OPEN ACCESS Freely available online

PLoS one

A Nonparametric Mean-Variance Smoothing Method to Assess *Arabidopsis* Cold Stress Transcriptional Regulator *CBF2* Overexpression Microarray Data

Pingsha Hu[1]*, Tapabrata Maiti[2]

1 Department of Energy-Plant Research Laboratory, Michigan State University, East Lansing, Michigan, United States of America, **2** Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, United States of America

Abstract

Microarray is a powerful tool for genome-wide gene expression analysis. In microarray expression data, often mean and variance have certain relationships. We present a non-parametric mean-variance smoothing method (NPMVS) to analyze differentially expressed genes. In this method, a nonlinear smoothing curve is fitted to estimate the relationship between mean and variance. Inference is then made upon shrinkage estimation of posterior means assuming variances are known. Different methods have been applied to simulated datasets, in which a variety of mean and variance relationships were imposed. The simulation study showed that NPMVS outperformed the other two popular shrinkage estimation methods in some mean-variance relationships; and NPMVS was competitive with the two methods in other relationships. A real biological dataset, in which a cold stress transcription factor gene, *CBF2*, was overexpressed, has also been analyzed with the three methods. Gene ontology and cis-element analysis showed that NPMVS identified more cold and stress responsive genes than the other two methods did. The good performance of NPMVS is mainly due to its shrinkage estimation for both means and variances. In addition, NPMVS exploits a non-parametric regression between mean and variance, instead of assuming a specific parametric relationship between mean and variance. The source code written in R is available from the authors on request.

Citation: Hu P, Maiti T (2011) A Nonparametric Mean-Variance Smoothing Method to Assess *Arabidopsis* Cold Stress Transcriptional Regulator *CBF2* Overexpression Microarray Data. PLoS ONE 6(5): e19640. doi:10.1371/journal.pone.0019640

Editor: Miguel A. Blazquez, Instituto de Biología Molecular y Celular de Plantas, Spain

Received December 8, 2010; **Accepted** April 4, 2011; **Published** May 17, 2011

Funding: This work was supported by the National Science Foundation (NSF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: phu@msu.edu

Introduction

Microarray has become a powerful tool for biological and medical science to monitor transcriptome changes under different treatments. However, because of high price of microarray experiments, replicates for each experiment are restricted in most cases. The feature of small replicates and large gene numbers, e.g., about 6,000 in yeast and 23,000 in Arabidopsis, in microarray data usually results in poor estimation of gene-specific variances. Several methods have been suggested for modification of gene specific variances or covariances to improve the estimation. For example, Efron et al. [1] suggested modifying the denominator of the *t*-statistic to allow estimation less sensitive to gene-specific variances. Smyth [2] proposed smoothing gene-specific variances to a common value. Cui et al. [3] and Tong and Wang [4] developed shrinkage estimators for gene specific variances using Stein-type estimation under squared error loss function which were used to construct traditional *t*- type and *F*- type statistics. In all the above estimators, gene specific means were assumed to be independent of variances. It has been observed that means are related to variances in microarray experiments; usually genes with high expression level show high variances, while genes with low expression level display small variances (Figure 1).

Recently, Hu and Wright [5] suggested a linear model to estimate gene-specific variances based on means. However, the relationship between mean and variance is not always linear. Figure 1 shows real biological datasets from *Arabidopsis thaliana* in Gene Expression Omnibus database (GSE5566, GSE9955, and GSE5520 for dataset 1, 2, and 3, respectively) and clearly suggests a non-linear relationship between mean and variance. Here, we propose NPMVS (Non-Parametric Mean Variance Smoothing), a method to estimate the mean and variance relationship, which is more general and can capture a wider range of non-linear relationships that exist in microarray experiments (Figure 1). We explore the mean-variance relationship by fitting a nonlinear curve using penalized splines [6,7]. In addition, inference is made upon shrinkage estimation of posterior means from Empirical Bayesian perspectives in our model, instead of *t*-statistic, which was used by Hu and Wright to test differential expression. Therefore, our approach has shrinkage estimation of both means and variances. First variances are smoothed using means, then means are smoothed assuming the variances are known. The simulation results showed that, under different mean-variance relationships, our method outperformed or was competitive with the other two popular shrinkage estimation methods, limma [2] and Gottardo et al. [8] generalized Bayesian statistic model *B*4, which assumes separate means and variances under different treatments for a given dataset. We also applied the three methods to a real biological dataset [9] to identify genes in cold stress regulatory pathways. With NPMVS, we detected more genes in the pathways

