The Korean Communications in Statistics
Vol. 13 No. 3, 2006, pp.701-718

Statistical Method for Implementing the Experimenter Effect in the Analysis of Gene Expression Data

Inyoung Kim[1], Sun Young Rha[2] and Byung Soo Kim[3]

Abstract

In cancer microarray experiments, the experimenter or patient which is nested in each experimenter often shows quite heterogeneous error variability, which should be estimated for identifying a source of variation. Our study describes a Bayesian method which utilizes clinical information for identifying a set of DE genes for the class of subtypes as well as assesses and examines the experimenter effect and patient effect which is nested in each experimenter as a source of variation. We propose a Bayesian multilevel mixed effect model based on analysis of covariance (ANACOVA). The Bayesian multilevel mixed effect model is a combination of the multilevel mixed effect model and the Bayesian hierarchical model, which provides a flexible way of defining a suitable correlation structure among genes.

Keywords : Analysis of covariance; Bayesian hierarchical model; cDNA micro-array; multilevel mixed effect model.

1. Introduction

DNA microarray technology is a major tool for high throughput analysis that simultaneously measures the expression levels of thousands of genes in biological sciences, in general and in cancer studies, in particular. The expression levels are of ten measured for each gene and each patient with various tumor subtypes under different experimenters. One usually normalizes the data in order to remove the systematic variation of various sources, which includes difference in labeling efficiency between two fluorescent dyes, spatial effect such as print-tip effect, array effect and etc (Yang *et al.*, 2002). However, so far as we know, there is no report on the experimenter effect or patient effect which is nested in each experimenter as a potential source of variation. The aim of this study is to identify the experimenter effect or patient effect which is nested in each

[1] Postdoctoral associate, Department of Epidemiology and Public Health, School of Me-dicine, Yale university, New Haven, CT 06520, USA.

[2] Professor, Brain Korea 21 Project for Medical Science, College of Medicine, Yonsei University, Seoul, 120-752, Korea.

[3] Professor, Department of Applied Statistics, Yonsei University, Seoul, 120-749, Korea.
Correspondence : bskim@yonsei.ac.kr

experimenter as a new source of variation and to develop a proper statistical model for its identification based on a cDNA microarray of colorectal cancers.

One of the major interests of cancer microarray experiments is to identify a set of DE genes among various tumor subtypes. An analysis of variance (ANOVA) is a natural method of examining the data. The ANOVA modeling approaches have been proposed both for cDNA microarray data (Kerr *et al.*, 2000; Kerr and Churchill, 2001; Wolfinger *et al.*, 2001) and for oligonucleotide data (Chu *et al.*, 2002; Hsieh *et al.*, 2003). The ANACOVA model extends ANOVA by adding covariates in the model to adjust potential confounding factors among subtypes.

A Bayesian approach was developed to simultaneously analyze a large number of genes and to identify DE genes between two comparison groups, e.g. normal and cancer groups, by Ibrahim *et al.* (2002). Townsend and Hartl (2002) also proposed Bayesian framework to analyze normalized microarray data acquired by any replicated experimental design in which any number of treatments were studied using a continuous chain of comparisons. Tadesse *et al.* (2003) formulated a hierarchical Bayesian model, which provided a flexible way of defining a suitable correlation structure among genes. Tadesse *et al.* (2003) modeled gene expression measures as the censored data accounting for undetected or unreliable transcripts by the quantification limits of technology. However these approaches considered a patient or an experimenter as a replication and did not account for heterogeneous error components on different patients or different experimenters in microarray data.

The experimenter or patient effect which is nested in each experimenter often shows quite heterogeneous error variability, which should be estimated for identifying a source of variation. In the cDNA microarray experiment of colorectal cancer we observe the possibility that the experimenter effect is a major source of variation. The cluster analysis indicates that the experimenter effect still exists even after the normalization. These results are given in <Section 4>. Clinical information of patients has been used to select a set of DE genes between tumors in several cancer studies including the breast cancer (van de Vijiver *et al.*, 2002) and the lymphoma (Rosenwald *et al.*, 2002). We have three clinical variables of interest, namely the location (colon, rectum), the CEA value and the stage (B, C ,D) of colorectal cancer. CEA stands for carcinoembryonic antigen which clinicians use to monitor the progress of the colorectal cancer after the initial treatment. The threshold value of CEA is usually set to be 5. Each of these three clinical variables defines a class of subtypes. The primary interest in the subtype analysis is to detect a set of DE genes between subtypes of a given clinical variable and to validate the chosen set of DE genes for the classification. For example, we are interested in identifying a set of genes that discriminate the colon cancer from the rectal cancer. For the stage variable, we are interested in the pairwise comparison. e.g. B versus C and so on.

