Statistical Applications in Genetics and Molecular Biology

Volume 10, Issue 1

2011

Article 42

Fully Moderated T-statistic for Small Sample Size Gene Expression Arrays

Lianbo Yu, *The Ohio State University*

Parul Gulati, *The Ohio State University*

Soledad Fernandez, *The Ohio State University*

Michael Pennell, *The Ohio State University*

Lawrence Kirschner, *The Ohio State University*

David Jarjoura, *The Ohio State University*

Recommended Citation:

Yu, Lianbo; Gulati, Parul; Fernandez, Soledad; Pennell, Michael; Kirschner, Lawrence; and Jarjoura, David (2011) "Fully Moderated T-statistic for Small Sample Size Gene Expression Arrays," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 42.

DOI: 10.2202/1544-6115.1701

Available at: <http://www.bepress.com/sagmb/vol10/iss1/art42>

©2011 Berkeley Electronic Press. All rights reserved.

Fully Moderated T-statistic for Small Sample Size Gene Expression Arrays

Lianbo Yu, Parul Gulati, Soledad Fernandez, Michael Pennell, Lawrence Kirschner, and David Jarjoura

Abstract

Gene expression microarray experiments with few replications lead to great variability in estimates of gene variances. Several Bayesian methods have been developed to reduce this variability and to increase power. Thus far, moderated t methods assumed a constant coefficient of variation (CV) for the gene variances. We provide evidence against this assumption, and extend the method by allowing the CV to vary with gene expression. Our CV varying method, which we refer to as the fully moderated t -statistic, was compared to three other methods (ordinary t , and two moderated t predecessors). A simulation study and a familiar spike-in data set were used to assess the performance of the testing methods. The results showed that our CV varying method had higher power than the other three methods, identified a greater number of true positives in spike-in data, fit simulated data under varying assumptions very well, and in a real data set better identified higher expressing genes that were consistent with functional pathways associated with the experiments.

KEYWORDS: empirical Bayes, microarray data analysis, variance smoothing

Author Notes: This research was partially supported by NIH grants P30 CA016058-32 and UL1RR025755. We thank the reviewers for suggested improvements.

1 Introduction

In testing for differential expression with microarrays, several methods have been proposed which borrow information across genes for estimating variances (Long et al., 2001, Tusher et al., 2001, Efron et al., 2001, Lonnstedt and Speed, 2002, Smyth, 2004, Cui et al., 2005, Fox and Dimmic, 2006, Sartor et al., 2006). Most of these references show that smoothing methods increase power relative to the ordinary t-test. In addition, use of the ordinary t-test means that variance estimates are gene specific, and the distribution of gene specific variance estimates across many thousands of genes will have some very low values due to sampling error. With small sample sizes, many gene specific variance estimates can be close to zero, leading to very small p-values associated with small and uninteresting differential expression values.

Smyth (2004) generalized the empirical Bayes model of Lonnstedt and Speed (2002) into an approach that can be used for small sample sizes. The method borrows information across genes to smooth variances and uses a single posterior variance estimate in a t-test. It performs well by increasing power, relative to the ordinary t-test, but assumes no dependence between gene variances and their corresponding expression values. Microarray data often show a relationship between gene expression and gene variance and this relationship has been considered in other Bayesian methods (Long et al., 2001, Fox and Dimmic, 2006, Sartor et al., 2006). Long et al.'s (2001) CyberT method uses a t-statistic with moving average variance estimates. However, the method requires an a priori value for the prior degrees of freedom rather than empirical estimates, and a chosen window size for smoothing the prior variances. Fox and Dimmic (2006) proposed an improvement to the CyberT. Their method uses a similar moving average approach to estimate the variances, but the value of prior degrees of freedom is not assigned a priori. Like the CyberT method, Fox and Dimmic's method requires users to fix the window size for variance smoothing.

Sartor et al. (2006) proposed the intensity-based moderated t-statistic (IBMT) as an extension of Smyth's moderated t-statistic (SMT). It uses local regression to model a functional relationship between gene variances and expression levels. IBMT provides better posterior variance estimates than the SMT method, because it shrinks sample variances towards expression specific values. As a result, IBMT reduces the variance of gene variances compared to SMT, which increases the prior degrees of freedom for the t-statistic, thereby increasing power. This is especially noticeable with small sample sizes. Although IBMT allows variances to change across gene expression levels, it fixes the prior degrees of freedom to one value. This means that the coefficient of variation (CV) of gene variances is assumed to be constant across genes (See Appendix). This assumption of constant CV for the

gene variances is too restrictive in microarray experiments as we show with examples. In particular, filtering methods at low expression values and ceiling effects at high ones can lead to more or less homogeneous variances.

