C\gamma_p \ .$$

For the special case of comparing two conditions, with n_1 and n_2 arrays from conditions 1 and 2, respectively, the design matrix D is an $(n_1 + n_2) \times 2$ matrix. For example, with $n_1 = 3$ and $n_2 = 4$ we can use

$$D^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \gamma_p = \begin{bmatrix} \gamma_{p1} \\ \gamma_{p2} \end{bmatrix}$$

and thus $\mu_p = (\gamma_{p1}, \gamma_{p1}, \gamma_{p1}, \gamma_{p2}, \gamma_{p2}, \gamma_{p2}, \gamma_{p2})^T$. With $C = [-1 \ \ 1]$ we have $\delta_p = \gamma_{p2} - \gamma_{p1}$, thus δ_p is the logged fold change between conditions 2 and 1.

However, instead of estimating the parameters of model (1) we use a reduced model derived from model (1) through a linear transformation of the vector y_p. Define the $n \times n$ and $n \times 1$ matrices

$$A_0 = I - D(D^T D)^{-1} D^T \quad \text{and} \quad B = D(D^T D)^{-1} C^T \ .$$

Since A_0 is of rank $n - k$ only we let A be an $n \times (n - k)$ matrix whose column space equals that of A_0. With $q = n - k + 1$ form the $n \times q$ transformation matrix M and the vector z_p of length q

$$M = [A; B] \quad \text{and} \quad z_p = M^T y_p \tag{2}$$

giving the reduced model

$$\begin{aligned} z_p | c_p &\sim \mathrm{N}_q\Big((0, \ldots, 0, \delta_p)^T, c_p \Sigma_z\Big) \\ c_p &\sim \Gamma^{-1}(\tfrac{1}{2}m, \tfrac{1}{2}m \cdot \nu(\bar{\mu}_p)) \end{aligned} \tag{3}$$

where $\Sigma_z = M^T \Sigma M$.

Model (3) is fitted using the EM algorithm (Dempster et al. 1977) as described in Section 3. The c_p's are treated as missing data and we replace the unknown intensity-level for probe p, $\bar{\mu}_p$, with the observed mean intensity across arrays, $\bar{y}_p$. Given estimators of the parameters Σ_z, m, and β we proceed as if these parameters are known, and weighted moderated t-tests are computed for each probe p. The unbiased minimum variance estimator of δ_p is

$$\hat{\delta_p} = (\lambda^T \Sigma_z^{-1} \lambda)^{-1} \lambda^T \Sigma_z^{-1} z_p \tag{4}$$

where λ is the vector $(0, \ldots, 0, 1)^T$ of length q. The weighted moderated t-statistic is defined as

$$\tilde{t}_p = \sqrt{\frac{q + m - 1}{(\lambda^T \Sigma_z^{-1} \lambda)^{-1}}} \frac{\hat{\delta_p}}{\sqrt{m \exp\{H(\bar{y}_p)^T \beta\} + \mathrm{RSS}_p}} \tag{5}$$

and under H_0: $\delta_p = 0$ it can be shown that $\tilde{t}_p$ is t-distributed with $q + m - 1$ degrees of freedom. Here

$$\mathrm{RSS}_p = z_p^T \Big(\Sigma^{-1} - \Sigma^{-1} \lambda (\lambda^T \Sigma^{-1} \lambda)^{-1} \lambda^T \Sigma^{-1}\Big) z_p \tag{6}$$

is the weighted residual sum of squares. See Kristiansson et al. 2006, for details. The PLW statistic for the probe-set $\mathcal{G}$ is then defined as

$$\mathrm{PLW}_{\mathcal{G}} = \mathrm{median}\Big\{\tilde{t}_p : p \in \mathcal{G}\Big\} \ . \tag{7}$$

3 Parameter estimation

The $q \times q$ covariance matrix Σ_z of model (3) is divided according to

$$\Sigma_z = \begin{bmatrix} \Sigma_A & \Sigma_{AB} \\ \Sigma_{AB}^T & \sigma_B^2 \end{bmatrix}$$

where Σ_A is the covariance matrix for all but the last dimension of z_p and σ_B^2 is the variance of the last dimension (indexes A and B refer to the corresponding sub-matrices of the transformation matrix M in (2)). Model (3) is fitted in two steps. First the parameters m, β and the sub-matrix Σ_A are estimated by dropping the last dimension of the vectors z_p. Since model (3) is not identifiable without a restriction on the function ν or the covariance matrices Σ_z we use the restriction $\text{trace}(\Sigma_A) = q - 1$. Secondly, the parameters m and β are held fixed and Σ_z is estimated using the complete z_p vectors. Temporarily the assumption of no regulated genes is used ($\delta_p = 0$ for all probes) and Σ_z is estimated under the restriction that the trace of the Σ_A part should be equal to $q - 1$.