We develop a new Bayesian method which utilizes clinical information for identifying a set of DE genes for the class of subtypes as well as assesses and examines the experimenter effect or patient effect which is nested in each experimenter as a source of variation. We propose a Bayesian multilevel mixed effect model based on ANACOVA. The Bayesian multilevel mixed effect model is a combination of the multilevel mixed effect model and the Bayesian hierarchical model, which provides a flexible way of defining a suitable correlation structure among genes. We consider gene, subtype and interaction between gene and subtype as fixed effects. Unlike other Bayesian approaches (Ibrahim *et al.*, 2002; Townsend and Hartl, 2002; Tadesse *et al.*, 2003), which consider a patient or an experimenter as a replication, we treat both the experimenter and the patient which is nested in the experimenter as random effects. The clinical variables are used as covariates.

The paper is organized as follows. In section 2 we describe the material of microarray experiment and microarray data pre-processing. We also discuss how the interaction term becomes a key parameter for identifying a set of DE genes for the class of subtypes. In section 3 we describe a Bayesian multilevel mixed effect model. In section 4 we apply the Bayesian approach to the cDNA microarray experiments of the 68 colorectal cancer patients. We observed that the experimenter effect was a major source of the variation but the patient which was nested in the experimenter was not. We compared two approaches of gene selection; one is based on the Bayesian multilevel mixed effect model and the other is based on the permutation test with the t-statistic. The Bayesian multilevel mixed effect model could make the mild improvement over the permutation test in prediction rate. We believe that this improvement is due to reducing error variability by explicitly incorporating in the model the experimenter effect and using clinical information such as CEA and stages. Section 5 contains concluding remarks

2. Materials and Data

Cancer tissues were obtained during the surgical operations from each of 67 colorectal cancer patients at Severance Hospital, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul, Korea, from May to December 2002. We conducted a cDNA microarray experiment using a common reference design with cDNA microarrays containing 17,000 human genes. We pooled eleven cancer cell lines and used it for the common reference. These eleven cancer cell lines are as follows: AGS, MDA-MB-231, HCT-116, SH-Hep-1, A549, HL-60, MOLT-4, HeLa, HT-1080, Caki-2 and U87MG (American Type Culture Collection). The fresh specimens of cancer tissues obtained from colorectal cancer patients during surgery were snap-frozen in liquid nitrogen right after the resection and stored at -70°C until required. After total RNAs were extracted from fresh frozen tissues, 50g of purified RNAs were labeled and hybridized to cDNA microarrays based on the protocol established in Cancer Metastasis Research Center (Yonsei University, Korea)(Park *et al.*, 2004).

Both a within-print tip group intensity dependent and scale adjustment normalizations between arrays were used to normalize the log intensity ratio, $M = \log_2 \frac{R}{G}$, for the evaluation of the relative intensity, where R and G represent the cy5 and cy3 fluorescent intensities, respectively. The genes whose signals were missing in more than 20% of the specimens were removed from the analysis. The missing values were estimated using the 10-nearest neighbor method. Imputation methods other than the KNN method were developed recently (Kim H. *et al.*, 2005; Jörnsten *et al.*, 2005; Scheel *et al.*, 2005). The values for multiple spots were also averaged. Finally, a data set represented by a 12850×68 matrix was obtained for the analysis, where 12850 represents the number of genes and 68 stands for the number of tumor arrays. <Table 1> shows the numbers of tumors in three classes of subtypes; the location (colon versus rectum), CEA level (CEA≤5 versus CEA>5), and the stage (B, C and D).

