Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 10

Normalization Method for Transcriptional Studies of Heterogeneous Samples – Simultaneous Array Normalization and Identification of Equivalent Expression

Li-Xuan Qin* Jaya M. Satagopan†

*Memorial Sloan-Kettering Cancer Center, qinl@mskcc.org

†Memorial Sloan-Kettering Cancer Center, satagopj@mskcc.org

Normalization Method for Transcriptional Studies of Heterogeneous Samples – Simultaneous Array Normalization and Identification of Equivalent Expression*

Li-Xuan Qin and Jaya M. Satagopan

Abstract

Normalization is an important step in the analysis of microarray data of transcription profiles as systematic non-biological variations often arise from the multiple steps involved in any transcription profiling experiment. Existing methods for data normalization often assume that there are few or symmetric differential expression, but this assumption does not always hold. Alternatively, non-differentially expressed genes may be used for array normalization. However, it is unknown at the outset which genes are non-differentially expressed. In this paper we propose a hierarchical mixture model framework to simultaneously identify non-differentially expressed genes and normalize arrays using these genes. The Fisher's information matrix corresponding to array effects is derived, which provides useful intuition for guiding the choice of array normalization method. The operating characteristics of the proposed method are evaluated using simulated data. The simulations conducted under a wide range of parametric configurations suggest that the proposed method provides a useful alternative for array normalization. For example, the proposed method has better sensitivity than median normalization under modest prevalence of differentially expressed genes and when the magnitudes of over-expression and under-expression are not the same. Further, the proposed method has properties similar to median normalization when the prevalence of differentially expressed genes is very small. Empirical illustration of the proposed method is provided using a liposarcoma study from MSKCC to identify genes differentially expressed between normal fat tissue versus liposarcoma tissue samples.

KEYWORDS: gene expression, normalization, mixture models, Fisher's information

---

*We thank two anonymous reviewers for insightful comments. We also thank Dr. Sam Singer at MSKCC for providing the liposarcoma data. This work was supported in part by NIH grants 2P01CA047179-15A2 (LXQ) and UL1RR024996 of the Clinical and Translation Science Center at Weill Cornell Medical College (LXQ and JS).

1 Introduction

Microarray is a high-throughput tool that can simultaneously measure the expression level of thousands of transcripts on a genome-wide scale (Schena et al. 1995; Lipshutz et al. 1999). It is increasingly used to determine the underlying biological differences in disease subtypes or treatment effects (Spellman et al. 1998; Perou et al. 2000; LaTulippe et al. 2002; Singer et al. 2007). A microarray experiment involves a complex multi-step process, including extraction of mRNAs, reverse transcription to cDNAs, denaturation of cDNAs, hybridization to probes on a microarray, and image scanning of fluorescence (Schena et al. 1995; Lipshutz et al. 1999; Nguyen et al. 2002). Owing to the complexity of the underlying process, the resulting data consist of multiple sources of variation, including systematic variation due to biological effects (the effect of interest), systematic variation due to experimental process, and stochastic noise. Common causes of systematic non-biological variation are background fluorescence, array batch difference, print-tip spatial effects, and dye effects (for two-color cDNA arrays).

The process of estimating and subsequently removing the effects due to experimental process is called preprocessing (Nguyen et al. 2002; Irizarry et al. 2003). Assumptions need to be introduced to make non-biological systematic effects identifiable from biological systematic effects. Preprocessing often involves multiple steps. The primary goal of this paper is to investigate methods for removing array effects so that the expression measures are comparable across arrays. We refer to this process as "normalization". When done appropriately, normalization can improve the accuracy of the subsequent statistical analysis, such as differential expression detection (Reilly et al. 2003). As pointed out by a referee, when analyzing real data, one needs to also consider other typical features of microarray data, such as the skewed distribution of intensity measurements (Purdom and Holmes 2005), the additive-multiplicative noise problem (Rocke and Durbin 2001), and the variance stabilization problem (Durbin and Rocke 2004; Huber et al. 2003). Addressing all these issues simultaneously using a single model may be an ambitious goal, particularly since the sample size of these studies is substantially smaller than the number of probes on the array. Therefore, it may be pragmatic to address them in multiple steps so as to avoid any identifiability issues that may arise under simultaneous modeling. For example, there is a large body of literature on data transformations (Atkinson 1985) that may be applied to the intensity data to address issues such as skewness, non-linearity, and variance stabilization. In this paper we focus on the array normalization issue, assuming that separate steps have been undertaken previously for other data pre-processing needs.

Existing methods for normalization often follow one of the two strategies.

1. **All-gene normalization**. This strategy makes the distribution of the data similar across arrays by using all genes on each array for normalization. It is based on the implicit assumption that few or symmetric over-/under-expression exists among genes. Methods based on this strategy include median normalization, non-linear normalization (often intensity-dependent) (Yang et al. 2002), and quantile normalization (Bolstad et al. 2003). In situations where this assumption does not hold, these methods tend to attenuate biological effects and hide differentially expressed genes. For example, under median normalization, the estimates of biological effects might be biased, as the median of an array (denoted as m) is determined by the following equation: $P(y > m) = P(\text{equivalent expression})P(y > m|\text{equivalent expression}) + P(\text{differential expression})P(y > m|\text{differential expression})$.
2. **Some-gene normalization**. This strategy selects a subset of genes (called "control genes") and makes the mean of their data distribution similar across arrays. Choices of control genes include spiked-in genes, house-keeping genes, and rank-invariant genes (Li and Wong 2001). These methods assume that the expression of each control gene is constant across the samples under study; hence the reliability of the control genes is critical. Spike-in genes are typically chosen to be genes with constant expression patterns across a panel of tissue types or treatments in prior studies. House-keeping genes are those believed to hold important biological function in cells and expected to be consistently expressed (for example, GAPDH and beta-Actin), but fluctuations of their expression do occur (Thellin et al. 1999). An example is an eight-tissue study conducted by Affymetrix, which compares the expression of GAPDH, beta-Actin, and the 100 (spiked-in) normalization control genes on HG-U133A arrays (Affymetrix document: Performance and Validation of the Genechip Human Genome U133 Set, available at Affymetrix websit http://www.affymetrix.com/index.affx). The validity of rank-invariant genes, whose ranks are consistent across arrays, for normalization depends on the assumption of independence between differential expression and expression intensity. For example, it is plausible that genes having high expression are more likely to be over-expressed, which is ignored under the rank-invariant formulation.

Cancer is a complex disease characterized by within-patient and, most notably, between-patient tumor heterogeneity. To normalize microarrays for

DOI: 10.2202/1544-6115.1339

such heterogeneous samples, the assumption of all-gene methods might not hold and the choice of control genes is not straightforward. Normalization methods based on non-differentially expressed genes have been used for two-channel array data (Zhao et al. 2005; Reilly et al. 2003). In this paper, we employ a hierarchical Gaussian mixture model to identify differentially expressed genes in the single-channel oligonucletide arrays. The normalization factor is a parameter of this model. The proposed model has parallels to penalized regression approach (Hastie et al. 2001). We derive the Fisher information corresponding to the normalization parameter, which provides intuition and mathematical justification guiding the choice of array normalization method. Simulation studies are conducted to evaluate properties of the proposed method.