We propose a t-test that more fully moderates the denominator of the test statistic (fully moderated t-statistic or FMT) than IBMT or SMT. Using local linear regression, we allow both gene variances and the CV of gene variances to change as a function of gene expression resulting in improved estimates of the prior degrees of freedom as well as gene variances. Consequently, the FMT method should have more accurate degrees of freedom and false positive and negative control across different expression levels. The proposed method is applied to simulated and spike-in data and its performance is compared with three other t-test methods. We also analyzed a small sample size real data set to show the differences in the gene lists produced by these methods.

2 Hierarchical Model

Our FMT method is adapted from the hierarchical Bayesian model proposed by Smyth (2004). In this model, assuming a linear model for log transformed expression data, the residual variance for gene g , s_g^2 , is assumed to follow a scaled chi-square distribution with d_g degrees of freedom:

$$s_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2,$$

where σ_g^2 is the gene variance, which is assigned a scaled inverse chi-square prior:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_{0g} s_{0g}^2} \chi_{d_{0g}}^2, \quad (1)$$

and where d_{0g} and s_{0g}^2 are the prior degrees of freedom and location, respectively. Under this model, the posterior mean of σ_g^2 given s_g^2 , is

$$\tilde{s}_g^2 = \frac{d_{0g} s_{0g}^2 + d_g s_g^2}{d_{0g} + d_g}$$

which is incorporated into the denominator of our fully moderated t-statistic:

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{2/n}}$$

where $\hat{\beta}_g$ is the difference in average log-expression values for gene g under the two groups of interest, $\tilde{s}_g^2$ is the posterior mean of the residual variance and n is

the number of samples within each group. Under the null hypothesis, the moderated t-statistic follows a t-distribution with augmented degrees of freedom $d_g + d_{0g}$, where $d_g = 2n - 2$ and assumed values of d_{0g} and s_{0g}^2 . Below we describe methods of estimating these parameters as a function of expression levels. We present the moderated t for the 2 group case here, for simplicity of exposition. It is generalizable to handle any linear model contrast(s), i.e., β_g can be any contrast across multiple groups including models with covariates. Details are provided in Smyth (2004), where our $2/n$ in the denominator gets replaced by his v_{gj} .

For our FMT method, we allow both d_{0g} and s_{0g}^2 to vary across genes. Note that the SMT method assumes they are constant. The consequence of assuming s_{0g}^2 constant under SMT is that all variances are shrunk toward one value. When s_{0g}^2 varies with expression level, SMT not only biases estimates of s_{0g}^2 , but also inflates the variance of σ_g^2 . Inflated estimates of $\text{Var}(\sigma_g^2)$ produce underestimates of the prior degrees of freedom, which reduces the power of the test statistic. Sartor et al.'s IBMT method (2006) extends SMT by allowing s_{0g}^2 to vary with expression levels. However, like SMT, IBMT assumes that the prior degrees of freedom are constant. The consequence of this assumption is that prior degrees of freedom may be under or overestimated at different expression levels resulting in reduced power or increased false positive error at different levels. In addition, estimates of s_{0g}^2 require prior estimates of d_{0g} , which means that s_{0g}^2 estimates could be biased by assuming constant d_{0g} . As seen below, our FMT method allows both the prior variance and prior degrees of freedom to vary flexibly with expression levels, thus removing the restrictions imposed by these previous methods. We have frequently found that prior degrees of freedom are higher at higher expression levels and lower at low expression levels.

3 Estimation of Hyperparameters

The hyperparameters d_{0g} and s_{0g}^2 are estimated from the data using an empirical Bayes approach similar to that implemented in SMT and IBMT (Smyth, 2004, Sartor et al., 2006). Under our model with the addition of a varying d_{0g} , the marginal distribution of the sample variance s_g^2 is a scaled F-distribution (Smyth, 2004):

$$s_g^2 \sim s_{0g}^2 \cdot F_{d_g, d_{0g}}$$

The log of the sample variance, $\log(s_g^2)$, is distributed as a constant plus Fisher's Z distribution (Johnson and Kotz, 1970). The first two moments of $\log(s_g^2)$ are

$$E(\log s_g^2) = \log s_{0g}^2 + \Psi(d_g/2) - \Psi(d_{0g}/2) + \log(d_{0g}/d_g), \quad (2)$$

and

$$\text{Var}(\log s_g^2) = \Psi'(d_g/2) + \Psi'(d_{0g}/2), \quad (3)$$

where $\Psi()$, the digamma function, is the first derivative of the log gamma function and $\Psi'()$, the trigamma function, is the second derivative of the log gamma function. Define e_g as an adjusted log-variance of the g th gene:

$$e_g = \log s_g^2 - \Psi(d_g/2) + \log(d_g/2),$$

which (based on Equation 2) has expected value

$$E(e_g) = \log s_{0g}^2 - \Psi(d_{0g}/2) + \log(d_{0g}/2), \quad (4)$$

and variance

$$\text{Var}(e_g) = \Psi'(d_g/2) + \Psi'(d_{0g}/2).$$

The adjusted log variances, e_g , are modeled as a function of the mean log-expression values using LOESS local regression (Cleveland, 1979, Cleveland and Devlin, 1988) to obtain the smoothed adjusted log variance values, $\text{pred}(e_g)$. In our approach, we also model d_{0g} as a function of gene expression. We first calculate the mean squared residuals of e_g by a moving average with a small window size along rank ordered gene expression mean values:

$$S^2(\log s_g^2) = \text{moving average}[e_g - \text{pred}(e_g)]^2. \quad (5)$$

We replace $\text{Var}(\log s_g^2)$ of Equation 3 by $S^2(\log s_g^2)$ and solve for $\Psi'(d_{0g}/2)$:

$$\Psi'(\widehat{d_{0g}}/2) = S^2(\log s_g^2) - \Psi'(d_g/2). \quad (6)$$

For the moving average, we chose a window size of 40 genes, but this does not seem to be a critical choice as the same results are obtained for a wide range of window sizes. We present this result in our simulations section. The estimated values $\Psi'(\widehat{d_{0g}}/2)$, obtained from Equation 6, are then modeled as a function of the mean log-expression values using LOESS local regression, and the smoothed $\Psi'(d_{0g}/2)$ values provide $\text{pred}(\Psi'(d_{0g}/2))$. We found that LOESS span parameters in the interval [0.8, 1.0] gave similar results and easily captured the true functions of simulation models. This was also obvious visually. When lower spans were

used, smoothing seemed insufficient. Finally, we solve for d_{0g} using the smoothed predicted values

$$\widehat{d_{0g}} = 2[\Psi'^{-1}(\text{pred}(\Psi'(d_{0g}/2)))],$$

where Ψ'^{-1} is the inverse of the trigamma function. The gene-specific s_{0g}^2 are obtained by substituting the smoothed log variance values, $\text{pred}(e_g)$, and the estimated prior degrees of freedom, $\widehat{d_{0g}}$, in Equation 4 and solving for s_{0g}^2 :

$$\widehat{s_{0g}^2} = \exp\{\text{pred}(e_g) + \Psi(\widehat{d_{0g}}/2) - \log(\widehat{d_{0g}}/2)\}.$$

Note that the moving average in Equation 5 uses multiple deviances between e_g and $\text{pred}(e_g)$, which proved valuable in producing symmetry for LOESS smoothing. This is unnecessary for obtaining $\widehat{s_{0g}^2}$ from $\text{pred}(e_g)$, because the s_g^2 are already based on multiple deviances between the original sample observations and their mean.

4 Simulations and Examples

4.1 Simulated data

Expression levels of 20,000 genes were simulated under three different model assumptions (see Simulation Steps in Appendix). Forty percent were simulated as non-expressed genes. Of the remaining 60% expressed genes, 300 or 2.5% were simulated as differentially expressed. Simulation Model 1 was based on our FMT method's model assumptions and allowed the hyperparameters, prior location and prior degrees of freedom of σ_g^2 , to vary across the genes as a function of their expression levels. We also studied the performance of the FMT, when the assumptions of the other methods (IBMT, SMT) were met. Simulation Model 2 was based on the SMT model assumptions where the prior degrees of freedom and prior location of σ_g^2 were constant for all the genes. Simulation Model 3 was based on the IBMT model assumptions where the prior degrees of freedom of σ_g^2 were constant for all the genes, but the prior location changed with gene expression level. All the simulations were set to resemble real data from prior gene expression array Affymetrix experiments. Under each of the three models, we generated data sets for three different sample sizes (3, 5, and 10 samples in each of 2 groups). A total of 100 simulations were run for each model and sample size combination. We found 100 simulations were sufficient because the empirical coefficient of variation of the estimators we report here were less than 5%. A threshold filtering method

was applied to each simulated data set to filter out genes which were simulated to be non-expressed. This method compares the signal from each gene in each sample to a cut-off value (close to background level), and filters out genes if that gene has signal levels smaller than the cut-off in all or a very high percentage of the samples. Consideration of filtering is critical when smoothing variances. Without filtering, variances and their CVs would be smoothed toward the values of genes whose expression levels are close to background noise which will downwardly bias variance estimates of the genes that are actually expressed.

Figure 1: Relationship between hyperparameters and gene expression for Simulation Model 1. A) Prior variance (s_{0g}^2). B) Prior degrees of freedom (d_{0g}).

The FMT simulation model (Model 1) and the IBMT simulation model (Model 3) simulation steps differ from those of the SMT model (Model 2) only in the assumption about the two hyperparameters, prior degrees of freedom and prior location. Under the IBMT model, we induced a dependency structure between s_{0g}^2 and the average log expression values (Figure 1A), but kept the prior degrees of freedom constant ($d_0 = 4$) for all the genes. Under the FMT model, both s_{0g}^2 and d_{0g} are a function of average log expression (Figure 1). The range of prior location and prior degrees of freedom chosen for Model 1 are based on our experience with real data examples. In each of the figures presented here, we used a sample of $n = 3$ per group.