<Table 1> The number of tumors in each subtype under each experimenter

	Location		Stage			CEA	
Experimenter	Colon	Rectum	B	C	D	≤5	>5
1	22	20	18	16	6	19	23
2	13	13	17	8	3	20	6

3. Model and Method

3.1 Multilevel mixed effect model

A multilevel mixed effect model to describe y_{ijkl} which represents the log intensity ratio for the ith gene of the jth patient in the kth group within the lth experimenter is

$$y_{ijkl}=\mu_{ik}+\sum_{m=1}^{M}\lambda_{km}^{'}x_{jkl}^{m}+\tau_l+s_{j(l)}+\epsilon_{ijkl},\ \tau_l\sim N(0,\sigma_\tau^2),\ s_{j(l)}\sim N(0,\sigma_s^2),\ \epsilon_{ijkl}\sim N(0,\sigma_\epsilon^2), \qquad (3.1)$$

where μ_{ik} is the mean expression level for the ith gene in the kth group. We assume that μ_{ik} can be expressed as a linear combination of main gene effects α_i, group effects β_k and the interaction effects γ_{ik}, i.e. $\mu_{ik}=\alpha_i+\beta_k+\gamma_{ik}$. The τ_l is the experimenter random effect and $s_{j(l)}$ is the jth patient random effect which is nested in the lth experimenter effect. That is, the experimenter is the first level group and the patient is the second level group which is nested in the first level group. The x_{jkl}^{m} is a Q-vector of the mth power of the covariates for jth patient of kth group in lth experimenter, where Q is the number of covariates. The λ_{km} is the corresponding Q-vector of regression parameters. The experimenter random effects τ_l are assumed to be independent of the different l, the patient random effects nested in the experimenter random effect $s_{j(l)}$ are assumed to be independent for different j and l and to be independent of the τ_l. Within-group error terms ϵ_{ijkl} are assumed to be independent for different i, j, k and l and to be independent of the random effects.

3.2 Bayesian hierarchical method

We develop a Bayesian hierarchical method to fit the model (3.1). Let $\theta=(\alpha,\beta,\gamma,\lambda,\sigma_s^2,\sigma_\tau^{2}{}'\sigma^2)$ denote the vector of the parameters, where $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_p)$, $\beta=(\beta_1,\ldots,\beta_K)$, $\gamma=(\gamma_{11},\ldots,\gamma_{pK})$ and $\lambda_m=(\lambda_{11m},\ldots,\lambda_{pQm})$. Let $D=(y,\tau_l,s_{j(l)})$ denote the data, where y is a $p\times n$ $(n=n_{11}+\ldots+n_{KL})$ matrix with elements y_{ijkl}. Let $f_\tau(\tau_l)$ and $f_s(s_{j(l)})$ be the normal density functions of τ_l and $s_{j(l)}$, respectively. The likelihood function for θ is then given by

$$L(\theta|D)=\prod_{l=1}^{L}\prod_{k=1}^{K}\prod_{j=1}^{n_{kl}}\prod_{i=1}^{p}\frac{1}{\sqrt{2\pi\sigma^2}}$$
$$\times\exp\left[-\frac{1}{2}\left(\frac{y_{ijkl}-(\alpha_i+\beta_k+\gamma_{ik}+\sum_{m=1}^{M}\lambda_{im}^{'}x_{jkl}^{m}+s_{j(l)}+\tau_l)}{\sigma}\right)^2\right]\times f_\tau(\tau_l)f_s(s_{j(l)}).$$

We specify flexible priors that define the suitable correlation structures between genes (α_i and $\alpha_{i'}$) or between group categories (γ_{ik} and $\gamma_{ik'}$) and allow an efficient calculation. These parameters have the following prior distributions:

$$\alpha_i|(\alpha_0,\sigma_\alpha^2)\sim N(\alpha_0,\sigma_\alpha^2),\qquad \alpha_0|(a_0,v_\alpha^2)\sim N(a_0,v_\alpha^2),\qquad \beta_k|(\beta_{0k},\sigma_\beta^2)\sim N(\beta_{0k},\sigma_\beta^2),$$
$$\beta_{0k}|(b_{0k},v_\beta^2)\sim N(b_{0k},v_\beta^2),\qquad \gamma_{ik}|(\gamma_{i0},\sigma_\gamma^2)\sim N(\gamma_{i0},\sigma_\gamma^2),\qquad \gamma_{i0}|(d_{i0},v_\gamma^2)\sim N(d_{i0},v_\gamma^2),$$
$$\lambda_{iqm}|(\lambda_{i0},\sigma_\lambda^2)\sim N(\lambda_{i0},\sigma_\lambda^2),\ \lambda_{i0}|(e_{i0},v_\lambda^2)\sim N(e_{i0},v_\lambda^2),\ s_{j(l)}|\sigma_s^2\sim N(0,\sigma_s^2),\ \sigma_s^2\sim IG(A_s,B_s),$$

$\tau_l|\sigma^2_\tau \sim N(0,\sigma^2_\tau)$, $\sigma^2_\tau \sim IG(A_\tau,B_\tau)$ and $\sigma^2 \sim IG(A,B)$, where $\sigma^2 \sim IG(A,B)$ denote a the inverse-gamma distribution with density

$$\frac{B^A}{\Gamma(A)}(\sigma^2)^{-(A+1)}\exp(\frac{-B}{\sigma^2}).$$

Since the conjugate prior families for $(\alpha,\beta,\gamma,\lambda_m)$ and $(\sigma^2_\tau,\sigma^2_s,\sigma^2)$ consist of normal distributions and inverse gamma distributions, respectively, the posterior samples can then easily be drawn using Gibb sampling or Markov chain Monte Carlo (MCMC) sampling. The derivation of the posterior distributions is given in Appendix. We fit the model (3.1) with $M=1$, $L=2$, $K=2$, $p=12850$. The n_{kl}'s are numbers of tumors in each class of subtypes two comparison subtypes under each experimenter. The hyperprior variances, σ^2_α and σ^2_γ control the strength of the prior correlation among the gene effects and among the interaction effects, respectively. The prior correlations are given by

$$Corr(\alpha_i,\alpha_{i'}|a_0,\sigma^2_\alpha,v^2_\alpha) = \frac{v^2_\alpha}{\sigma^2_\alpha+v^2_\alpha},\ Corr(\gamma_{ik},\gamma_{ik'}|d_{i0},\sigma^2_\gamma,v^2_\gamma) = \frac{v^2_\gamma}{\sigma^2_\gamma+v^2_\gamma}.$$

For noninformative priors we set σ^2_α, σ^2_β and σ^2_γ to 100. The hyperprior means are set to 0 and the hyperprior variances v^2_α and v^2_γ are chosen to yield prior correlations of order 10^{-2}. A burn-in of 10,000 and a main run of 10,000 were taken.

3.3 Gene selection criterion

The most interesting question is that of which genes are differentially expressed in each class of subtypes. The key parameter of interest is a contrast of interaction effects which is the first two terms in equation (3.3.1), as was noted in Kerr and Churchill (2001). The parameter for identifying a set of DE genes between group categories k and k' in each class of subtypes could be formulated as

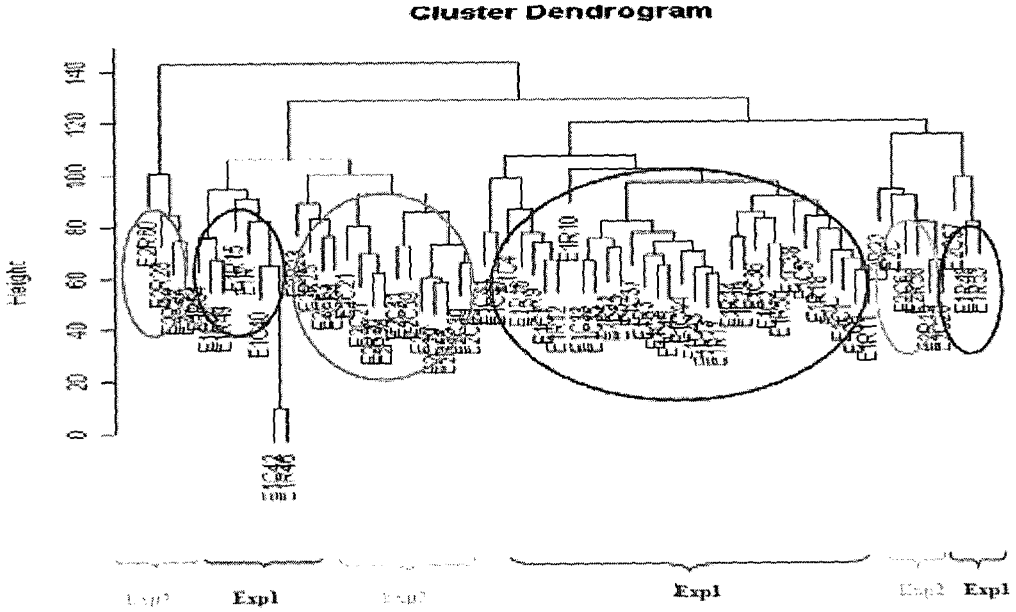
$$\gamma_{ik'} - \gamma_{ik} - (\gamma_{ck'} - \gamma_{ck}), \qquad (3.3.1)$$