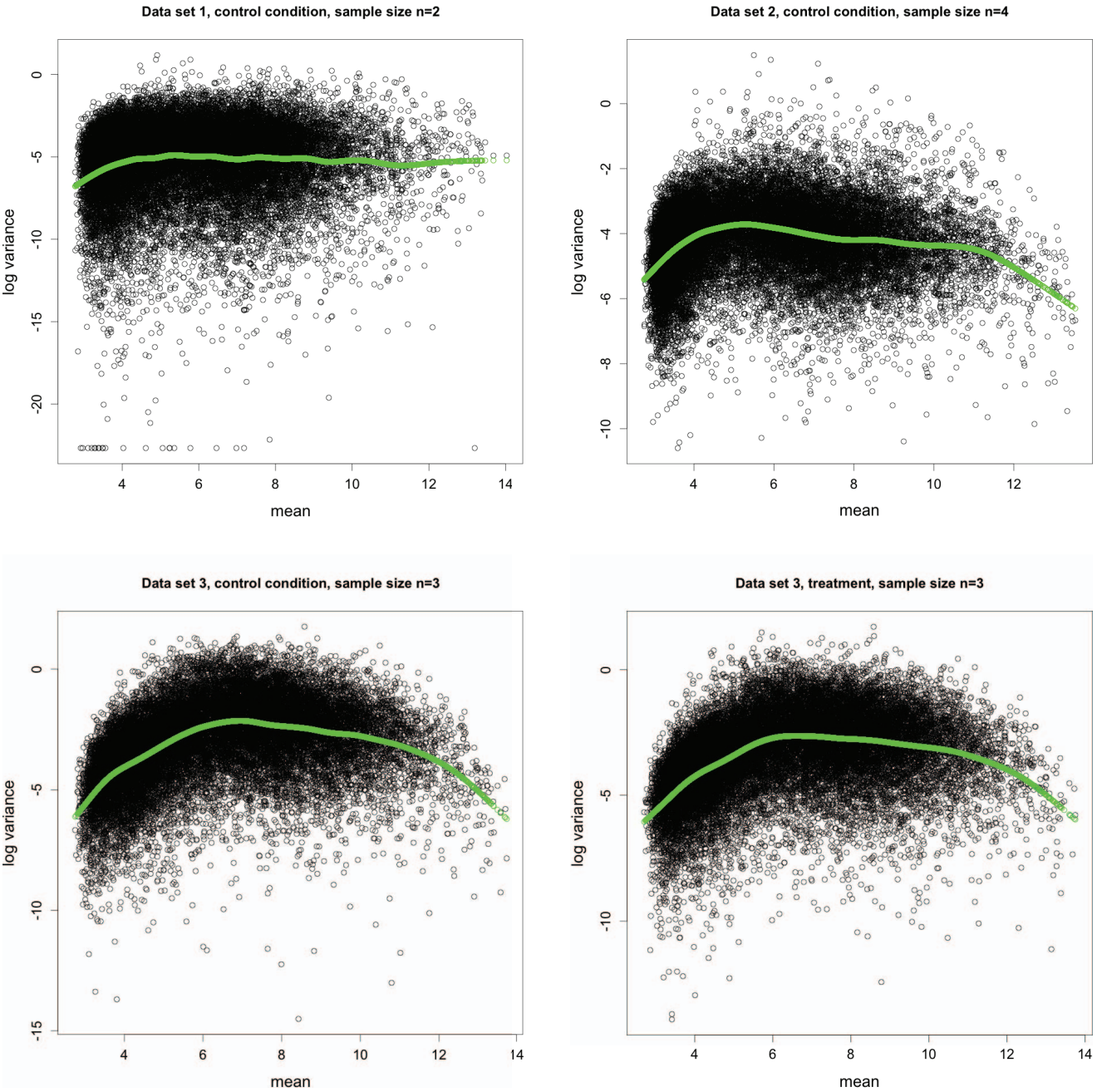

Figure 1. Relationship between sample mean and sample variance. Sample mean versus log sample variance plots of three different datasets from either control or treatment conditions. Smoothed variances using a non-paramteric method [6,7] is displayed with green lines. Sample size n is indicated for each dataset. The data sets were normalized with RMA method.
doi:10.1371/journal.pone.0019640.g001

and uniquely identified transcriptional changes in cell wall metabolism-related components under overexpression of a key transcription factor for freezing tolerance, *CBF2*.

Results and Discussion

Analysis of simulated data

To evaluate the performance of NPMVS, we compared it to the other two established methods, limma [2] (http://bioconductor.org/packages/2.5/bioc/html/limma.html), a linear model approach with variance shrinkage, and the $B4$ statistic from a Bayesian method [8]. We compared the performance of the three methods using simulated datasets. For each simulated dataset, a pair of data for control and treatment was generated with 10,000 genes. For the control data, we assumed that all genes expression level has a normal distribution with mean at 8 and standard deviation of 1. For the treatment data, 200 ($p=0.02$), 500 ($p=0.05$), and 1000 ($p=0.1$) genes were assigned as differentially expressed (DE) ones in different data designs. The up- and down-regulated DE genes were created with uniform distributions with different mean ranges. For up-regulated genes, 25% out of the total DE ones were assigned with mean from 8.1 to 11, another 25% from 11.1 to 14; For down-reulated genes, 25% were assigned with mean from 5 to 7.9 and another 25% from 2 to 4.9.

Samples for each genes were simulated as independent normal observations with four mean (μ_g) and variance (σ_g^2) relationships (Figure 2A) (see methods for details). We simulated datasets with 3 replicates. For each dataset design, different DE gene percentage and mean-variance relationship were imposed. One hundred simulated datasets were generated per design.

A plot of type I and type II error curve was used to compare the performance of the three methods (Figure 2B). For datasets with different percentage of DE genes, we had similar results. Here, we showed a representative result when the DE gene rate $p = 0.02$. Figure 2B shows the performance of the three methods in different mean-variance relationships. NPMVS and limma are competitive in the three mean-variance scenarios except case 0, in which limma displayed higher type I and II error rate than the other two methods. Compared with *B*4 statistic, NPMVS had better performance. In three cases (case 1 to 3), NPMVS outperformed the *B*4 statistic. In case 0, where non-linear relationship was displayed for variance and mean, NPMVS has better performance than *B*4 given a false positive rate less than 40% (Figure 2B).