False Positive Control

We tested for differential expression in each simulated data set controlling the per family error rate (PFER). PFER has been shown to have superior stability in terms

of the variance of both the number of true positives and the number of total positives when compared to Benjamini and Hochberg's (BH) procedure (Gordon et al., 2007). This method focuses on the expected number of false positives rather than on the probability of false positives (Tukey, 1953). PFER is defined as:

$$\text{PFER} = E(V) = E \sum_{g=1}^{G_0} I_{\{P_g \leq \frac{\gamma}{G}\}} = \sum_{g=1}^{G_0} \Pr\{P_g \leq \frac{\gamma}{G}\} = G_0 \frac{\gamma}{G} \leq \gamma,$$

where $E(V)$ is the expected number of false positives and G_0 is the total number of true null hypotheses, G is the total number of hypotheses tested, P_g ($g = 1, 2, \dots, G$) are p-values (the random variables here), γ is the significance cut-off value, and $I_{\{E\}}$ denotes an indicator function of event E . The above equality holds under just one weak assumption that the mean of the p-values across all the null genes is less than or equal to the cut-off p-value, $\frac{\gamma}{G}$.

The mean number of false positives for the SMT, IBMT and FMT testing methods were slightly less than or equal to the expected values of V . The same false positive control was seen for simulation Models 2 and 3. These results are exactly paralleled by the more commonly used false discovery rate (FDR) method (Benjamini and Hochberg, 1995, Gordon et al., 2007). SMT, IBMT, and FMT methods under all 3 models exhibited the same levels of FDR control.

Power and prior degrees of freedom

Power under Simulation Models 1 and 2 are shown in Figures 2. Under the FMT simulation model (Model 1), the FMT testing method performs the best with respect to power, followed by IBMT, SMT and the ordinary t-test, respectively (Figure 2A). It is not surprising that the FMT performed the best as it accurately estimated the prior degrees of freedom (Figure 3) across the gene expression range while the IBMT and SMT methods underestimated the prior degrees of freedom for genes whose average log2 expression levels were greater than 10. The estimated prior degrees of freedom were averaged over 100 simulations, and per gene averages were displayed for the FMT method and compared to the true curve. As expected, both IBMT and FMT accurately estimated the prior location of the variance while SMT underestimated its value at low expression levels (data not shown). Under the SMT simulation model (Model 2), all three moderated t-statistic methods displayed similar power, and higher than that obtained from the ordinary t-test (Figure 2B). This demonstrates the flexibility of IBMT and FMT methods; because they produce the same results as the SMT method when its stronger assumptions are met. So that the only cost to use FMT or IBMT over SMT is the additional calculations.

Because IBMT and SMT underestimated the degrees of freedom at the larger expression levels under Model 1, these methods should have lower power

than FMT to detect true positives among highly expressed genes. To confirm this, we grouped genes into three expression categories: low (1: average signal < 7), moderate (2: $7 < \text{average signal} < 11$) and high (3: average signal > 11). With the expression level categories of the 300 true positives were fixed over the 100 simulations, the average number of true positives detected at each expression level category were calculated. Figure 4 gives these results when the expected number of false positives was controlled at 5. All three moderated t-tests outperformed the ordinary t-test at each intensity level. At low intensity levels, the three moderated t-test methods were similar. At moderate intensity levels, the FMT and IBMT methods outperformed the SMT, because they correctly estimated the hyperparameters there while SMT underestimated them (Figure 3). At high intensity levels, FMT outperformed the other two moderated t-tests, because it correctly estimated prior degrees of freedom there, while IBMT and SMT underestimated them.

Figure 5 displays the number of false positives detected within each expression level category when the expected number of false positives was controlled at 5. FMT and IBMT provided very similar false positive control. SMT had most of its false positives observed at low intensity levels; this is because the method underestimates the variance at these expression levels (data not shown).

Gene rank and false positive numbers

For each testing method, the genes were ranked by their p-values (lowest to highest) and the numbers of false positives among the most significant 300 genes were calculated. The results for Model 1 are presented in Figure 6. FMT accumulated the fewest false positives followed by IBMT, SMT and t-test, respectively. Under the SMT simulation model (Model 2), all three moderated t-statistic methods displayed similar false positives.

Power and sample size

When 5 samples per group were simulated under the FMT model, the FMT method still outperformed the other three testing methods, but the differences were not as large as seen with a sample size of 3 per group. As the sample size increases, the total degrees of freedom also increase and hence modest underestimation or overestimation of the degrees of freedom does not have as large effect on power (data not shown).