In step 1, we let x_p denote the sub-vector of z_p obtained by dropping the last element. Under model (3) x_p is distributed according to model (1) with $\Sigma = \Sigma_A$, $\mu_p = 0$, $n = q-1$, and using the EM-algorithm an iterative procedure for estimating m, β and Σ_A is obtained. Given estimates of the previous iteration, m_0, β_0 and Σ_{A0}, updated estimates are found as follows. Let

$$w_p = \frac{m_0 + q - 1}{x_p^T \Sigma_{A0}^{-1} x_p + m_0 \exp\{H(\bar{y}_p)^T \beta_0\}} \ .$$

The updated estimate of Σ_A is

$$\hat{\Sigma}_A = \frac{1}{P} \sum_{p=1}^{P} w_p x_p x_p^T \tag{8}$$

and the updated estimate of β is found by numerical maximization of the function

$$h(\beta) = \frac{1}{P} \sum_{p=1}^{P} \Big(H(\bar{y}_p)^T \beta - w_p \exp\{H(\bar{y}_p)^T \beta\} \Big) \ .$$

With $\hat{\beta}$ equal to the updated estimate of β let

$$S = h(\hat{\beta}) + \psi\left(\frac{m_0 + q - 1}{2}\right) - \log(m_0 + q - 1) + \frac{1}{P} \sum_{p=1}^{P} \log(w_p)$$

where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function. The updated estimate of m is then found using numerical maximization of the function

$$f(m) = m\Big(\log(m) + S \Big) - 2 \log\Big(\Gamma(m/2) \Big) \ .$$

In step 2 a similar iterative procedure is used to estimate Σ_z. With Σ_{z0} denoting the estimate of Σ_z from the previous iteration and with w_p re-defined as

$$w_p = \frac{\hat{m} + q}{z_p^T \Sigma_{z0}^{-1} z_p + \hat{m} \exp\{H(\bar{y}_p)^T \hat{\beta}\}}$$

where $\hat{m}$ and $\hat{\beta}$ are the estimates obtained in step 1, an updated estimate of Σ_z is computed according to (8) with z_p replacing x_p. In order for the estimators of Σ_A and Σ_z, in step 1 and 2, respectively, to comply with the trace restriction the updated estimates are scaled at the end of each iteration. See supplementary material for more details.

4 Results and discussion

4.1 Data sets

The two data sets used in Figures 1 and 2 are publicly available at the Gene Expression Omnibus repository (http://www.ncbi.nlm.nih.gov/geo/) with series or sample reference number indicated below. Data set A consists of the 18 arrays from the severe group of the COPD data set (Spira et al. 2004) (series reference number GSE1650), where Affymetrix arrays of type HG U133A were used. In data set B the 18 arrays with normal tissue where selected from a lung tumor data set (Stearman et al. 2005) (sample reference numbers GSM47958-GSM47976, excluding GSM47967). Here the HG-U95A arrays were used.

Five spike-in data sets were used to evaluate the proposed methods. In the Affymetrix U95 and 133A Latin Square data sets (http://www.affymetrix.com/support/datasets.affx) arrays of type HG-U95A and HG-U133A, respectively, were used. The Affymetrix U95 data set consists of data from 59 arrays divided into 19 groups of size 3, and one group of size 2. From the 20 groups there are 178 possible pair-wise group comparisons each with 16 (Cope et al. 2004) known differentially expressed genes among the 12626 genes present on the arrays. The Affymetrix 133A data set comprise data from 42 arrays with a total of 22300 probe-sets of which 42 were spiked in at known concentration. The 42 arrays are divided into 14 groups of size 3 and thus there are 91 possible pair-wise group comparisons. As done in the Affycomp II assessments (Cope et al. 2004) we exclude 271 probe-sets which are likely to cross-hybridize to spike-in probe-sets. The sequence of each spike-in clone was blasted against all HG-U133A target sequences ($\sim$600bp regions from which probes are selected). A threshold of 100bp identified 271 probe-sets which are available in the affycomp R-package.