1.1 Motivating Example

Our work is motivated by an ongoing study of gene expressions in liposarcoma at Memorial Sloan-Kettering Cancer Center. Liposarcoma is a rare type of tumor that arises in fat cells. It has five major variants: well-differentiated, de-differentiated, myxoid, myxoid/round cell (MRC), and pleomorphic. A microarray study was performed to measure gene expression among liposarcoma tumors and normal fat tissues using Affymetrix HG-U133A arrays consisting of 22,215 probe sets, 100 of which are control probesets (Affymetrix website). In this paper we consider data from 8 MRC tumor samples and 12 normal fat samples. Figure 1 shows the un-normalized probe-level intensities for all genes (22,215 probesets) and for the control genes (100 probesets). Each curve represents the empirical density of the gene expressions from a single array (that is, sample). It is evident that there is substantial variation among arrays even within the normal fat group. Clearly, the expression levels of the control genes are not similar across arrays. These observations suggest the need for appropriate normalization of the arrays to identify differentially expressed genes.

This paper is organized as follows. Section 2 describes the proposed hierarchical mixture model for normalization and discusses the identifiability of model parameters. The Fisher information corresponding to the normalization parameter is derived and used to provide further insights into the proposed method. Section 3 illustrates the operating characteristics of the proposed method using simulated data. Application of the proposed method to the liposarcoma data is detailed in Section 4, and compared with median normalization, control-gene median normalization, and quantile normalization. Section 5 provides concluding remarks and recommendations for practice.

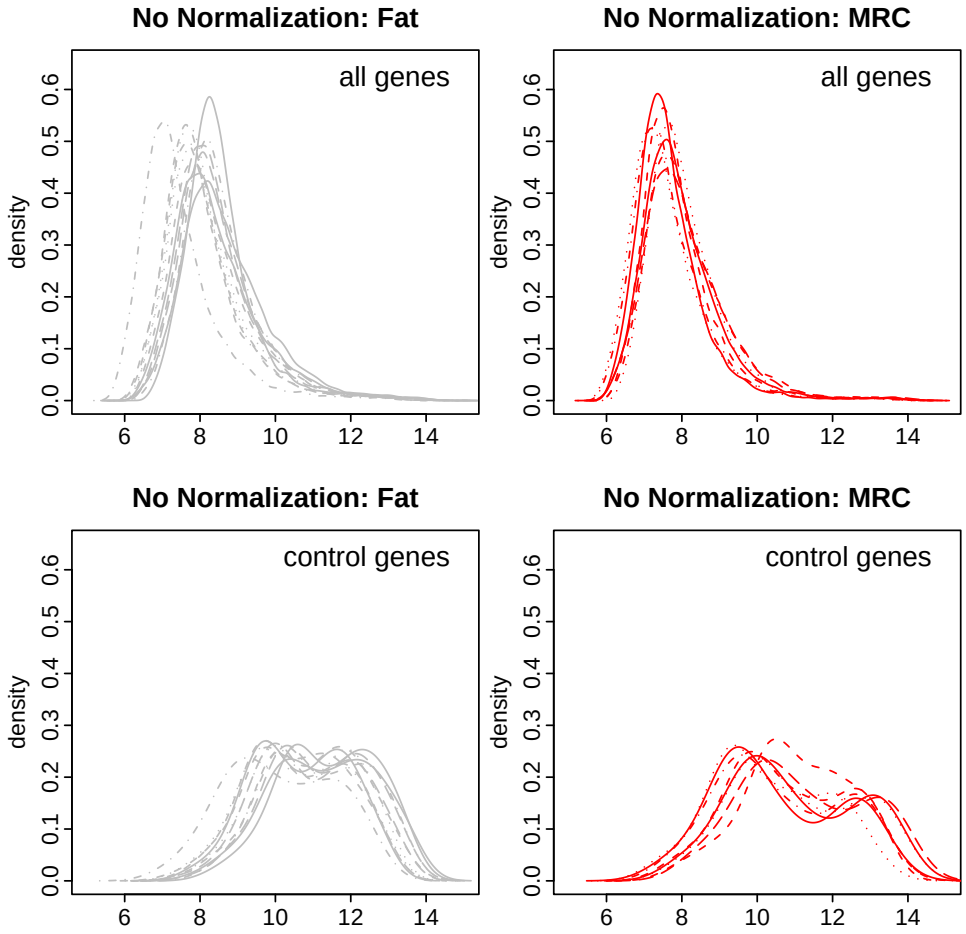

Figure 1: Density plots of normal fat (left panels) and liposarcoma (right panels) arrays in absence of normalization. Top panels are for all genes and bottom panels are for the 100 control genes. In each panel, each curve represents one array.

DOI: 10.2202/1544-6115.1339

2 Method

2.1 The Model

We will present the proposed model in a two-class setting (for example, tumors vs. normal tissues). Denote y_{igp} as the expression intensity (typically log2 transformed for variance stabilization) for sample i, gene g, and probe p (nested within gene g), and x_i as the indicator of disease status for sample i (0 for normal tissues and 1 for tumors). The gene expression y_{igp} is modeled using analysis of variance (ANOVA) with the following components: array effect α_i, gene effect β_g, probe effect δ_{gp} (nested within gene effect), interaction between array effect and gene effect γ_{ig}, and measurement error ϵ_{igp} (see Equation 1 below). These components can be interpreted as follows. The array effect α_i represents the normalization parameter or the systematic non-biological variation, averaged over all the genes on the array. This parameter needs to be estimated accurately so that the differentially expressed genes can be identified with adequate sensitivity. The gene effect β_g represents the baseline expression level of gene g. The probe effect δ_{gp} is the contribution of an individual probe p to the expression level of gene g. In this paper we ignore probe-specific effects and set $\delta_{gp} = 0$ (Equation 2). The interaction γ_{ig} is the additional effect of gene g on the expression level that arises due to the disease status x_i, and we represent this as $\gamma_{ig} = x_i\gamma_g$ (Equation 3). Thus, the expected expression of gene g is β_g among the controls and $\beta_g + \gamma_g$ among the cases. And $\gamma_g = (\beta_g + \gamma_g) - \beta_g$ is the extent to which gene g is differentially expressed in the cases relative to the controls. The component ϵ_{igp} is random noise, assumed to have an independent $N(0, \sigma^2)$ distribution.

$$
\begin{aligned}
y_{igp} &= \alpha_i + \beta_g + \delta_{gp} + \gamma_{ig} + \epsilon_{igp} && (1)\\
&= \alpha_i + \beta_g + \gamma_{ig} + \epsilon_{igp} && (2)\\
&= \alpha_i + \beta_g + x_i\gamma_g + \epsilon_{igp} && (3)\\
\epsilon_{igp} &\sim N(0, \sigma^2)\\
\beta_g &\sim N(0, \tau^2)\\
\gamma_g &\sim (1-\pi)I\{0\} + \pi\lambda N(\mu_o, \psi^2) + \pi(1-\lambda)N(\mu_u, \xi^2)\\
&\quad \text{where } \mu_o > 0 > \mu_u
\end{aligned}
$$

The normalization parameter α_i may be interpreted as the systematic non-biological variation, averaged over all the genes on the array. The effect β_g may be assumed to be 0 once the non-biological variation is eliminated. However, it is conceivable that some genes may naturally be over- or under-expressed in the control population and, hence, the assumption of $\beta_g = 0$

may not be uniformly applicable to all genes. Under such uncertainty, we may postulate a stochastic framework for the underlying true expression β_g as $\beta_g \sim N(0, \tau^2)$, where τ^2 represents the uncertainty about the assumption of zero gene effect among the controls. A gene g may be equivalently expressed ($\gamma_g = 0$), over-expressed ($\gamma_g > 0$), or under-expressed ($\gamma_g < 0$) among the cases relative to the controls. Denote μ_o and μ_u ($\mu_o > 0 > \mu_u$) as the mean over- and under-expression of the differentially expressed genes. As before, we can conceptually postulate a stochastic framework for the over- and under-expressed genes. Denoting π as the proportion of differentially expressed genes and λ as the proportion of over-expressed genes among those differentially expressed, we posit a mixture distribution for the effect γ_g: a mass at 0 with probability $1 - \pi$, a $N(\mu_o, \psi^2)$ distribution with probability $\pi\lambda$, and a $N(\mu_u, \xi^2)$ distribution with probability $\pi(1 - \lambda)$. Here the variances ψ^2 and ξ^2 reflect the uncertainty about the mean over- or under-expression effects of the differentially expressed genes.