A)

B)

Figure 2: Power plots of four testing methods under 2 different simulation models. Power averaged over 100 simulated datasets was calculated separately for the simulated data by using 4 different testing methods: t-test (purple), SMT (blue), IBMT (red), and FMT (black). A) FMT simulation model (Model 1). B) SMT simulation model (Model 2).

Figure 3: Prior degrees of freedom estimated by different moderated testing methods under the FMT simulation model (Model 1). Prior degrees of freedom were estimated using: SMT (blue), IBMT (red), and FMT (green) under the FMT simulation model (Model 1) and averaged over 100 simulated datasets.

Moving average and window size

To assess the sampling variability of LOESS smoothing curves of the moving average estimate $\Psi'(d_{0g}/2)$ (Equation 6) and select appropriate moving average window sizes, we calculated the mean and standard deviation of LOESS estimates by intensity levels over 100 simulations (Figure 7). The average CV across the 100 simulations of LOESS estimates is less than 5% for all intensity levels and all selected window sizes ($m = 10, \dots, 2000$). The window size of 40 genes seems the best choice for these data, but any moderate window size gave the same results.

Gene correlation

To understand the effect of gene association on the tests' performances, we induced a block diagonal correlation structure among genes under Model 1. We randomly divided the 12,000 expressed genes into 120 blocks of 100 genes each. Within each block we fixed the correlation to 0.5. The power and false positive number obtained from these data are highly similar to those obtained under data simulated assuming independence (Figures 4 and 5).

Figure 4: Observed true positives detected at different expression levels. Observed true positives averaged over 100 simulated datasets were plotted under 4 different testing methods: t-test (purple), SMT (blue), IBMT (red), and FMT (black) at three expression levels (low (1), medium (2), and high (3)) when PFER equals 5 under Model 1. (Numbers in parentheses in the legend indicate the number of true positives).

Violation of shape of variance distribution

Another set of simulations were performed to determine the sensitivity of the moderated t tests to the inverse chi-square assumption about the residual variance. The residual variance, σ_g^2 , for each of the 12,000 expressed genes was randomly generated from either a scaled inverse chi-square distribution or a normal distribution with distribution selection probability 0.7 and 0.3, respectively. The mean of the normal distribution was fixed at the 80th percentile of the scaled-inverse chi-square distribution and the standard deviation was fixed at 0.3. The resulting distribution is bi-modal and has a much heavier tail than the inverse scaled chi-square distribution assumed by the moderated t methods. Under this mixture model, false positives were controlled at their nominal level. Because a constant variance and prior degrees of freedom were used, no differences among the moderated t methods were observed (nor expected).

Figure 5: Observed false positives detected at different expression levels. Observed false positives averaged over 100 simulated datasets were counted by using 4 different testing methods: t-test (purple), SMT (blue), IBMT (red), and FMT (black) at three expression levels (low (1), medium (2), and high (3)) when PFER equals 5 under Model 1. (Numbers in parentheses in the legend indicate the numbers of true negatives).

4.2 Affymetrix spike-in data

The performance of the four testing methods (SMT, IBMT, ordinary t-test and FMT) were next compared using a controlled "spike-in" Affymetrix data set. In this data set, the numbers of true negatives and true positives are known. The "spike-in" data were generated by Choe et al. (2005) and mimic a two group (treatment vs. control) microarray experiment. The data set consists of 3,860 individual cRNAs of known sequence which were divided into two samples: 'constant' (C) and 'Spike' (S). Each sample was hybridized in triplicate on Affymetrix GeneChips (six chips total). The data set also includes expression levels of 10,131 empty probe sets (blanks) in each of the six chips. Out of 3,860 cRNAs, 1,309 true positives were spiked-in with different known concentrations in the S and C samples. The remaining 2,551 cRNAs had the same concentrations in each sample (true nulls). Choe et al. (2005) generated 152 expression data sets by using different combinations of

Figure 6: Comparison of false positives among top 300 ranked genes. Genes were ranked by p-values, and the corresponding false positive numbers averaged over 100 simulated datasets were obtained separately for four different testing methods: t-test (purple), SMT (blue), IBMT (red), and FMT (black) when PFER equals 5 under Model 1.

preprocessing methods and ranked the preprocessed data sets based on their ability to detect differentially expressed genes. As in Choe et al. (2005) and Sartor et al. (2006), we focused on the average of the 10 data sets with the highest ranking (best) preprocessing steps.

Figure 8 displays prior degrees of freedom estimates of the three moderated t-statistics (SMT, IBMT, and FMT). Because SMT and IBMT assume constant prior degrees of freedom, d_0 , they estimated a single d_0 value (SMT=1.19, IBMT=2.30) for all the genes. Our method allowed the prior degrees of freedom to vary across the genes and they ranged from 1.19 to 19.74.