From the Gene Logic (http://www.gene logic.com/newsroom/studies/) Tonsil and AML data sets all groups with 3 replicated arrays were used, giving a total of 12 and 10 groups, respectively. For these data there are 11 genes spiked in at known concentration, which can be studied in 66 and 45 pair-wise group comparisons, respectively. Both data sets were obtained using the Affymetrix HG-U95A arrays having 12626 genes.

The Golden Spike (Choe et al. 2005) consists of 6 arrays of type Drosgenome1 divided into 2 groups of equal size. The samples used in this experiment consist of mRNA from 3866 genes, of which 1331 are differentially expressed between the groups. The Drosgenome1 array has a total of 14010 genes, thus 10144 of these should not be expressed, 2535 should be expressed but not regulated, and 1331 should be expressed and regulated. It has been observed that the 2535 genes make these data atypical (Irizarry et al. 2006). We have chosen to exclude the 2535 genes from the analysis, thus only using 11475 genes of which 1331 are known to be regulated. Also, since all 1331 genes are up-regulated it is necessary to take special care in the normalization. In place of the quantile normalization method, the default for the RMA method, we used the contrast normalization (Åstrand 2003) and fitted the normalization curve using PM probes from the 11475 unregulated genes only. To present comparable result only, methods relying on a normalization procedure at the probe-set level (PPLR and BGX) and methods for which the default normalization method is not the quantile method (logit-t) were excluded when analyzing this data set.

4.2 Comparison with existing methods

Using the spike-in data sets listed above the proposed methods, PLW and LMW were compared with 11 existing methods for ranking genes. The 11 methods include ranking with respect to: observed fold change (FC), ordinary t-test, the moderated t-test in the R-package LIMMA (Smyth 2004), the weighted moderated t-test in the R-package WAME.EM (Åstrand et al. 2007), Efron's penalized t-test (Efron et al. 2001) and the Shrink-t method (Opgen-Rhein and Strimmer 2007) in the R-package st, the SAM method (Tusher et al. 2001) in the R-package

Table 1: Area under ROC curves up to 100 false positives rounded to nearest integer value with an optimum of 100. Numbers within parenthesis are within data set ranks for the methods compared. Methods are ordered with respect to mean rank across data sets.

Method	Affy-metrix U95	Affy-metrix 133A	Golden Spike	Gene Logic Tonsil	Gene Logic AML
PLW	96(1)	93(5)	40(1)	87(1)	86(1)
LMW	96(2)	94(1)	32(3)	84(3)	80(4)
LPE	94(4)	93(9)	38(2)	84(2)	85(2)
WAME	95(3)	94(2)	32(7)	81(5)	78(7)
Efron-t	94(5)	93(3)	32(5)	79(7)	79(5)
FC	92(10)	93(4)	31(9)	83(4)	85(3)
LIMMA	94(6)	93(6)	32(8)	76(8)	75(8)
logit-t	94(8)	92(10)	-(-)	80(6)	79(6)
SAM	94(7)	93(7)	32(5)	74(10)	74(9)
Shrink-t	94(9)	93(8)	32(4)	75(9)	73(10)
PPLR	88(11)	90(11)	-(-)	71(11)	69(11)
t-test	85(12)	86(12)	25(10)	57(12)	52(12)
# of genes	12626	22029	11475	12626	12626
# of spikes	16	42	1331	11	11
# of groups	20	14	2	12	10

samr, and the Local-pooled-error test (Jain et al. 2003) in the R-package LPE. All these methods (including LMW) were applied to RMA expression indexes obtained using the R-package affy, while PLW was applied to logged PM intensities, background corrected and normalized using the default methods of RMA. With LMW 4-6 spline-knots (depending on the number of probe-sets) were used for the function ν, whereas 12 knots were used in PLW (the spline-knots are set using an internal function in the R package plw). Note that the RMA method was applied only to the arrays involved in each group comparison, as opposed to running the RMA method using all arrays of each data set.

We also compared with the PPLR method (Liu et al. 2006) applied to the expression index and probe-level measurement error of the multi-mgMOS model (Liu et al. 2005) available in the R-package puma, the logit-t procedure implemented in the R-package plw according to the description in Lemon et al. 2003, and the BGX method (Hein et al. 2005) as implemented in the R-package bgx.