Analysis of *CBF2* overexpression line data

Higher plants have complex regulatory mechanisms to temperature changes. Cold acclimation is a process by which plants increase their freezing tolerance in response to low, non-freezing temperatures. Previous studies have demonstrated that in Arabidopsis cold acclimation rapidly induces the expression of *CBF* genes, key transcription factors in response to low temperature. CBFs can increase freezing tolerance through activating downstream target genes (the CBF regulons) by binding to the target genes promoter region. To gain a better understanding of the CBF regulatory network, gene expression profiles were generated between *CBF2* overexpression lines (*CBF2_OX*), which constitutively overexpresses *CBF2* and can tolerance freezing without prior cold acclimation, and wild type (wt) [9,10]. The microarray data was analyzed by limma, *B*4 and NPMVS methods, respectively.

We retrieved DE genes with an adjusted cut-off *p* value less than 0.01 from limma, and a cut-off posterior probability greater than 0.99 from both *B*4 and NPMVS. The DE genes were further filtered with gene's average log 2 fold change greater than 1 or less than -1. As a result, NPMVS discovered more DE genes in both up- and down-regulated gene sets than limma and *B*4 (Figure 3) did. Limma identified 105 up-regulated genes, while *B*4 and NPMVS identified 191 and 238 up-regulated genes, respectively. In the down-regulated genes, limma only identified 17 genes, while *B*4 and NPMVS retrieved 48 and 88 genes, respectively. In addition, all genes identified by NPMVS were also found in the gene set identified by *B*4 and limma. Some genes discovered by NPMVS but not limma, like transcription factor *RAV1*, sugar related genes and cell wall synthesis genes, have been identified as CBF and cold responsive genes previously [9,11].

To evaluate the DE genes identified by the three methods, we accessed DE gene functions by gene ontology and cis-regulatory elements analysis to see if they are related to CBF and cold responsive pathway. Gene ontology enrichment analysis was performed on the three DE gene sets produced by the three methods (Table 1). Genes response to stress are enriched in all three up-regulated gene sets discovered by the different methods. The above result is consistent with the function of CBFs, which activates cold responsive genes as well as other abiotic stress

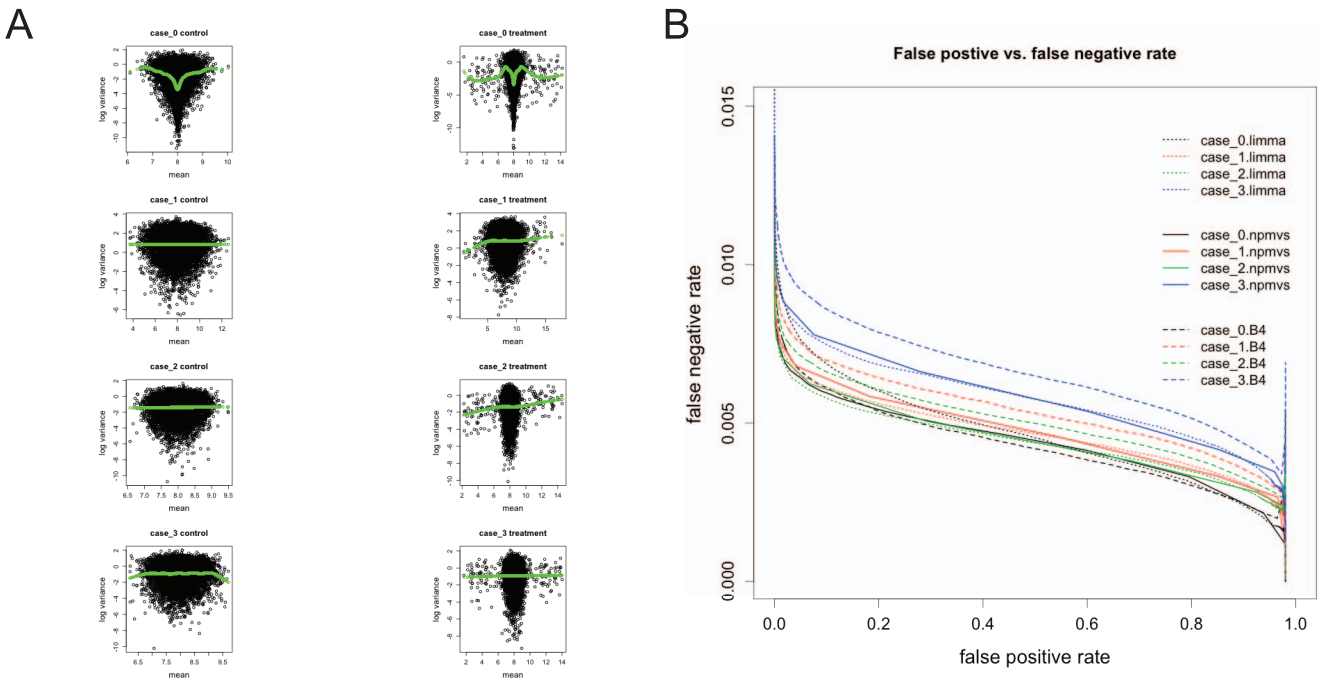

Figure 2. Simulation results from four mean-variance relationships. (A) The plots display mean versus log variance relationship in the four simulated data from case 0 to case 4. Simulated control data are presented on the left, and differentially expressed data are on the right. Smoothed variances using a non-paramteric method is displayed with green lines. (B) The plot displays false negative versus false positive rate for identifying DE genes in the simulated data using different methods. The false positive and false negative rate are the average rate from 100 simulated datasets. They were estimated over a range of cut-off values for each method. Dashed line, solid line, and dotted line represent Gottardo et al. [8] Bayseian (*B*4) method, NPMVS and limma, respectively. Four mean and variance relationships, case 0, case 1, case 2, and case 3 are represented by black, red, green and blue colours, respectively.
doi:10.1371/journal.pone.0019640.g002