We compared the four testing methods with respect to their power and number of false positives among the most significant 1,309 genes. When genes were ranked by p-values, our FMT method had the minimum number of false positives among the 1,309 genes with smallest p-values followed by IBMT, the t-test and SMT (Figure 9). This relationship was most apparent for genes 800 – 1,309. As noted by others (Storey and Dabney, 2006, Irizarry et al., 2006), type I error is not adequately controlled in these data sets because of experimental dependencies.

Figure 7: Summary plots of LOESS smoothing curves for estimating prior degrees of freedom. A) Mean of LOESS curves over 100 simulations for the window sizes of $m = 10, 40, 200, 600$, and 2000 genes. Black curve represents the true model. B) Standard deviation of LOESS curves over 100 simulations for the window sizes of $m = 10, 40, 200, 600$, and 2000 genes.

Figure 8: Spike-in data prior degrees of freedom estimates. SMT (blue), IBMT (red), and FMT (black) methods were used to estimate prior degrees of freedom over average log expressions.

Figure 9: Comparison of false positives among top ranked genes. Gene ranks based on p-values were obtained separately for 4 different testing methods: t-test (purple), SMT (blue), IBMT (red), and FMT (black). False positives counts were determined from among the top ranked genes.

4.3 Real data example

To demonstrate the advantage of our proposed method for identifying differentially expressed genes, we applied it to a real microarray dataset and compared it with the other three methods as in the simulation studies and analysis of the Spike-in data. Affymetrix mouse exon arrays were used for measuring gene expression from the thyroids of 4 control mice and 4 littermates carrying a thyroid-specific deletion of the *Prkar1a* gene (Pringle et al., 2011). *PRKAR1A/Prkar1a* encodes the type 1A regulatory subunit of the cAMP-dependent Protein Kinase (Protein Kinase A, PKA), and mutation of this gene is tumorigenic in both mice and men (Kirschner et al., 2009). Thus, this microarray compares normal to neoplastic thyroid tissue.

Figure 10 shows the fitted LOESS local regression curves of variance estimates of log gene variance against log expression levels for SMT, IBMT, and FMT. The estimated prior degrees of freedom for FMT increased from 4.34 to 8.30 across the log expression level range of 3.33 to 13.66. In addition, Figure 11 displays the fitted LOESS local regression curves of prior variance estimates against log expression levels for SMT, IBMT, and FMT. For testing differential expression, per family error rate was 10 (Gordon et al., 2007). We found 536 significant genes under FMT, 524 significant genes under IBMT, 533 significant genes under SMT, and 244 significant genes under the ordinary t-test. Figure 12 displays overlaps and differences of the 4 significant lists. FMT detected 18 unique genes that are not detected by IBMT. Note that 17 out of these 18 genes had moderate or high expression. IBMT detected 6 unique genes that were not detected by FMT and all of them were low expression genes close to background noise. The main distinction between these two methods is that IBMT estimated a fixed prior degrees of freedom of 5.40, which was higher than the prior degrees of freedom estimated by FMT at low expression, but was smaller than FMT's estimates at the moderate and high expression levels thereby explaining the increase in positives observed for FMT at these expression levels. Similar differences were seen between FMT and SMT. Even though SMT has the second largest significant gene list, 25 unique SMT genes that are not detected by FMT are all low expression genes close to background noise.

We studied the differences in gene functions of the FMT and IBMT significant gene lists through Ingenuity Pathway Analysis (IPA), which uses prior published data on gene function to identify associations. The main gene function networks for both IBMT and FMT gene lists are identical. In addition, we found that most of the function networks of the 18 genes uniquely identified by FMT are consistent with the function networks identified in the overall FMT and IBMT significance lists. In other words the IPA validated that these 18 genes should be considered positives. However most of the function networks of the 6 genes unique to IBMT are not consistent with those same function networks, which suggests that

these 6 genes are likely false positives. We had similar findings for the 25 genes unique to SMT. Details of this analysis are to be published. An important difference in the FMT gene list is that it included Pde7b, a high-affinity cAMP-specific phosphodiesterase anticipated to be upregulated in response to elevated PKA activity (Hetman et al., 2000), and Sphingosine kinase 1, which is functionally upregulated in response to PKA activation (Rius et al., 1997).

Figure 10: Estimated prior degrees of freedom and variance of log variances against log intensity in the real data example. A) Estimated prior degrees of freedom by FMT (green), IBMT (red), and SMT (blue). B) Estimated variance of log variances by FMT. Moving average (a window size of 40 genes) estimate (green) was obtained by LOESS local regression with span=0.95.

5 Discussion and Conclusions

We proposed a novel improvement of the moderated t-test for identifying differentially expressed genes in small sample size gene expression experiments. Previous moderated t-test methods, SMT and IBMT, are too restrictive in that they assume a constant coefficient of variation of gene variances. In contrast, the FMT method we propose allowed the CVs of gene variances to change as a function of gene expression levels.