where the gene c has a fairly constant expression across tissue types to ensure that $\gamma_{ck'} - \gamma_{ck}$ is close to zero. The same gene c will be used as a reference against which all genes are compared. We simply make the criterion $Z_i = \gamma_{ik'} - \gamma_{ik}$, following the line of Tadesse *et al.* (2003), who showed that the criterion was quite robust to the different reference genes.

4. Results

4.1 Hierarchical clustering analysis using colorectal cancer data

Our interest lies in identifying DE genes in each class of subtypes, e.g., colon versus rectal cancers, CEA≥5 versus CEA<5, stages B versus C, stages B versus D, and stage C versus D. The number of tumors in each subtype, the location (colon versus rectum), CEA level (CEA≥5 versus CEA<5) and the stage (B, C, and D), are given in <Table 1>. Kim B.S. *et al.* (2005) noted that in contrast to detecting an overwhelming number of DE genes between normal and tumor tissues they failed to detect a significant number of genes in subtypes analysis.



As an initial step, we employed hierarchical clustering analysis for various class of subtypes, colon versus rectal cancers, CEA≥5 versus CEA< 5, stages B versus C, stages B versus D, and stages C versus D. We observed that the experimenter effect still existed even after normalizations as is shown in <figure 1> for the case of colon versus rectal cancers. For other classes of subtypes we also noticed similar observations through cluster dendrograms (data not shown). We applied the Bayesian multilevel mixed effect model (3.1) to test the significance of the experimenter effect and to incorporate clinical information for identifying a set of DE genes between two comparison subtypes.

<Figure 1> Cluster dendrogram of colon versus rectum tumors. ''E1C'' means colon tumor by experimenter 1 (Exp1) and ''E2R'' means rectum tumor by experimenter 2 (Exp2).

4.2 Gene selection based on the Bayesian multilevel mixed effect model

We use CEA as a continuous variable and stage as a categorical variable. To detect DE genes between colon and rectal cancers we utilize CEA and stage as covariates in the model (3.1). The CEA was used as a covariate for detecting DE genes among stages. The stage was employed as a covariate when identifying DE genes between CEA$\geq$5 and CEA<5.

To fit the Bayesian multilevel mixed effect model we set σ_α^2, σ_β^2 and σ_γ^2 to 100 for noninformative prior. The hyperprior means were set to 0 and the hyperprior variances v_α^2 and v_γ^2 were chosen to yield prior correlations of order 10^{-2}. After a burn-in of 10,000 iterations, a main run of 10,000 was used. We estimated the parameters by taking the averages of MCMC samples. The estimates and intervals of σ_T, σ_s and σ_ϵ are given in <Table 2>.