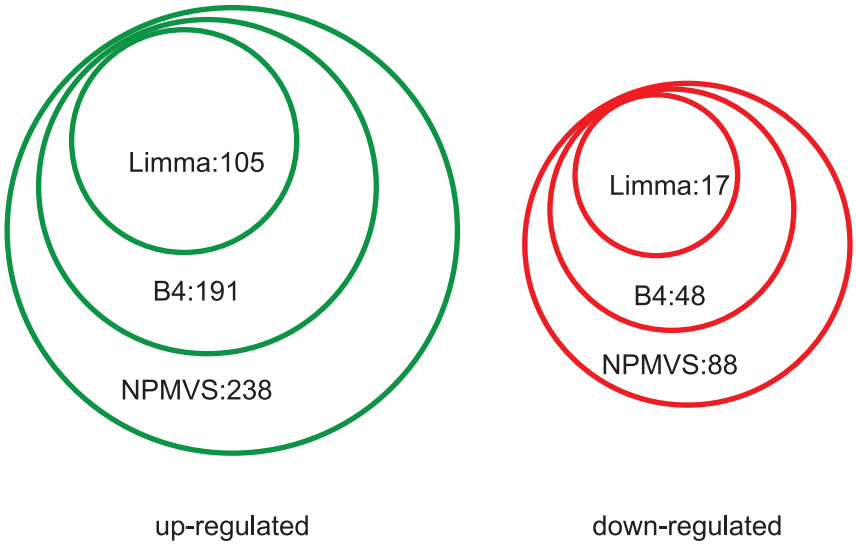

Figure 3. Identification of *CBF2_OX* differentially expressed genes. Up- and down-regulated genes greater than 2 fold changes are uncovered by three different methods with cut-off p (adjusted by Benjamini & Hochberg method) value less than 0.01 for limma, and a cut-off posterior probability greater than 0.99 for *B*4 and NPMVS, respectively.
doi:10.1371/journal.pone.0019640.g003

responsive genes [12]. The gene set from NPMVS showed the most significant enrichment with the smallest p value compared to the gene sets identified by the other two methods (Table 1). Some important CBF2 target genes, such as *RAV1*, were missed when using limma but identified by NPMVS. Other genes like *RCI2A*(Rare Cold Inducible 2A), which is induced under various stress conditions [13], was also uniquely discovered by NPMVS. In the down-regulated gene set, limma did not find any enriched GO terms. From NPMVS and *B*4, the GO term respond to stress was also enriched in the down-regulated gene sets.

Enrichment of cell wall components has been uniquely discovered by NPMVS in the down-regulated genes. Most of these genes are involved in cell wall metabolism (Table S1). This result is in agreement with the previous report that cell wall related genes were down-regulated in the later time points (8 hour to 168 hour) under cold stress [11]. It has also been reported that cold acclimation resulted in increase of cell wall weight and a change in cell wall composition [14]. All above suggest CBF mediated cold responsive pathway is involved in the cell wall re-organization.

We also investigated enrichment of cis-regulatory elements in the up- and down-regulated genes identified from the three methods (Table 2). A cis-regulatory element usually is a 4-12 word nucleotide motif in a gene promoter region. Transcription factors activate/repress expression of their regulons by binding to cis-regulatory elements in the promoter of their target genes. The transcription factor CBF2 binds to a conserved cis-regulatory element, the CRT/DRE, which contains a CCGAC core motif, presented in CBF target gene promoter regions, and thus activates transcription of the CBF regulons. We expected that the CRT/DRE cis-element would be enriched in the up- and/or down-regulated gene sets, assuming that the direct CBF target genes were in the discovered gene sets. As a result, the most highly enriched cis-element is the CRT/DRE CCGAC core motif, which has been discovered in all up-regulated gene sets from the three methods. NPMVS identified 24 and 59 more genes containing CCGAC than *B*4 and limma did, respectively (Table 2). There are another two motifs identified in all three up-regulated gene sets with a cut-off p value less 0.0001 (Table 2). The NPMVS gene set also showed the most significant p value of the second motif, ACGTG, which is an ABRE-like element that has been found in the promoter regions of cold, high-salinity and drought stress regulated genes [12]. Two additional enriched motifs have been identified by NPMVS and *B*4 but not limma. Three motifs have been uniquely discovered by NPMVS in the up-regulated dataset. Two out of three motifs, CCACG and CACGTG, which contain the BOXII and CACGTGMOTIF core elements, respectively, are related to light response [15,16]. The two motifs have not been reported in the previous CBF regulon motif studies [9]. No significant cis-acting elements have been found in all three down-regulated gene sets. The gene ontology and cis-regulatory element analysis indicated that NPMVS identified more stress responsive genes than limma and *B*4 did.