Due to long computer run times the comparison with the BGX method is restricted to the Gene Logic AML data set using a subset of probe-sets only (the run time for one single analysis of 6 arrays with all 12626 probe-sets is more than 24 hours). The subset of size 1011 consists of probe-sets number 6000-7002 (excluding 6030, 6367, and 6463) together with the 11 spiked probe-sets and the same subset was used in Hein et al. 2005. The probe-set numbering is as obtained when loading data into R using the R-package affy.

For each spike-in data set and method ROC-curves were calculated. Also, for the analysis using a complete set of probe-sets, the area (AUC) under the ROC curve up to 25, 50, 100 and 200 false positives was computed. In the comparison with BGX using only 1011 probe-sets, AUC was computed up to 2, 4, 8 and 16 false positives in order to cover the same false positive range as for the complete probe-set comparisons.

ROC curves for a subset of the compared methods are found in Figure 3 and AUC values up to 100 false positives from the complete probe-set analysis are found in Table 1 (ROC curves for

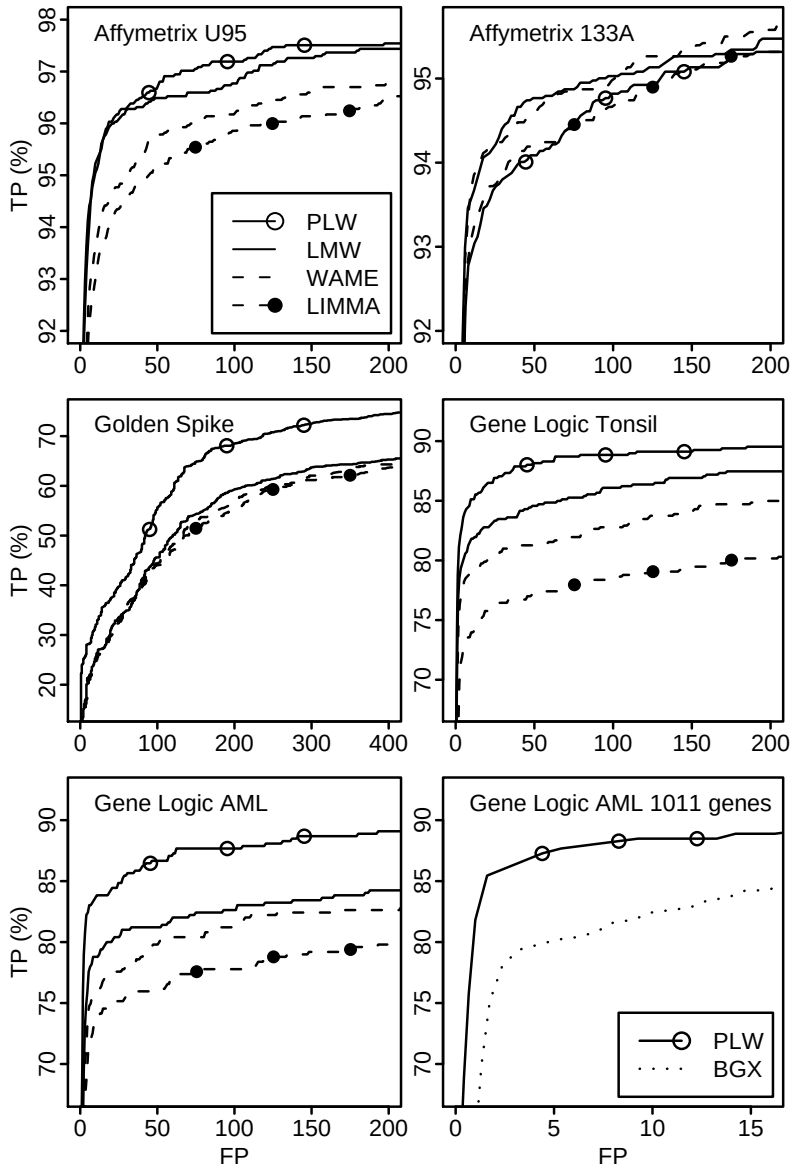

Figure 3: ROC curves for a subset of the compared methods. The horizontal axis shows the number of false positives (FP) and the vertical axis the proportion of true positives found (TP).