A more convenient mathematical construct, which will be helpful for obtaining parameter estimates, can be set up by introducing binary variables o_g and u_g, where $o_g = 1$ if gene g is over-expressed and 0 otherwise, and $u_g = 1$ if gene g is under-expressed and 0 otherwise. Hence, $\gamma_g = o_g\gamma_{og} + u_g\gamma_{ug}$, where $\gamma_{og} \sim N(\mu_o, \psi^2)$ and $\gamma_{ug} \sim N(\mu_u, \xi^2)$. Further, o_g and u_g have a multinomial distributions with probabilities $\pi\lambda$ and $\pi(1 - \lambda)$, respectively.

$$\begin{aligned} \gamma_g &= o_g\gamma_{og} + u_g\gamma_{ug} \\ (1 - o_g - u_g, o_g, u_g) &\sim Multinomial(1, (1 - \pi, \pi\lambda, \pi(1 - \lambda))) \\ \gamma_{og} &\sim N(\mu_o, \psi^2) \\ \gamma_{ug} &\sim N(\mu_u, \xi^2) \end{aligned} \tag{4}$$

2.2 Motivation of the Use of Gaussian Mixture Model

The Gaussian mixture model in our work is motivated by the following observations. When analyzing a large number of putative risk factors (such as gene expressions) in relation to an outcome of interest, it is now widely accepted that analyzing one risk factor at a time may not be a useful strategy (for example, Kendziorski et al. 2003). It can lead to imprecise estimates of the effects and can easily result in false positive findings. Penalized regression techniques have been proposed as a useful strategy for addressing such issues (for example, Hastie et al. 2001). This approach estimates the effects by imposing suitable stability constraints, and has been successfully used for both class comparison and class prediction problems.

DOI: 10.2202/1544-6115.1339

Two very popular and useful penalized regression methods are: ridge regression (Hoerl 1962) and the LASSO (Tibshirani 1996). Ridge regression imposes a constraint on the sum of the squares of the effects. This is equivalent to imposing an exchangeable normal prior distribution for the effects. The variance of this prior distribution is intimately related to the ridge constraint. In contrast, the LASSO imposes a constraint on the sum of the absolute values of the effects. This is equivalent to imposing an exchangeable double exponential (equivalently, Laplace) prior distribution for the effects. The variance of this prior is intimately related to the LASSO constraint.

Both ridge regression and LASSO provide shrinkage estimates of the effects. It is well-known that LASSO places higher a priori mass around 0 (Tibshirani 1996). Thus, LASSO can identify null effects with better specificity than ridge regression. LASSO is also closely related to robust estimation techniques. Carroll (1980) showed that mixture distributions of the form $(1-\epsilon)\Phi + \epsilon H$ can provide robust inferences. Here Φ is a standard normal distribution and H is any symmetric distribution. Carroll (1980) termed this the "normal centre-exponential tails" distribution, and used this approach for robust inferences when applying Box-Cox type of power transformations to the outcome to achieve normality. One can plot the "normal centre-exponential tails" distribution with H as an indicator function having mass at 0 and by considering various choices of ϵ. From such a plot, it can be easily seen that this mixture distribution has similarities to a Laplace prior.

2.3 Identifiability of Model Parameters

Given the equivalent-expression, over-expression, or under-expression status of each gene, the unknown parameters of the proposed mixture model are (a) the normalization parameters α_i's, (b) the means of over-expressed genes μ_o and under-expressed genes μ_u, (c) the variances of treatment effects for over-expressed genes ψ^2 and under-expressed genes ξ^2, (d) the variance of gene effects τ^2, and (e) the variance of measurement error σ^2. Before describing the algorithm to estimate these unknown parameters, it will be useful to understand if these parameters are indeed identifiable using the observed data. Table 1 gives the method of moments estimates, illustrating that the unknown parameters can be estimated unbiasedly using the gene-specific covariances and variances of the probe intensities.

- If sample i is a control (that is, $x_i = 0$), then $E(y_{igp}) = \alpha_i$. An unbiased estimate of the normalization factor α_i is the average of all the probe intensities on array i.

Table 1: Parameter identifiability in the mixture model for array normalization.

	Equal-Expression	Over-Expression	Under-Expression
$E(y_{igp})$	α_i	$\alpha_i + x_i\mu_o$	$\alpha_i + x_i\mu_u$
$var(y_{igp})$	$\sigma^2 + \tau^2$	$\sigma^2 + \tau^2 + x_i\psi^2$	$\sigma^2 + \tau^2 + x_i\xi^2$
$cov(y_{igp}, y_{igq})$	τ^2	$\tau^2 + x_i\psi^2$	$\tau^2 + x_i\xi^2$

- If sample i is a case (that is, $x_i = 1$), then $E(y_{igp}) = \alpha_i$, $E(y_{igp}) = \alpha_i+\mu_o$, or $E(y_{igp}) = \alpha_i + \mu_u$, depending upon whether gene g is equivalently expressed, over-expressed, or under-expressed. An unbiased estimate of α_i is the average probe intensity of the equivalently expressed genes on array i. Were we to know *a priori* that $\mu_o = -\mu_u$, then α_i may be unbiasedly estimated as the average intensity of all genes on array i.

- Once α_i is estimated, an unbiased estimate of μ_o (or μ_u) can be obtained as the difference between the average probe intensity of the over-expressed (or under-expressed) genes and the equivalently-expressed genes, since o_g and u_g are assumed known.

- The errors are independent. Therefore, when gene g is equivalently expressed or when sample i is a control, we have $\tau^2 = cov(y_{igp}, y_{igq})$, the covariance between probes within a gene. When gene g is over-expressed, we have $\tau^2 + x_i\psi^2 = cov(y_{igp}, y_{igq})$. Finally, when gene g is under-expressed, we have $\tau^2 + x_i\xi^2 = cov(y_{igp}, y_{igq})$. This suggests that all four unknown variance parameters can be estimated unbiasedly using the gene-specific covariances and variance of the probe intensities.

2.4 Parameter Estimation

Since differential expression status, $(o_g,\ u_g)$'s, are not observed, we use the EM algorithm to maximize the classification likelihood for the mixture model. (Details of the implementation are presented in Appendix A.)

- In the E-step, o_g and u_g are estimated for each gene in the form of posterior probabilities.

- In the M-step, array effects α_i's are estimated as the average among the non-differentially expressed genes for each array and the variances for random effects are estimated by fitting a linear mixed effects model.

DOI: 10.2202/1544-6115.1339

2.5 Fisher's Information of the Normalization Parameters

The variance of the parameter estimates can be derived using the Fisher's information matrix. We are particularly interested in estimating the normalization parameter α_i's as accurately as possible. The differential expression status of the genes are unknown at the outset. This missing piece of information can have an impact on the precision when the parameters are estimated. The precision is given by the Fisher's information, defined as the second derivative of the log likelihood function with respect to α_i, which can provide important guidance to assess trade-offs in estimating the α_i's. Here we calculate Fisher's information corresponding to α_i and evaluate the underlying insights.