Table 1. Gene ontology enrichment analysis for *CBF2_OX* up- and down-regulated genes.

	limma	*B*4	NPMVS
Enriched GO in up-regulated genes	response to stress (8.39E-09)	response to stress (2.82E-10)	response to stress (2.83E-12)
Enriched GO in down-regulated genes	N.A.	response to stress (0.034)	response to stress (0.00016)
	N.A.	cell wall (0.034)	cell wall (8.91E-06)

P values, which are indicated in parentheses, were adjusted by Benjamini & Hochberg method.
doi:10.1371/journal.pone.0019640.t001

Table 2. Enriched cis-regulatory elements in *CBF2_OX* up-regulated genes.

	Limma		*B*4		NPMVS		PLACE[20] Annotation
word	*p* value	counts	*p* value	counts	*p* value	counts	
CCGAC	7.45E-35	78	1.06E-43	113	1.36E-43	137	DRE
ACGTG	6.02E-06	69	2.48E-09	104	1.42E-11	133	ABRE
ATGTCG	9.30E-26	47	2.04E-33	65	6.64E-31	77	N.A.
CCACG			4.01E-06	73	1.09E-05	89	BOXIIPCCHS
CGGCA			9.04E-06	59	2.79E-06	71	N.A.
ACACG					9.95E-07	133	GADOWNAT
CACGTG					6.68E-05	55	CACGTMOTIF
CGTGTC					1.44E-05	51	N.A.

doi:10.1371/journal.pone.0019640.t002

Conclusions

In this paper we compared the three shrinkage estimation methods, limma, *B*4, and NPMVS. Limma and *B*4 only have shrinkage estimation on variances, while our method NPMVS has shrinkage estimation on both means and variances. The simulation study showed that NPMSV performed better than limma in case 0, and the two methods were competitive in other mean-variance relationships. NPMSV outperformed *B*4 in case 1, 2, 3 and a competitor in case 0. The real microarray data from an overexpression line *CBF2*, which is a major regulator in cold and abiotic stress responsive pathways, was explored by the three methods. NPMVS identified more genes than both limma and *B*4 did. In addition, the gene set discovered by NPMVS included all genes identified by the other two methods. The gene ontology analysis showed that genes additionally identified by NPMVS are also related to stress response, which is consistent with previous findings for CBF2 targeted genes, implying that the NPMVS method makes a considerable improvement for gene detection. In agreement with gene ontology analysis, search of cis-acting elements in the up- and down-regulated gene sets showed that NPMVS identified more genes containing the core CBF response element, CCACG. NPMVS uniquely discovered genes involved in cell wall re-organization, which is consistent with previous cold stress microarray data [11]. Cis-acting elements, Box II and CACGTGMOTIF, which are light responsive components, were also uniquely discovered by NPMVS.

The good performance of NPMVS is mainly due to its shrinkage estimation for both means and variances. Our model used "smoothed" estimation of variances, which combines information from other genes. In addition, our method exploits mean and variance relationship, which is generally not considered in standard procedure. There is no specific type of relationship assumed for mean and variance; instead a nonparametric regression has been used. All above features contribute to the robustness of NPMVS. However, we should be aware that our NPMVS is based on the assumption that there is a relationship between mean and variance. Application of NPMVS will not be justified well if the assumption does not hold, namely, means are independent of variances. Mean and variance relationship should be investigated before the application of NPMVS.

Methods

For shrinkage estimation of both means and variances, our objective first is to obtain smooth estimation of gene specific variances and then to use estimated variances in a hierarchical model assuming it is known. Therefore, our approach has two steps. First, variances are smoothed using means. Second, means were smoothed by a hierarchical model assuming their variances (improved estimated variances) are known. Here, we first present a general hierarchical model and how to make inferences about DE genes based on the Bayes rule. Then we propose non-parametric estimation of variances and a new hierarchical model, which assumes that smoothed variances are known and takes more general form of a prior. Finally, we present a multiple sample case hierarchical model with smoothed variances.

Hierarchical model for one sample case

Let $I_g \sim B(1,p), g=1,\cdots,G$, be the Bernoulli random variable indicating whether the gene g is differentially expressed ($\mu_g \neq 0$), i.e., $\text{Prob}(I_g=1)=p$, where

$$I_g = \begin{cases} 0 & \text{if } I_g = 0 \\ 1 & \text{if } I_g \neq 0 \end{cases} \quad (1)$$

and μ_g denotes the mean expression level for the gene g. For each gene g we are interested in knowing if the gene is differentially expressed given the data.

$$\begin{aligned} y_{gj}|(\mu_g,\sigma_g^2) &\overset{ind}{\sim} N(\mu_g,\sigma_g^2) \\ \mu_g|\delta^2, I_g=0 &\overset{ind}{\sim} N(0,\delta_0^2) \\ \mu_g|\mu,\delta^2, I_g=1 &\overset{ind}{\sim} N(\mu,\delta^2) \end{aligned} \quad (2)$$