all methods and AUC up to 25, 50, 200 false positives are available as supplementary material). Overall the methods taking the variability-to-intensity-level dependency into account (PLW, LMW and LPE) perform better than the other methods, with the proposed method PLW having the highest AUC on four of the five data sets. Ranking genes with respect to FC performs quite well on the Affymetrix U133A and the two Gene Logic data sets but not on the other two data sets. Among the penalized and moderated t-test methods, WAME and Efron-t consistently perform better than the other ones. However, the difference between these methods for the two Affymetrix Latin Square and the Golden Spike data sets are small, compared to the difference in AUC obtained using the two Gene Logic data sets. Thus, the two Gene Logic data sets appear slightly different from the other three. The PPLR method based on the multi-mgMOS model (Liu et al. 2005) was ranked as number eleven with only the ordinary t-test having lower AUC values.

The ROC curves obtained using the subset of 1011 probe-sets from the Gene Logic AML data set are found in the lower right panel of Figure 3. The PLW method shows consistently higher true positive rate compared with BGX and the AUC up to 8 false positives (scaled so that optimum is 100) is 84 and 75 for PLW and BGX, respectively.

The second proposed method LMW differs from existing moderated and penalized t-test in that the global variance estimator (which gene-specific estimators are adjusted towards) varies with intensity-level. Actually this is the only difference between LMW and the WAME method. The LPE method also uses a global variance estimator that varies with intensity-level. But opposed to using a weighted mean of the global and gene-specific estimator, only the global estimator is used in the denominator of the LPE statistic. Thus for genes with similar intensity-level, LPE is basically identical to ranking using fold change. Hence, since LMW consistently performs better than WAME, and LPE has higher AUC than fold change in four of the five data test, modeling the global variance estimator as a function of intensity is worthwhile doing. Further, having in mind that the GCRMA and MAS5 expression indexes showed a clear dependency between variability and intensity-level in Figure 1, whereas the RMA expression index only showed a weak dependency, this kind of variance modeling might be even more important in analysis based on GCRMA or MAS5 expression indexes.

Also, Figure 2 shows that the false discovery rate obtained by adjusting towards a global estimate that varies with intensity-level results in a much more stable false discovery rates compared to using a (truly) global estimate.

Both logit-t and PLW do inference on background corrected and normalized logged PM intensities resulting in multiple statistics which are then summarized into one by the median statistic for each probe-set, in contrast to first summarizing PM intensities and then doing inference. With this being the only difference between PLW and LMW, and one of the differences between logit-t and the ordinary t-test using RMA expression indexes (they also use different background correction and normalization methods), we find that computing statistics for each PM probe and then summarizing shows better performance compared to the other option.

More complicated models often come with the prize of longer computer run times. Of the methods evaluated the BGX model and the PPLR method together with the multi-mgMOS model are the most computer intense ones. The computer run time for one single two group analysis of 3+3 HG-U95A arrays with data from 12626 genes is more than 24 hours with BGX and 1.5 hours for PPLR+multi-mgMOS (using the recommended EM method of PPLR) on a 2.2 GHz AMD Opteron machine. The corresponding time (including pre-processing of PM and MM data) is 2-3 minutes for PLW and 9 seconds for the moderated t-test in LIMMA.

5 Conclusion

We have presented two new methods for ranking genes with respect to differential expression: Probe level Locally moderated Weighted median-t (PLW) and Locally Moderated Weighted-

t (LMW). Both methods perform very well compared to existing methods with PLW having the most accurate ranking of regulated genes in four out of five examined spike-in data sets. With LMW we show that introducing an intensity-level dependent scale parameter for the prior distribution of the gene-specific variances improves the performance of the moderated t-test. Also, compared to the moderated t-statistic, LMW shows a much more stable false discovery rate across intensity-levels when used on MAS5 expression indexes. In the PLW method inference is performed directly on logged PM intensities and the median of the resulting moderated t-statistics for each probe-set is used to find differentially expressed genes. Overall the PLW method performs better than all compared methods and thus probe-level inference appears to be preferable over the standard approach using gene expression indexes for inference.

Acknowledgments

The research was supported by the Gothenburg Mathematical Modeling Center and the Gothenburg Stochastic Centre.

References

Åstrand, M. (2003). Contrast normalization of oligonucleotide arrays. *J. Comput. Biol. 10*(1), 95–102.