We observe the probe level intensities y_{igp}'s. The differential expression status of each gene is unobservable or missing. Therefore, the information corresponding to α_i can be obtained using the probe intensity y_{igp} of array i as the difference between the complete data information and the missing data information ((Louis 1982); Appendix B). Denoting P as the number of probes per gene, the missing data information, I_m, is given by:

$$
\begin{aligned}
I_m &= P^2 \frac{x_i}{1+x_i} \sum_g \{B_{1g} + B_{2g} + B_{3g}\} \qquad \text{where} \qquad (5)\\
B_{1g} &= (\bar{y}_{ig.} - \alpha_i)^2 w_{0g}(1 - w_{0g})\\
B_{2g} &= \mu_o^2[(1 - w_{0g}) - (w_{1g} - w_{2g})^2]\\
B_{3g} &= 2(\bar{y}_{ig.} - \alpha_i)\mu_o w_{0g}(w_{1g} - w_{2g})
\end{aligned}
$$

Note that w_{0g}, w_{1g}, and $w_{2g} = 1 - w_{0g} - w_{1g}$ are the posterior probabilities of equivalent-, over-, and under-expression of gene g, respectively. It is desirable to have the missing data information as small (that is, preferably as close to 0) as possible, so that the estimate of α_i is precise.

Recall that $x_i = 1$ for cases and 0 for controls. The term $\frac{x_i}{(1+x_i)}$ in Equation 5 is 0 when $x_i = 0$, suggesting that missing information for the array effect is only a concern for cases. This is consistent with intuition that gene expression remains at the expected level among controls and differential expression results in altered expression level of a gene among the cases. Hence, array normalization will be critical for cases. The first term (B_1) within the summation vanishes for those genes whose the differential expression probability w_{0g} is 0 or 1. Hence, genes with ambiguous differential expression probabilities contribute to missing information, particularly when their normalized expression level is large. This observation suggests the need to estimate the differential expression status of gene g with minimum or no ambiguity to the extent possible. The

second term (B_2) can be written as $\mu_0^2[w_{1g}(1-w_{1g})+w_{2g}(1-w_{2g})+2w_{1g}w_{2g}]$. Genes with ambiguous probabilities of over- or under-expression contribute to this term. This suggests that excluding differentially expressed genes from normalization can reduce the missing data information, further supporting our proposal that normalization be based solely on equivalently expressed genes. For every gene g, the third term (B_3) vanishes when there is symmetry of differential expression (that is, $w_{1g} = w_{2g}$) or when the differential expression status is known without ambiguity (that is, $w_{0g} = 0$ or 1), again suggesting the need to estimate the differential expression status as accurately as possible.

In summary, the missing data information suggests normalizing the arrays, particularly of cases, using equivalently expressed genes.

3 Simulation

Evaluating the performance of different normalization methods depends upon the scientific question such as detection of differential expression and estimation of the amount of differential expression. Here we consider the problem of detecting differential expression, and evaluate the sensitivity (true positive) and specificity (true negative) rates based on three normalization methods. More specifically, we simulate data as outlined below and analyze them as follows.

- First, the data are analyzed using the proposed method. The posterior probability of association between each gene and the disease status can be obtained from the proposed EM algorithm (quantified by $w_{1g} + w_{2g}$; see Appendix B). The genes are ranked in the decreasing order of this posterior probability, and the sensitivity and specificity are calculated. One hundred datasets are simulated, and sensitivity and specificity are averaged across them. The ROC curve ((Hanley and McNeil 1982)) is then plotted and the area under the curve (AUC) is calculated. Large areas correspond to favorable performance of the proposed method.

- Secondly, the data are analyzed by applying median normalization followed by the two-sample t-test to evaluate differential expression. The genes are ranked based on the increasing order of the resulting p-values. A ROC curve and the corresponding AUC are obtained as outlined above.

- Finally, the data are analyzed by applying no normalization and using the two-sample t-test for each gene. The genes are ranked based on the

DOI: 10.2202/1544-6115.1339

increasing order of the resulting p-values. A ROC curve and its AUC are obtained as outlined above.

The AUCs of the proposed method, median normalization, and no normalization are compared to examine which method provides favorable gene ranking.

In the first set of simulations, data are generated according to Equation 3 for 50 arrays (25 cases and 25 controls), 10000 probesets, 4 probes per probeset. Array effect α_i's are generated from a normal distribution with mean 7 and standard deviation 2. The mean of over-expression μ_o is set to 2 and the mean of under-expression μ_u to -2. The variances of the random effects are set to 0.5^2, 2.5^2, 1^2, and 0.75^2 for measurement error ϵ_{igp}, gene effect β_g, over-expression disease effect γ_{og}, and under-expression disease effect γ_{ug}, respectively.

We simulated data under three scenarios for various patterns of differential expression: (1) many and highly asymmetric differential expression ($\pi = 0.3$ and $\lambda = 0.9$), (2) some and highly asymmetric differential expression ($\pi = 0.1$ and $\lambda = 0.9$), and (3) many and slightly asymmetric differential expression ($\pi = 0.3$ and $\lambda = 0.6$).

In the first set of simulations, array effects are sorted to be higher in controls than cases (that is, array effects is non-randomized). As shown in Figure 2 (left-hand panels), normalization is necessary to detect differentially expressed genes and the proposed normalization out-performs median normalization. Clearly, when the assumption of median normalization does not hold, it might not only hide differentially expressed genes but also lead to non-differentially expressed genes claimed otherwise. The improvement of the proposed normalization depends on the proportion of differential expression (π), the degree of asymmetry of over-/under-expression (λ), and the average size of differential expression (μ_o and μ_u; see results in supplementary Figure 1). These three factors determine the true median of the expression intensity on an array, given that differential expression exists. An example of non-randomization in real data is when all the cases are hybridized together on one batch and all controls are hybridized together on another batch (separately from the cases).

We performed a second set of simulations for the same three differential expression patterns but with array effects randomized between controls and cases. The results are consistent with intuition that, when the array effects are randomized, normalization is not critical. Further, the proposed normalization method slightly improves the detection of differential expressed genes, while median normalization could deteriorate the detection if the differential expression are many and asymmetric (Figure 2, right-hand panels).