Then one can make inference on the basis of posterior probability

$$Prob(I_g=1|data) = \frac{pProb(_g|I_g=1)}{pProb(_g|I_g=1) + (1-p)Prob(_g|I_g=0)} \quad (3)$$

where y_g is the vector of measurement for gene g. It is easy to see that the posterior mean of μ_g is $\gamma_g \bar{y}_g + (1-\gamma_g)\mu$, where $\gamma_g = \dfrac{\delta^2}{\delta^2 + \sigma_g^2/n}$. This type of hierarchical model was considered by Baldi *et al.* [17], Lonnstedt *et al.* [18,19] and Gottardo *et al.* [8]. Moreover, in some of the above papers, δ^2 was taken in the form of $\kappa\sigma_g^2$ in which case the posterior mean does not even depend on

gene specific variances and the shrinkage factors (γ_g) are constants. Although the estimators preserve the shrinkage, the assumption $\delta^2=\kappa\sigma_g^2$ is hard to justify. In the case of $\delta^2=\kappa\sigma_g^2$, the posterior mean has a closed form expression. For example, the structure of the *B*4 estimator is $\gamma\bar{y}_g+(1-\gamma)\mu$, the shrinkage factor γ is constant over all genes. The only advantage of this is that the Bayesian computations get easier.

Proposed Hierarchical Model and Smoothing Variances

We examined Arabidopsis Affymetrix microarray data and plotted the log variances with their mean (Figure 1). The plot immediately suggested no linear relationship is appropriate. Thus we fitted a nonlinear curve using penalized spline [6,7]. Clearly, the spline did a very good fitting. Thus our objective is to first obtain the smooth estimate of the gene specific variances and then use them into a hierarchical model assuming they are known.

Smoothing variances:

We assume that the gene expressions y_{gj} for gth gene and jth replicate are normally distributed with mean μ_g and variance σ_g^2; $g=1,\cdots,G; j=1,\cdots,n$, and define $s_g^2=\frac{1}{n-1}\sum_j\left(y_{gj}-\bar{y}_g\right)^2$, $\bar{y}_g=\frac{1}{n}\sum_g y_{gj}$, where n is the sample size in the given data. We assume the following two level model

$$\begin{aligned} log\left(s_g^2\right) &= log\left(\sigma_g^2\right)+\varepsilon_g \\ log\left(\sigma_g^2\right) &= \boldsymbol{X}_g\beta+\boldsymbol{Z}_g u_g \end{aligned} \tag{4}$$

where the $\boldsymbol{X}_g$ ans $\boldsymbol{Z}_g$ are constructed from sample means and their quantiles [7]. It is easy to obtain the best linear unbiased predictor of $log(\sigma_g^2)$ as $\boldsymbol{X}_g\hat{\beta}+\boldsymbol{Z}_g\hat{\boldsymbol{u}}_g$. Note that all G genes are being used in this smoothing process. Let A_g be the estimated values. To estimate the probability of differential expression, we modified (2) in the hierarchical model (1)-(3) as described below.

$$\begin{aligned} y_{gk}|\mu_g &\overset{ind}{\sim} N(\mu_g,A_g), k=1,\cdots,n\ A_g\ known \\ \mu_g|\delta^2,I_g=0 &\overset{ind}{\sim} N(0,\delta_0^2) \\ \mu_g|\mu,\delta^2,I_g=1 &\overset{ind}{\sim} N(\mu,\delta_1^2) \end{aligned} \tag{5}$$

Lonnstedt and Speed [18], Gottardo *et al.* [8] and Lonnstedt and Britton [19] took $\delta_0^2=\delta_1^2=constant\times A$. The above structure facilitates the posterior calculations in a closed form. However, we do not see much of reasoning that the between variance would be a constant multiple of the within variance. Therefore, our model is more general.

Identifying the posterior distribution of $\mu_g|I_g=0$ as $N(0,A_g+\delta_0^2)$ and $\mu_g|I_g=1$ as $N(\mu,A_g+\delta_1^2)$, one can find

$$\begin{aligned} f(\ _g|I_g=0) &= \int\prod_{k=1}^{n} N(0,A_g+\delta_0^2)\pi(\delta_0^2)d\delta_0^2 \\ f(\ _g|I_g=1) &= \int\prod_{k=1}^{n} N(\mu,A_g+\delta_1^2)\pi(\delta_1^2)d\delta_1^2 \end{aligned} \tag{6}$$

where $\pi(\delta^2)$ is the prior distribution of δ^2. Previous works used Inverse gamma as a natural choice and the hyperparameter values were supplied. Note that again, unless δ^2 is a multiple of A_g, closed form expression does not exist even with IG prior distribution. Without any other prior information, we propose the uniform prior $\pi\left(\delta^2\right)=\frac{A_g}{\left(A_g+\delta^2\right)^2}$. In case of *IG*(*a*,*b*) prior distribution, the above two conditional distribution takes the form

$$\begin{aligned} &f\left(\boldsymbol{y}_g|I_g=0\right) \\ &=\frac{A_g\Gamma(n/2)}{\pi^{n/2}\left(\sum_k y_{gk}^2\right)^{(n/2)}}\int_{A_g}^{\infty} IG\left(\frac{n+a}{2},\frac{b+\sum_k y_{gk}^2}{2}\right)dt \\ &f\left(\boldsymbol{y}_g|I_g=1\right) \\ &=\frac{A_g\Gamma(n/2)}{\pi^{n/2}\left(\sum_k\left(y_{gk}-\mu\right)^2\right)^{(n/2)}}\int_{A_g}^{\infty} IG\left(\frac{n+c}{2},\frac{d+\sum_k\left(y_{gk}-\mu\right)^2}{2}\right)dt \end{aligned} \tag{7}$$

Choose *a*,*b*,*c* and *d* arbitrary positive number, say 0.5 or 0.01. They only involve evaluating the cumulative distribution function of inverse gamma distribution. Instead of using any prior distribution about μ, we shall use some pre-assigned quantity. The natural choice is to use grand mean expression value over all the genes.