Åstrand, M., P. Mostad, and M. Rudemo (2007). Improved covariance matrix estimators for weighted analysis of microarray data. Technical report, Chalmers University of Technology and Göteborg University, Department of Mathematical Statistics, http://www.math.chalmers.se/Math/Research/Preprints/2007/27.pdf.

Baldi, P. and A. D. Long (2001). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics 17*(6), 509–519.

Choe, S., M. Boutros, A. Michelson, G. Church, and M. Halfon (2005). Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biology 6*(2), R16.

Comander, J., S. Natarajan, M. Gimbrone, and G. Garcia-Cardena (2004). Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics 5*(1), 17.

Cope, L. M., R. A. Irizarry, H. A. Jaffee, Z. Wu, and T. P. Speed (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics 20*(3), 323–331.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the *em* algorithm. *J. Roy. Statist. Soc. Ser. B 39*, 1–38.

Eaves, I. A., L. S. Wicker, G. Ghandour, P. A. Lyons, L. B. Peterson, J. A. Todd, and R. J. Glynne (2002). Combining Mouse Congenic Strains and Microarray Gene Expression Analyses to Study a Complex Trait: The NOD Model of Type 1 Diabetes. *Genome Res. 12*(2), 232–243.

Efron, B., R. Tibshirani, J. Storey, and V. Tusher (2001, December). Empirical bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc. 96*, 1151–1160.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning* (First ed.), Volume 1. Springer.

Hein, A.-M. K., S. Richardson, H. C. Causton, G. K. Ambler, and P. J. Green (2005). BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data. *Biostatistics 6*(3), 349–373.

Hu, J. and F. A. Wright (2007). Assessing differential gene expression with small sample sizes in oligonucleotide arrays using a mean-variance model. *Biometrics 63*(1), 41–49.

Irizarry, R., L. Cope, and Z. Wu (2006). Feature-level exploration of a published affymetrix genechip control dataset. *Genome Biology 7*(8), 404.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4*(2), 249–264.

Jain, N., J. Thatte, T. Braciale, K. Ley, M. O'Connell, and J. K. Lee (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics 19*(15), 1945–1951.

Kristiansson, E., A. Sjögren, M. Rudemo, and O. Nerman (2005). Weighted analysis of paired microarray experiments. *Stat. Appl. Genet. Mol. Biol. 4*(1), article 30.

Kristiansson, E., A. Sjögren, M. Rudemo, and O. Nerman (2006). Quality optimised analysis of general paired microarray experiments. *Stat. Appl. Genet. Mol. Biol. 5*(1), article 10.

Lemon, W., S. Liyanarachchi, and M. You (2003). A high performance test of differential gene expression for oligonucleotide arrays. *Genome Biology 4*(10), R67.

Liu, X., M. Milo, N. D. Lawrence, and M. Rattray (2005). A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformatics 21*(18), 3637–3644.

Liu, X., M. Milo, N. D. Lawrence, and M. Rattray (2006). Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics 22*(17), 2107–2113.

Lönnstedt, I. and T. P. Speed (2002). Replicated microarray data. *Statistica Sinica 12*(1), 31–46.

Opgen-Rhein, R. and K. Strimmer (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol. 6*(1), article 9.

Sjögren, A., E. Kristiansson, M. Rudemo, and O. Nerman (2007). Weighted analysis of general microarray experiments. Technical report, Chalmers University of Technology and Göteborg University, Department of Mathematical Statistics, http://www.math.chalmers.se/Math/Research/Preprints/2007/29.pdf.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol. 3*(1), article 3.

Spira, A., J. Beane, V. Pinto-Plata, A. Kadar, G. Liu, V. Shah, B. Celli, and J. S. Brody (2004). Gene Expression Profiling of Human Lung Tissue from Smokers with Severe Emphysema. *Am. J. Respir. Cell Mol. Biol. 31*(6), 601–610.

Stearman, R. S., L. Dwyer-Nield, L. Zerbe, S. A. Blaine, Z. Chan, J. Bunn, Paul A., G. L. Johnson, F. R. Hirsch, D. T. Merrick, W. A. Franklin, A. E. Baron, R. L. Keith, R. A. Nemenoff, A. M. Malkinson, and M. W. Geraci (2005). Analysis of Orthologous Gene Expression between Human Pulmonary Adenocarcinoma and a Carcinogen-Induced Murine Model. *Am J Pathol 167*(6), 1763–1775.

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS 98*(9), 5116–5121.