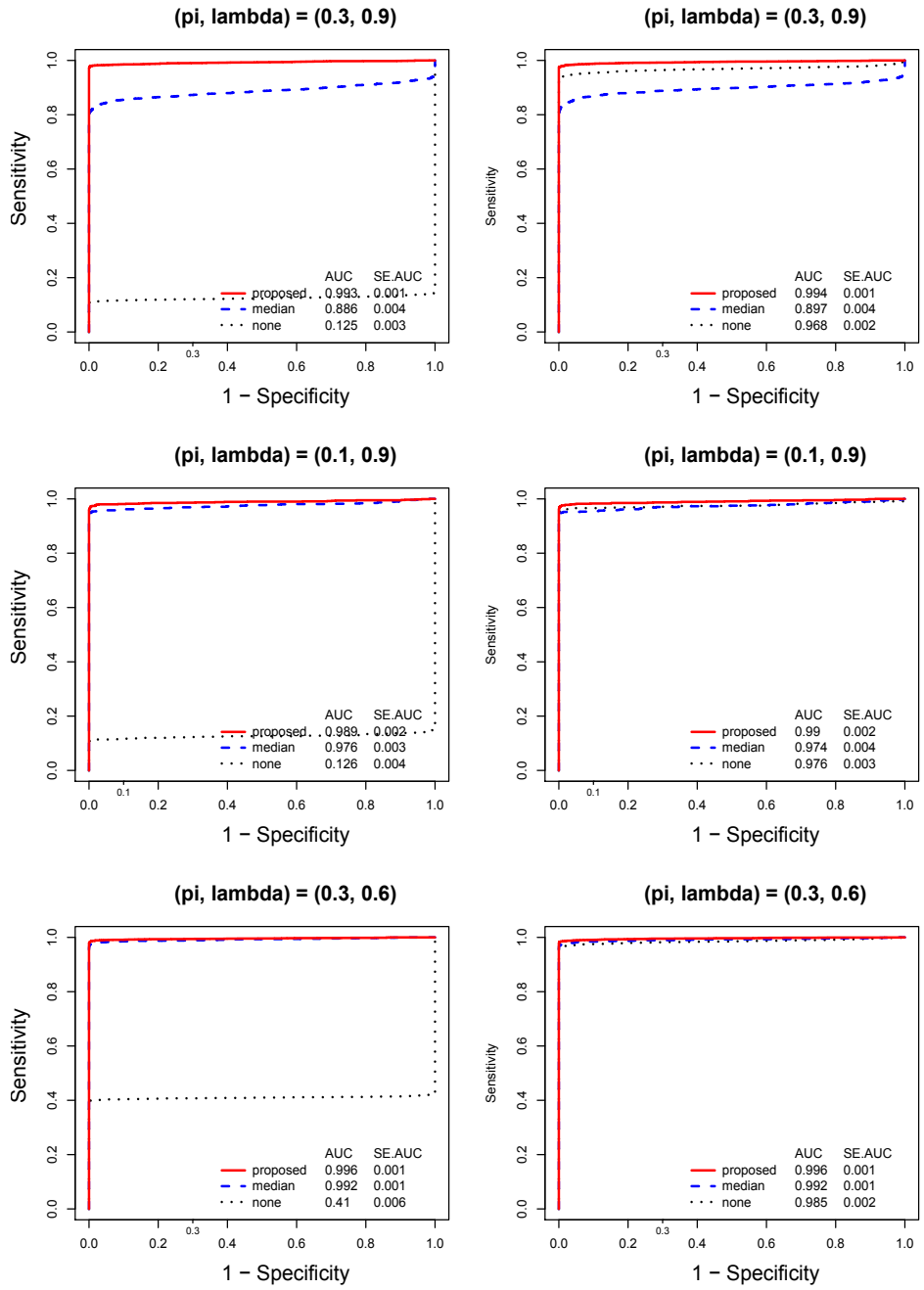

Figure 2: ROC curves for the posterior probability based on the proposed normalization and the two-sample t-test p-values following no or median normalizations. True array effects are non-randomized for the left panels and randomized for the right panels. From top to bottom, (π, λ) are (30%, 90%), (10%, 90%), and (30%, 60%).

DOI: 10.2202/1544-6115.1339

In order to examine whether the above simulation results favor the proposed normalization due to the fact that the data are generated based on the proposed model, we conducted a third set of simulations under the following parametric configurations: (1) the gene effect β_g is generated from a uniform distribution; and (2) the measurement error ϵ_{igp} is generated from a t distribution with degrees of freedom of 3. The results of these simulations are similar to those observed above. The corresponding figures are given as supplementary materials.

Finally, we also conducted additional simulations with $\sigma^2 = 1$ to study the operating characteristics when the measurement error is not very low. The results of these simulations (shown in supplementary materials) are similar to those observed above.

In summary, the proposed method out-performs median normalization when there are many and asymmetric differentially expression, while it performs similarly when there are few differentially expression. Hence, the proposed method is robust to the assumption of few or symmetric differential expression.

4 Data Application

We applied the proposed normalization, median normalization, control-gene normalization, and quantile normalization to a subset of the liposarcoma data, including 8 MRC tumors and 12 normal fat tissues. The arrays were generated as patient samples became available; that is, the tumors and normal tissues were not separately batched and also not purposely randomized.

Figure 3 shows the density plot of the arrays after normalization using each of the four normalization methods. It shows that the normalized data based on the proposed method is most similar to that based on median normalization. Also control-gene normalization does not seem to sufficiently remove the array effects and it results in negative values for most genes. In this particular dataset, the array effects based on the proposed method are slighly smaller than those based on median normalization, with a difference bigger among normal fat arrays than among tumor arrays. In order to evaluate how well the proposed model fits the data, we calculated the predicted expression using the maximum likelihood estimates of the fixed effects and the BLUPs of the random effects (Robinson 1991), and then obtained a QQplot for the predcited expression versus the observed expression among normal fat samples and tumor samples (see supplementary material). Due to the large number of probesets, we show the QQplots for a random set of one tenth of the probesets

on the array (that is, 2228 probesets). The QQplots show that the proposed model fit the data well.

A per-gene two-sample t-test was then applied to identify differentially expressed genes between tumors and normal fat tissues. Figure 4 compares the distribution of the p-values for the four normalization methods. They show that the proposed normalization, median normalization, and quantile normalization provide very similar p-values for this dataset, while no normalization and control-gene median normalization give very different results. In particular, based on the proposed normalization, the proportion of differentially expressed genes is estimated to be $\pi = 32.0\%$; among these genes, the proportion of up-regulated genes is estimated to be $\lambda = 37.0\%$; the mean of over-expression is estimated to be $\mu_o = 0.35$ and that of under-expression to be $\mu_u = -0.05$. These results suggest that there are many and moderately asymmetric differential expression between MRC tumors and normal fat tissues in the liposarcoma data. Further the magnitude of over- and under-expression are different. This corresponds to a scenario that is approximately similar to the simulations represented in the bottom left panel in supplementary Figure 1, which shows that the performances of the proposed method and median normalization are fairly similar and superior to no normalization. In this manner, the data analysis results are consistent with the simulation findings.

There have been limited studies published to date providing a detailed investigation on the genetic basis of liposarcoma. Since there is no gold standard method providing the genes of relevance for liposarcoma, we examined the functional relevance of the genes identified by the proposed method and median normalization in an effort to obtain insights into the practical utility of the two methods. A significance cutoff of p-value=0.00001 was applied to the per-gene p-values to identify significant genes for each of the normalization methods. There are 1893 and 2007 significant genes for median normalization and proposed normalization, respectively. Among them, 1734 genes are overlapping. A total of 159 genes were identified by the proposed method but not by the median normalization. According to the EASE analysis on the functional themes, these 159 unique genes are enriched in signal transduction and cellular process. These functions have been previously implicated in genetic studies of liposarcoma (Gauthier et al. 2003; Chibon et al. 2004; Muller et al. 2007; Guo et al. 2008). A total of 273 unique genes were identified solely by the median normalization method. These genes are enriched in nucleoside metabolism or binding, as suggested in other published studies (Barone et al. 1994). The detailed EASE results can be found in the supplementary materials. These results suggest that the proposed normalization method is able to identify genes of known functional relevance for liposarcoma.