If we use the uniform prior $\pi(\delta^2)$, then the conditional distributions take the form

$$f\left(y_g|I_g=0\right)=\int\frac{exp\left(-\frac{\sum_k y_{gk}^2}{2\left(A_g+\delta_0^2\right)}\right)}{(2\pi)^{\frac{n}{2}}\left(A_g+\delta_0^2\right)^{\frac{n}{2}}}\frac{A_g}{\left(A_g+\delta_0^2\right)^2}d\delta_0^2 \tag{8}$$

$$f\left(y_g|I_g=1\right)=\int\frac{exp\left(-\frac{\sum_k\left(y_{gk}-\mu\right)^2}{2\left(A_g+\delta_1^2\right)}\right)}{(2\pi)^{\frac{n}{2}}\left(A_g+\delta_1^2\right)^{\frac{n}{2}}}\frac{A_g}{\left(A_g+\delta_1^2\right)^2}d\delta_1^2 \tag{9}$$

This is what we have used in our study.

Multiple Sample Case

$$\begin{aligned} y_{gjk}|\mu_{gj} &\overset{ind}{\sim} N\quad(\mu_{gj},A_{gj}),\ A_{gj}, j=1,\cdots,J.\ known \\ \mu_{gj}|\delta^2,I_g=0 &\overset{ind}{\sim} N(\mu,\delta_0^2) \\ \mu_{gj}|\mu,\delta^2,I_g=1 &\overset{ind}{\sim} N(\mu_j,\delta_j^2) \end{aligned} \tag{10}$$

Using the similar prior distribution for variance parameters δ^2 used in (7), we can easily shown that the distributions are

$$\begin{aligned} &f\left(\boldsymbol{y}_{gj}|I_g=0\right) \\ &=\prod_{j=1}^{J}\frac{A_{gj}\Gamma\left(n_j/2\right)}{\pi^{n_j/2}\sum_k\left(y_{gjk}^2\right)^{n_j/2}}\int_{A_{gj}}^{\infty} IG\left(\frac{n_j+a}{2},\frac{1}{2}\left(\sum_k\left(y_{gjk}^2\right)+b\right)\right)dt \end{aligned} \tag{11}$$

$$f\left(\boldsymbol{y}_{gj}|I_g=1\right)=\prod_{j=1}^{J}\frac{A_{gj}\Gamma\left(n_j/2\right)}{\pi^{n_j/2}\left(\sum_k\left(y_{gjk}-\mu_j\right)^2\right)^{n_j/2}}\int_{A_{gj}}^{\infty}IG\left(\frac{n_j+c}{2},\frac{1}{2}\left(d+\sum_k\left(y_{gjk}-\mu_j\right)^2\right)\right)dt \quad (12)$$

If we use the uniform prior $\pi(\delta^2)$ then the above conditional distributions are

$$f\left(\boldsymbol{y}_{gj}|I_g=0\right)=\int\prod_{j=1}^{J}\frac{exp\left(-\frac{\sum_k y_{gjk}^2}{2\left(A_{gj}+\delta_0^2\right)}\right)}{(2\pi)^{\frac{n_j}{2}}\left(A_{gj}+\delta_0^2\right)^{\frac{n_j}{2}}}\frac{A_{gj}}{\left(A_{gj}+\delta_0^2\right)^2}d\delta_0^2 \quad (13)$$

$$f(\boldsymbol{y}_{gj}|I_g=1)=\int\prod_{j=1}^{J}\frac{exp\left(-\frac{\sum_k(y_{gjk}-\mu)^2}{2(A_{gj}+\delta_j^2)}\right)}{(2\pi)^{\frac{n_j}{2}}(A_{gj}+\delta_j^2)^{\frac{n_j}{2}}}\frac{A_{gj}}{(A_{gj}+\delta_j^2)^2}d\delta_j^2 \quad (14)$$

Note that *apriori* the δ_j's are assumed to be independent. There is a number of possible modifications can be easily done. For example, one might assume same variance for all the conditions for non regular gene means or different variances for regular gene means.

We evaluated the one dimensional integral in (13) and (14) using 20 point Gauss-Hermite procedure.

Method Evaluation

To evaluate the performance of NPMVS, we compared it to other two established methods, limma [2] (http://bioconductor.org/packages/2.5/bioc/html/limma.html) and Gottardo et al. [8] Bayesian method. We compared the performance of the three methods using simulated datasets and a real biological dataset, in which an overexpression line *CBF2_OX* was compared with wild type control.

Simulated Data. We applied different mean-variance relationships for the simulated data. First, expression means were generated from normal distribution for non-differentially expressed genes, $\mu_g \overset{ind}{\sim} N(8,1)$; For DE genes, $\mu_g \overset{ind}{\sim} U(a,b)$, where $a=8.1,b=11$ or $a=11.1,b=14$ for up-regulated and highly up-regulated genes, respectively, and $a=5,b=7.9$ or $a=2,b=4.9$ for down-regulated and deeply down-regulated genes, respectively. Secondly, variance σ_g^2 was generated, $\sigma_g^2 \overset{ind}{\sim} f(\mu_g|\beta)$. The plots of four mean-variance relationships are displayed in Figure 2A. Last, expression data y_g was generated from normal with mean and variance produced in the first two steps, $y_{gk} \overset{ind}{\sim} N(\mu_g,\sigma_g)$, where k = 3 is for three replicates. The most non-linear relationship is symbolized as case 0. In the simulation data, the choice of parameters β was based on the variance range observed from real datasets. The variance range produced in the simulation data was closed to the one surveyed from four real datasets (GSE9955, GSE5520, GSE5536, and GSE5727). Note that, since we used a nonparametric method, there is no need to estimate the beta parameters in real data application.