When data were simulated under the SMT model, the three methods were comparable in terms of power, false positive control, and accuracy of hyperparameter estimates. However, when SMT assumptions were not met, differences among the methods were evident. This was expected, because SMT ignores the dependency of gene variances and the CV of gene variances on gene expression levels

Figure 11: Estimated prior variance against log intensity in the real data example. Prior variance estimates by FMT (green) and IBMT (red) were obtained through fitting a LOESS local regression on the adjusted log-variance e_g with span=0.75. Differences between FMT and IBMT are mainly due to differences in estimated prior degrees of freedom. Prior variance estimate for SMT (blue) is constant.

and results in incorrect estimation of these hyperparameters. Specifically, we found that SMT consistently underestimates the prior degrees of freedom in these situations, which is especially important in small sample size experiments. We also found that when the data were generated under the IBMT and FMT models where variances are larger at the lower expression levels, SMT generates most of its false positives at lower expression levels. While the performance of IBMT and FMT were comparable when data were simulated under the IBMT model, FMT was more powerful than IBMT when the CVs of gene variances were dependent on expression levels. Under an increasing trend in prior degrees of freedom, the differences in power were greatest at high intensity levels. The power of the moderated t methods were more comparable at larger sample sizes, because as the residual degrees of freedom increase, the posterior mean of the gene variance is pulled toward the empirical gene-specific estimate. Also, when the residual degrees of freedom are larger, the increase in degrees of freedom provided by the moderated t-tests is relatively smaller. We also compared the four testing methods using an Affymetrix controlled spike-in data set. Our method identified the largest number of true positives, followed by IBMT, t-test and SMT, respectively. This is due to the higher degrees of freedom associated with our method at moderate to high expression levels, where we gained most in power. These results were confirmed in an analysis of

Figure 12: Venn Diagram of the 4 significant gene lists from 4 different testing methods:t-test, SMT, IBMT, and FMT

a small sample size experiment. In this analysis, FMT identified 18 unique genes at moderate to high expression levels. IPA analysis revealed that the function of these 18 genes were consistent with the other significant genes. An R-package to implement this method is in preparation, and will be called FMT. A temporary R function can be found on <http://www.biostatistics.osu.edu> under the first author's page.

6 Appendix

6.1 Relationship between the prior degrees of freedom and the coefficient of variation

With σ^2 following a scaled inverse chi-square distribution, the probability density function over the domain $x > 0$ is

$$f(x; d_0, s_0^2) = \frac{(s_0^2 d_0 / 2)^{d_0 / 2} \exp\left[-\frac{d_0 s_0^2}{2x}\right]}{\Gamma(d_0 / 2) x^{1 + d_0 / 2}}$$

where d_0 is the degrees of freedom parameter and s_0^2 is the scale parameter. The mean is $\frac{d_0 s_0^2}{d_0 - 2}$ for $d_0 > 2$ and the variance is $\frac{2d_0^2 s_0^4}{(d_0 - 2)^2 (d_0 - 4)}$ for $d_0 > 4$. Therefore the coefficient of variation (CV) is simply

$$CV = \sqrt{\frac{2}{d_0 - 4}}.$$

6.2 Simulation steps

- 1: For 20,000 genes ($g = 1, 2, \dots, 20,000$) simulate the average log expression, α_g , from a three parameter log-normal distribution: $\ln(\alpha_g - 1.75) \sim N(1.35, 0.5)$.
- 2: **FMT Model (Model 1)**: For the 12,000 expressed genes; calculate the gene-specific prior degrees of freedom d_{0g} and gene-specific prior location of σ_g^2 , s_{0g}^2 , as a function of the average log expression values (α_g) obtained from Step 1. We modeled the prior location and prior degrees of freedom as:

$$s_{0g}^2 = 0.5e^{-1.9(\alpha_g - 5.25)} + 0.1$$

$$d_{0g} = 1.2\sqrt{-1.1e^{0.1(\alpha_g - 6)} + 0.6(\alpha_g - 7)^2 + 20} + e^{0.95(\alpha_g - 14)} - 2$$

SMT MODEL (Model 2): For the 12,000 expressed genes, we fixed the prior degrees of freedom at $d_0 = 4$ and the prior location of the gene variance at $s_0^2 = 0.24$.

IBMT Model (Model 3): For the 12000 expressed genes, we fixed the prior degrees of freedom at $d_0 = 4$ and modeled the gene-specific prior location of σ_g^2 as a function of the average log expression values, α_g , obtained from Step 1 using the same expression as simulation Model 1.