DOI: 10.2202/1544-6115.1339

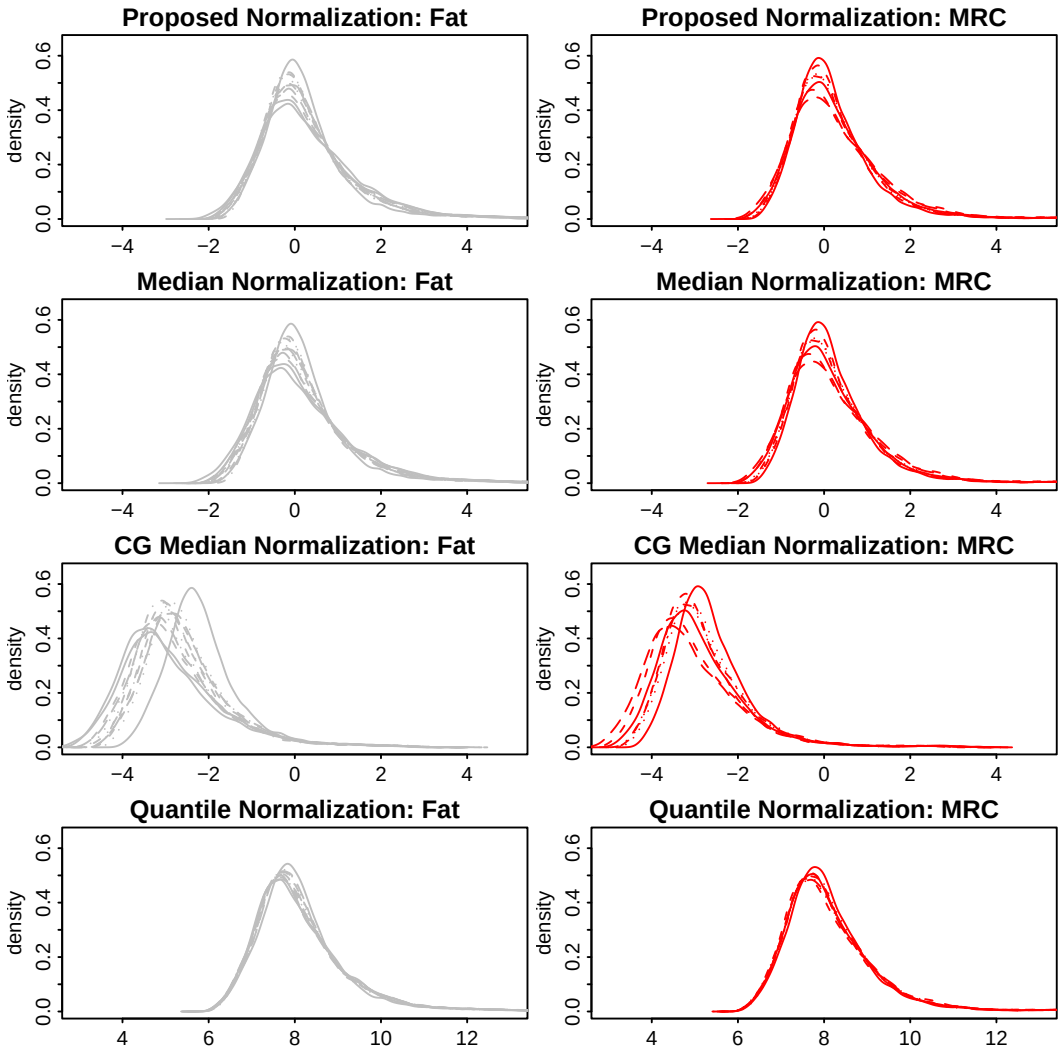

Figure 3: Density plots of normal fat (left panels) and liposarcoma (right panels) arrays after different normalizations (proposed, median, control-gene (CG) median, and quantile normalization).

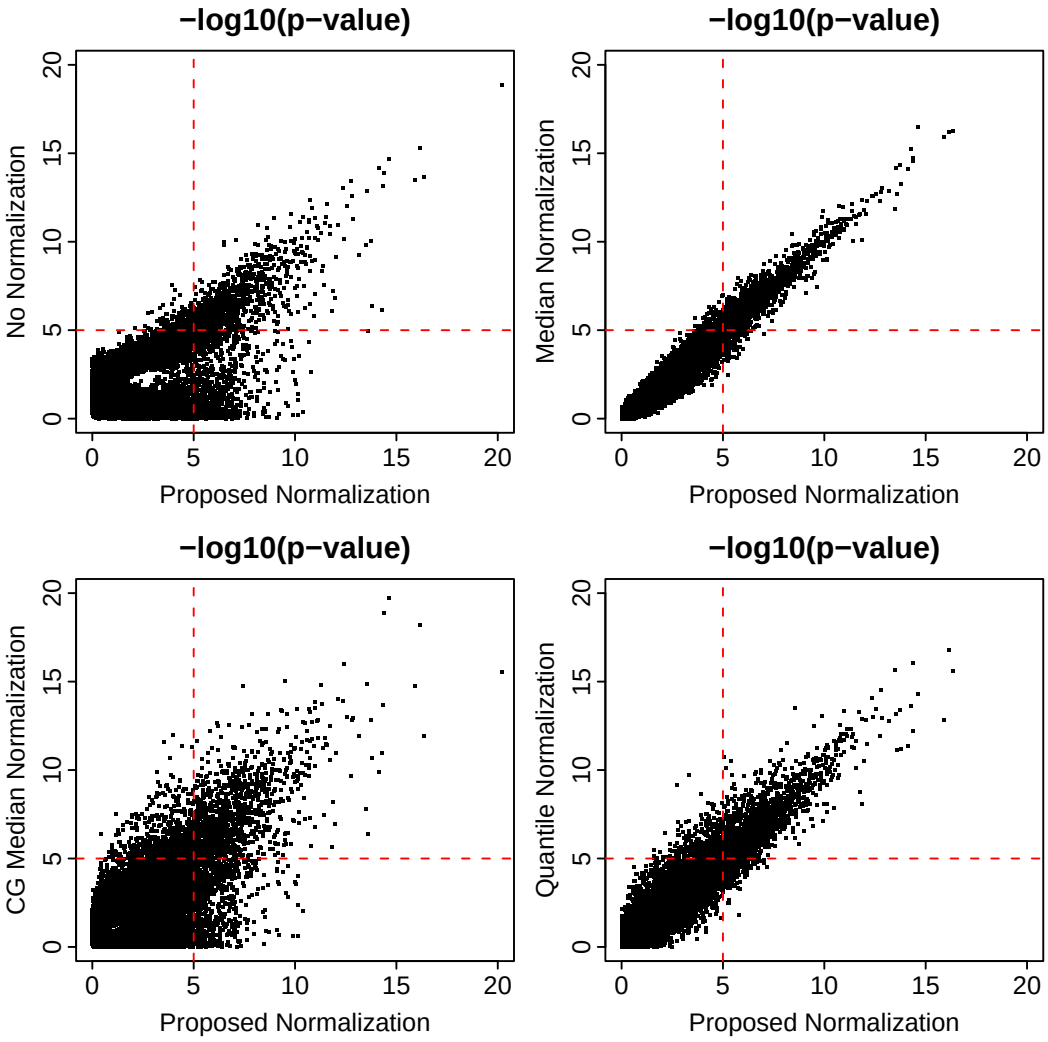

Figure 4: Scatter plots comparing the t-test p-values following proposed normalization with that following no normalization, median normalization, control-gene (CG) median normalization, and quantile normalization. P-values are plotted on the -log10 scale.