Real Biological Data. The CBF2 data (GEO series number: GSE5566) includes two genotypes, two independent *CBF2* overexpression lines and its corresponding *Arabidopsis thaliana* wild type, and two samples for each genotype. The microarray platform is Affymetrix ATH1 GeneChip. The raw CEL files were normalized by RMA. Fisher's exact test for one-tail (over-presented) was applied to the Gene Ontology enrichment analysis. The p value for over-presented GO terms i is $\sum_{x=r}^{k}f(r|N,M,k)=\frac{\binom{M}{r}\binom{N-M}{k-r}}{\binom{N}{k}}$, where M is the total DE gene number; N is the total gene number in the genome; k is the gene number in GO term i; and r is the number of genes which belong to DE gene list in GO term i. Benjamini and Hochberg false discovery rate correction was used for adjusting p values. Five hundred base pair promoter region sequence for each gene was used for cis-regulatory element analysis via a *de novo* motif searching tool ELEMENT (http://element.cgrb.oregonstate.edu/). An enumerative method in ELEMENT was used for counting 4-8 mer DNA words. By comparing a word frequency for a given gene set to samples from the whole *Arabidopsis* genome sequence (a bootstrap procedure), a corresponding Z score and p value (adjusted by Benjamin and Hochberg FDR method) were calculated in ELEMENT to estimate if the word is over-presented in the given gene set.

Supporting Information

Table S1
(CSV)

Acknowledgments

We thank Sarah Gilmour for critical comments on the manuscript.

Author Contributions

Conceived and designed the experiments: TM PH. Performed the experiments: PH. Analyzed the data: PH. Contributed reagents/materials/analysis tools: PH. Wrote the paper: PH TM.

References

1. Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical bayes analysis of a microarray experiment. Journal of the American Statistical Association 96: 1151–1160.
2. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology 3.
3. Cui X, Kerr MK, Churchill GA (2003) Transformations for cdna microarray data. Statistical applications in genetics and molecular biology 2.
4. Tong T, Wang Y (2007) Optimal shrinkage estimation of variances with applications to microarray data analysis. Journal of the American Statistical Association 102: 113–122.
5. Hu J, Wright F (2007) Assessing differential gene expression with small sample sizes in oligonu-cleotide arrays using a mean-variance model. Biometrics 63: 41–49.
6. Green P, Silverman B (1994) Nonparametric regression and generalized linear models. Glasgow: Chapman and Hall.
7. Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge: Cambridge University Press.
8. Gottardo R, Pannucci JA, Kuske CR, Brettin T (2003) Statistical analysis of microarray data: a Bayesian approach. Biostat 4: 597–620.
9. Vogel JT, Zarka DG, Van Buskirk HA, Flower SG, Thomashow MF (2005) Roles of the cbf2 and zat12 transcription factors in configuring the low temperature transcriptome of arabidopsis. The Plant Journal 41: 195–211.
10. Gilmour S, Fowler S, Thomashow MF (2004) Arabidopsis transcriptional activators cbf1, cbf2, and cbf3 have matching functional activities. Plant Molecular Biology 54: 767–781.
11. Fowler S, Thomashow MF (2002) Arabidopsis transcriptome profiling indicates that multiple reg-ulatory pathways are activated during cold

acclimation in addition to the CBF cold response pathway. Plant Cell 14: 1675–1690.
12. Shinozaki K, Yamaguchi-Shinozaki K, Seki M (2003) Regulatory network of gene expression in the drought and cold stress responses. Current Opinion in Plant Biology 6: 410–417.
13. Medina J, Catala R, Salinas J (2001) Developmental and stress regulation of RCI2A and RCI2B, two cold-inducible genes of arabidopsis encoding highly conserved hydrophobic proteins. Plant Physiol 125: 1655–1666.
14. Weiser RL, Wallner SJ, Waddell JW (1990) Cell wall and extensin mRNA changes during cold acclimation of pea seedlings. Plant Physiol 93: 1021–1026.
15. Terzaghi WB, Cashmore AR (1995) Light-regulated transcription. Annual Review of Plant Physiology and Plant Molecular Biology 46: 445–474.
16. Hudson ME, Quail PH (2003) Identification of promoter motifs involved in the network of phytochrome a-regulated gene expression by combined analysis of genomic sequence and microarray data. Plant Physiol 133: 1605–1616.
17. Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 17: 509–519.
18. Lonnstedt I, Speed T (2002) Replicated microarray data. Statistica Sinica 12: 31–46.
19. Lonnstedt I, Britton T (2005) Hierarchical bayes models for cdna microarray gene expression. Biostat 6: 279–291.
20. Higo K, Ugawa Y, Iwamoto M, Higo H (1998) Place: A database of plant cis-acting regulatory dna elements. Nucleic Acids Research 26: 358–359.