- 3: For the 12,000 expressed genes, we simulated σ_g^2 from an inverse chi square distribution (Equation 1) using the hyperparameters from Step 2.
- 4: For the 8,000 non-expressed genes, we randomly selected σ_g^2 from the genes sample variances from real data with noise level expression values.
- 5: Among the 12,000 expressed genes, 300 genes were simulated to be differentially expressed using the following steps:
 - a) We simulated the mean log-differences μ_g from $N(k, \eta_g^2)$, where $k = 1.5$ for 150 differentially expressed genes and $k = -1.5$ for the remaining 150 differentially expressed genes and $\eta_g^2 = 3\sigma_g^2$.
 - b) We simulated the raw data for each sample in group 1 from $N(\alpha_g + \mu_g/2, \sigma_g^2)$ and for each sample in group 2 from $N(\alpha_g - \mu_g/2, \sigma_g^2)$.
- 6: For the remaining 11,700 expressed genes and 8,000 non-expressed genes, we simulated the raw data for each sample from $N(\alpha_g, \sigma_g^2)$.

References

- Benjamini, Y. and Y. Hochberg (1995): “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *J. Roy. Statist. Soc. Ser. B*, 57, 289–300.
- Choe, S. E., M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon (2005): “Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset,” *Genome Biology*, 6(2), R16.
- Cleveland, W. S. (1979): “Robust locally weighted regression and smoothing scatterplots,” *Journal of the American Statistical Association*, 74, 829–836.
- Cleveland, W. S. and S. J. Devlin (1988): “Locally-weighted regression: An approach to regression analysis by local fitting,” *Journal of the American Statistical Association*, 83, 596–610.
- Cui, X., J. T. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill (2005): “Improved statistical tests for differential gene expression by shrinking variance components estimates,” *Biostatistics*, 6, 59–75.
- Efron, B., R. Tibshirani, J. Storey, and V. Tusher (2001): “Empirical bayes analysis of a microarray experiment,” *Journal of the American Statistical Association*, 96, 1151–1160.
- Fox, R. J. and M. W. Dimmic (2006): “A two-sample bayesian t-test for microarray data,” *BMC Bioinformatics*, 7, 126.
- Gordon, A., G. Glazko, X. Qiu, and A. Yakovlev (2007): “Control of the mean number of false discoveries, bonferroni and stability of multiple testing,” *Annals of Applied Statistics*, 1, 179–190.
- Hetman, J. M., S. H. Soderling, N. A. Glavas, and J. A. Beavo (2000): “Cloning and characterization of pde7b, a camp-specific phosphodiesterase,” *Proc. Natl. Acad. Sci. U.S.A.*, 97, 472–476.
- Irizarry, R., L. Cope, and Z. Wu (2006): “Feature-level exploration of a published affymetrix genechip control dataset,” *Genome Biology*, 7, 404.
- Johnson, N. L. and S. Kotz (1970): *Distributions in statistics: continuous univariate distributions*, New York: Wiley.
- Kirschner, L. S., Z. Yin, G. N. Jones, and E. Mahoney (2009): “Mouse models of altered protein kinase a signaling,” *Endocr Relat Cancer*, 16, 773–793.
- Long, A. D., H. J. Mangalam, B. Y. Chan, L. Toller, G. W. Hatfield, and P. Baldi (2001): “Improved statistical inference from dna microarray data using analysis of variance and a bayesian statistical framework. analysis of global gene expression in escherichia coli k12,” *J. Biol. Chem.*, 276, 19937–19944.
- Lonnstedt, I. and T. P. Speed (2002): “Replicated microarray data,” *Statistica Sinica*, 12, 31–46.

- Pringle, D. R., Z. Yin, G. N. Jones, P. K. Manchanda, D. F. Comiskey, L. Yu, D. Jarjoura, K. M. D. la Perle, and L. S. Kirschner (2011): "Thyroid specific ablation of the carney complex gene, *prkar1a*, results in hyperthyroidism and follicular thyroid cancer," *In Preparation*.
- Rius, R. A., L. C. Edsall, and S. Spiegel (1997): "Activation of sphingosine kinase in pheochromocytoma pc12 neuronal cells in response to trophic factors," *FEBS Letters*, 417, 173–176.
- Sartor, A. M., R. C. Tomlinson, C. S. Wesselkamper, S. Sivaganesan, D. G. Leikauf, and M. Medvedovic (2006): "Intensity-based hierarchical bayes method improves testing for differentially expressed genes in microarray experiments," *BMC Bioinformatics*, 7, 358.
- Smyth, G. K. (2004): "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, 3, Article 3.
- Storey, J. and A. Dabney (2006): "A reanalysis of a published affymetrix genechip control dataset," *Genome Biology*, 7, 401.
- Tukey, J. W. (1953): *The problem of multiple comparison. In: The collected works of John W. Tukey VIII. Multiple comparisons: 1948-1983 1-300*, New York: Chapman and Hall.
- Tusher, V. G., R. Tibshirani, and G. Chu (2001): "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Natl. Acad. Sci. U.S.A.*, 98, 5116–5121.