DOI: 10.2202/1544-6115.1339

5 Discussion

The essence of both all-gene and some-gene strategies is to identify non-differentially expressed genes with certain assumptions and use their expression for normalization across arrays. All-gene methods use all genes on an array, while some-gene methods define these genes *a priori*. In order to effectively choose the control genes *a priori*, the selection should be based on a randomized experiment with a sufficiently large sample size.

For both all-gene and some-gene strategies, nonlinear estimation can be applied (Yang et al. 2002). Most of them model the array effects as a nonlinear function of intensity levels, for example loess smoothing. Intensity normalization will work reasonably well if, at each level of intensity, the average up- and down-regulation are about equal. However, as the normalization method becomes more flexible, one needs to be aware of the risk of over-normalizing and washing out real biological effects.

Randomization has been shown in our simulation study to be an effective approach to minimize the effect of array differences and should be adopted in practice to the extent possible.

We obtained diagnostic QQplots to examine the goodness of fit for the proposed model when applied to the liposarcoma data. The QQplots provide evidence that the proposed model fit the data well.

Zhao et al (2005) employed a mixture model to identify equivalently expressed genes for normalization. While these authors focused on the normalization of cDNA microarrays using Gamma mixture distributions, our work pertains to oligonucleotide arrays with expressions modeled as Gaussian mixtures.

The current implementation of the proposed method took about 8 hours to normalize this liposarcoma data in a PC with 3.0GHz Pentium 4 processor and 1024 MB memory running Windows 2000. More computational efficient implementation will be explored as part of the future research.

In summary, array normalization is an important component of analyzing microarray data. Several normalization methods have been proposed in the literature, based on simplifying assumptions. We have developed a novel approach that utilizes genes that are not differentially expressed for normalization. Our results suggest that the proposed method has superior sensitivity for identifying differentially expressed genes, relative to median normalization and no normalization, when the arrays are not randomized. As expected, when the arrays are randomized, the sensitivity of the proposed method is comparable to no normalization.

6 Software

The proposed method has been implemented using R. The code is available from the first author (qinl@mskcc.org). This code may also be downloaded from our institutional web page http://www.mskcc.org/mskcc/html/60448.cfm.

7 Appendix

7.1 Appendix A. Parameter Estimation for the Proposed Method

The observed data is y_{igp}'s. The complete data is y_{igp}'s and $\{o_g, u_g\}$'s. The parameters to estimate are α_i's, μ_o, μ_u, τ^2, ψ^2, ξ^2, and σ^2.

In the E-step, the posterior probability for a gene to be equal-, over- and under-expressed are calculated as following. Denote $\boldsymbol{y}_g$ as the vector of expression levels for gene g, $\boldsymbol{x}$ as the vector of disease status corresponding to $\boldsymbol{y}_g$, $\boldsymbol{\Phi}$ as the pdf function of a multivariate normal distribution, $\boldsymbol{\Sigma}$ as a $n \times n$ covariance matrix, $\mathbf{0}$ as a vector of zero, and $\boldsymbol{I}$ as a unit diagonal matrix.

$$\begin{aligned} f(\boldsymbol{y}_g|o_g = 0, u_g = 0) &= \boldsymbol{\Phi}(\boldsymbol{y}_g; \boldsymbol{\mu}_e, \boldsymbol{\Sigma}_e) \\ f(\boldsymbol{y}_g|o_g = 1) &= \boldsymbol{\Phi}(\boldsymbol{y}_g; \boldsymbol{\mu}_o, \boldsymbol{\Sigma}_o) \\ f(\boldsymbol{y}_g|u_g = 1) &= \boldsymbol{\Phi}(\boldsymbol{y}_g; \boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u) \\ \boldsymbol{\mu}_e &= \mathbf{0} \\ \boldsymbol{\mu}_o &= \boldsymbol{x}\mu_o \\ \boldsymbol{\mu}_u &= \boldsymbol{x}\mu_u \\ \boldsymbol{\Sigma}_e &= \boldsymbol{I}\sigma^2 \\ \boldsymbol{\Sigma}_o &= \boldsymbol{I}\sigma^2 + \boldsymbol{x}^T\boldsymbol{x}\psi^2 \\ \boldsymbol{\Sigma}_u &= \boldsymbol{I}\sigma^2 + \boldsymbol{x}^T\boldsymbol{x}\xi^2 \end{aligned}$$

A gene is then assigned to the category that gives the highest pdf: equal-expression ($o_g = 0$ and $u_g = 0$), over-expression ($o_g = 1$ and $u_g = 0$), or under-expression ($o_g = 0$ and $u_g = 1$).

In the M-step, the parameters are estimated by fitting a linear mixed effects model as following.

$$\text{lme}(y' \sim -1 + x * o + x * u, \text{ random} = \sim 1 + x * o + x * u| \text{ group})$$

where $y'_{igp} = y_{igp} - \alpha_i$, *group* indicates the grouping of the observations such that observations for each equal-expressed gene are in a unique group and observations for each differential-expressed gene are in two unique groups, one

DOI: 10.2202/1544-6115.1339

for cases and the other for controls. For example, one configuration of *group* is the following.

$$\begin{aligned}
\text{group}_{igp} &= g, \text{ if } o_g = 0 \text{ and } u_g = 0 \\
\text{group}_{igp} &= g, \text{ if } o_g = 1 \text{ and } x_i = 0 \\
\text{group}_{igp} &= g, \text{ if } u_g = 1 \text{ and } x_i = 0 \\
\text{group}_{igp} &= -g, \text{ if } o_g = 1 \text{ and } x_i = 1 \\
\text{group}_{igp} &= -g, \text{ if } u_g = 1 \text{ and } x_i = 1
\end{aligned} \tag{6}$$

7.2 Appendix B. Information of α_i's

Let θ_{0g}, θ_{1g}, and θ_{2g} denote indicators for equal-, over-, and under-expression of gene g, respectively. Note that for any gene, $\theta_{0g} + \theta_{1g} + \theta_{2g} = 1$.

The complete data log likelihood for array i, given $\{\theta_{0g}, \theta_{1g}, \theta_{2g}\}$'s, can be written as follows. Let $\psi()$ denote the density function for a standard normal distribution.

$$\begin{aligned}
& l_c(\{y_{igp}\}, \alpha_i, \mu_o, \mu_u, \tau^2, \sigma^2, \psi^2, \xi^2) \\
= & \sum_g \sum_p \{\theta_{0g} \log \phi(\frac{y_{igp} - \alpha_i}{\sqrt{\tau^2 + \sigma^2}}) \\
+ & \theta_{1g} \log \phi(\frac{y_{igp} - \alpha_i - x_i\mu_o}{\sqrt{\tau^2 + \sigma^2 + x_i\psi^2}}) \\
+ & \theta_{2g} \log \phi(\frac{y_{igp} - \alpha_i - x_i\mu_u}{\sqrt{\tau^2 + \sigma^2 + x_i\xi^2}})\}
\end{aligned}$$

The complete data information is given by $-E(\frac{\partial^2 l_c}{\partial \alpha_i^2})$, that is,

$$I_c = GP\{\frac{\pi}{\tau^2 + \sigma^2} + \frac{(1-\pi)\lambda}{\tau^2 + \sigma^2 + x_i\psi^2} + \frac{(1-\pi)(1-\lambda)}{\tau^2 + \sigma^2 + x_i\xi^2}\}$$

where G and P are the total number of genes and the total number of probes per gene on the array, respectively.

The missing data are $\{\theta_{0g}, \theta_{1g}, \theta_{2g}\}'s$. The log likelihood of the missing data given the observed data can be written as the following.

$$\begin{aligned}
& l_m(\{y_{igp}\}, \alpha_i, \mu_o, \mu_u, \tau^2, \sigma^2, \psi^2, \xi^2) \\
= & \sum_g \theta_{0g} \log(w_{0g}) + \theta_{1g} \log(w_{1g}) + \theta_{2g} \log(w_{2g})
\end{aligned}$$

where $w_{kg} = P(\theta_{kg} = 1 | \{y_{igp}\}, \alpha_i, \mu_o, \mu_u, \tau^2, \sigma^2, \psi^2, \xi^2)$ for $k = 0, 1, 2$.

The missing data information corresponding to α_i is given by $-E(\frac{\partial^2 l_m}{\partial \alpha_i^2})$. It follows from straightforward algebra that

$$I_m = \sum_g \sum_k \frac{1}{w_{0g}} (\frac{\partial w_{0g}}{\partial \alpha_i})^2$$

For the sake of simplicity, we provide the formula for the special case where $\mu_u = -\mu_o$ and $\tau^2 + \sigma^2 = 1$. Note that w_{0g}, w_{1g}, and w_{2g} are posterior probabilities of equal-, over-, and under-expression, respectively; hence we can write the following equations using simple algebraic expansions. Denote $\bar{y}_{ig.} = \frac{1}{P} \sum_{p=1}^{P} y_{igp}$.

$$\begin{aligned}
\frac{\partial w_{0g}}{\partial \alpha_i} &= P w_{0g} \frac{x_i}{1+x_i} [(\bar{y}_{ig.} - \alpha_i)(1 - w_{0g}) + \mu_o (w_{1g} - w_{2g})] \\
\frac{\partial w_{1g}}{\partial \alpha_i} &= -P w_{1g} \frac{x_i}{1+x_i} [(\bar{y}_{ig.} - \alpha_i) w_{0g} + \mu_o (1 - w_{1g} + w_{2g})] \\
\frac{\partial w_{2g}}{\partial \alpha_i} &= P w_{2g} \frac{x_i}{1+x_i} [-(\bar{y}_{ig.} - \alpha_i) w_{0g} + \mu_o (1 + w_{1g} - w_{2g})]
\end{aligned}$$

Therefore, the missing data information for α_i's is the following.

$$\begin{aligned}
I_m &= P^2 \frac{x_i}{1+x_i} \sum_g \{B_1 + B_2 + B_3\} \\
B_1 &= (\bar{y}_{ig.} - \alpha_i)^2 w_{0g} (1 - w_{0g}) \\
B_2 &= \mu_o^2 [(1 - w_{0g}) - (w_{1g} - w_{2g})^2] \\
B_3 &= 2(\bar{y}_{ig.} - \alpha_i) \mu_o w_{0g} (w_{1g} - w_{2g})
\end{aligned}$$

DOI: 10.2202/1544-6115.1339

References

Atkinson, A. (1985). *Plots, transformations and regression: An introduction to graphical methods of diagnostic regression analysis*. Oxford University Press.

Barone, M., A. Crozat, A. Tabaee, L. Philipson, and D. Ron (1994). CHOP (GADD153) and its oncogenic variant, TLS-CHOP, have opposing effects on the induction of G1/S arrest. *Genes and Development 8*, 453–464.

Bolstad, B., R. Irizarry, M. Astrand, and T. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics 19*, 185–193.

Carroll, R. (1980). A robust method for testing transformation to achieve approximate normality. *Journal of the Royal Statistical Society, Series B, Methodological 42*, 71–78.

Chibon, F., O. Mariani, J. Derre, A. Mairal, J. Coindre, L. Guillou, and X. Sastre (2004). ASK1 (MAP3K5) as a potential therapeutic target in malignant fibrous histiocytomas with 12q14-q15 and 6q23 amplifications. *Genes, Chromosomes and Cancer 40*, 32–37.

Durbin, B. and D. Rocke (2004). Variance-stabilizing transformations for two-color microarrays. *Bioinformatics 20*, 660–667.

Gauthier, A., G. Vassiliou, F. Benoist, and R. McPherson (2003). Adipocyte low density lipoprotein receptor-related protein gene expression and function is regulated by peroxisome proliferator-activated receptor gamma. *The Journal of Biological Chemitry 278*, 11945–11953.

Guo, Y., J. Xie, E. Rubin, Y. Tang, F. Lin, X. Zi, and B. Huang (2008). Frzb, a secreted Wnt antagonist, decreases growth and invasiveness of fibrosarcoma cells associated with inhibition of Met signaling. *Cancer Research 68*, 3350–3360.

Hanley, J. and B. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*, 29–36.

Hastie, T., R. Tibshirani, and J. H. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hoerl, A. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress 58*, 54–59.

Huber, W., A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron (2003). Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology 2*, Article 3.

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

Kendziorski, C., M. Newton, H. Lan, and M. Gould (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine 22*, 3899–3914.

LaTulippe, E., J. Satagopan, A. Smith, H. Scher, P. Scardino, V. Reuter, and W. Gerald (2002). Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Research 62*(15), 4499–4506.

Li, C. and W. Wong (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA 98*, 31–36.

Lipshutz, R. J., S. P. Fodor, T. R. Gingeras, and D. J. Lockhart (1999). High density synthetic oligonucleotide arrays. *Nature Genetics Supplement 21*, 20–24.

Louis, T. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society, Series B, Methodological 44*, 226–233.

Muller, C., E. Paulsen, P. Noordhuis, F. Pedeutour, G. Saeter, and O. Myklebost (2007). Potential for treatment of liposarcomas with the mdm2 antagonist nutlin-3a. *International Journal of Cancer 121*, 199–205.

Nguyen, D. V., A. B. Arpat, N. Wang, and R. J. Carroll (2002). DNA microarray experiments: biological and technological aspects. *Biometrics*, 701–717.

Perou, C. M., T. Srlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, ystein Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lnning, A.-L. Brresen-Dale, P. O. Brown, and D. Botstein (2000). Molecular portraits of human breast tumours. *Nature 406*, 747–752.

Purdom, E. and S. P. Holmes (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology 4*, 16.

DOI: 10.2202/1544-6115.1339

Reilly, C., C. Wang, and M. Rutherford (2003). A method for normalizing microarrays using genes that are not differentially expressed. *Journal of the American Statistical Association 98*, 868–878.

Robinson, G. (1991). That blup is a good thing: the estimation of random effects. *Statistical Science 6*, 15–32.

Rocke, D. and B. Durbin (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology 8*, 557–569.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science 270*, 467–470.

Singer, S., N. Socci, G. Ambrosini, E. Sambol, P. Decarolis, Y. Wu, R. O'Connor, R. Maki, A. Viale, C. Sander, G. Schwartz, and C. Antonescu (2007). Gene expression profiling of liposarcoma identifies distinct biological types/subtypes and potential therapeutic targets in well-differentiated and dedifferentiated liposarcoma. *Cancer Research 67*(14), 6626–6636.

Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher (1998). Comprehensive identification of cell cycle–regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell 9*, 3273–3297.

Thellin, O., W. Zorzi, B. Lakaye, B. DeBorman, B. Coumans, G. Hennen, T. Grisar, A. Igout, and E. Heinen (1999). Housekeeping genes as internal standards: use and limits. *Journal of Biotechnology 75*, 291–295.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological 58*, 267–288.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. Speed (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research 30*(4), e15.

Zhao, Y., M.-C. Li, and R. Simon (2005). An adaptive method for cDNA microarray normalization. *BMC Bioinformatics 6*, 28.