<Table 2> The estimates of σ_T, σ_s, and σ_ϵ in Bayesian multilevel mixed effect model with M=1

Class of subtypes	σ_T	95% CI of σ_T	σ_s	95% CI of σ_s	σ_ϵ	95% CI of σ_ϵ
Colon vs Rectal cancers	0.43	[0.406,0.556]	0.05	[0.021,0.102]	0.63	[0.625,0.656]
CEA$\geq$5 vs <5	0.04	[0.019,0.185]	0.07	[0.051,0.105]	0.55	[0.544,0.571]
Stages B vs C	0.02	[0.002,0.146]	0.11	[0.090,0.147]	0.47	[0.459,0.483]
Stages B vs D	0.03	[0.004,0.158]	0.06	[0.044,0.108]	0.51	[0.495,0.525]
Stages C vs D	0.03	[0.006,0.159]	0.09	[0.066,0.136]	0.45	[0.436,0.468]

We noticed that the estimate of σ_T for colon and rectal cancers was much larger than estimates of σ_T for other classes of subtypes, namely, CEA$\geq$5 versus CEA<5, stages B versus C, stages B versus D, and stages C versus D. We also observed that the estimates of σ_s for all classes of subtypes were small. The results imply that the experimenter effect should be incorporated for the analysis of the data. We also note from the widths of the 95% confidence intervals in <Table 2> that σ is estimated relatively precisely, whereas σ_γ and σ_s vary.

The top 10 genes selected based on the proportion of $\frac{Z_i^S}{SE(Z_i)}[-z_{\frac{r}{2}}, z_{\frac{r}{2}}]$, $S=1,\ldots,10{,}000$ and the highest posterior density (HPD) intervals for each class of subtypes are given in <Tables 3-7>, where Z_i^S is the Sth MCMC samples of Z_i, $SE(Z_i)$ is the standard error of Z_i, and $z_{\frac{r}{2}}$ is the $\frac{r}{2}$th upper quantile of standard normal distribution. Genes will be chosen if the 100(1-r)% HPD intervals for Z_i do not contain 0, where r is arbitrarily set by the investigator. Since we could not select any gene if r was larger than 10^{-4} we set r to 10^{-4}. We identified two DE genes and three DE genes between colon versus rectal cancers and between CEA$\geq$5 versus CEA>5, respectively. We found three DE genes for stages B versus C, detected three DE genes for stages B versus D, and selected five DE genes for stages C versus D. With $r = 10^{-3}$, we noted that the fourteen DE genes and twelves DE genes were identified for colon versus rectal cancers and CEA$\geq$5 versus CEA>5, respectively. We also detected three DE genes, twenty-two DE genes, and eight DE genes between stages B and C, between stages B and D, and between stages C and D, respectively.

Furthermore we used two sample t-statistic with unequal variances. We calculated the family-wise error rate (FWER) using Dudoit *et al.*'s maxT procedure (2002), the false discovery rate (FDR) using Tusher *et al.*'s SAM procedure (2001), and positive false discovery rate (pFDR) using Story's Bayesian approach (2002). We could detect 36 DE genes for CEA$\geq$5 vs CEA<5 with FDR=0.31, and 9 DE genes between stages B and D with FDR=0.33. We failed to detect DE genes for colon versus rectal cancers. We ranked genes based on absolute values of the t-statistic. We noticed that all ten genes out of 10 genes were in the top 50 gene list selected from a permutation test with a t-statistic for colon versus rectal tumors. For CEA$\geq$5 versus CEA>5, there were six out of 10 genes in the top 50 gene list selected from a permutation test with a t-statistic. We also observed that all ten genes, three genes, four genes out of 10 genes were in the top 50 gene list between stages B versus C, stages B versus D, and stages C versus D, respectively.

<Table 3> Top 10 genes, which were selected by Bayesian multilevel linear mixed effect model, that best discriminate between colon and rectal cancers, and the gene ontology of top 10 genes. Z_i^S is the Sth MCMC samples of Z_i, $S = 1,\ldots,10000$. $z_{\frac{10^{-1}}{2}}$ is the $\frac{10^{-4}}{2}$th upper quantile of standard normal distribution. "Unknown" means that the gene name or ontology is unknown.

Subtype	Gene Id	Gene name	Z_i	Prop. of $\frac{Z_i^S}{SE(Z_i)} \notin (-Z_{\frac{10^{-4}}{2}}, Z_{\frac{10^{-4}}{2}})$	$(1-10^{-4})$% HPD	Rank of t-stat
Colon vs Rectum	AA431203	DnaJ (Hsp40) homolog subfamily B, member 6	0.19	2×10^{-4}	[0.012,0.358]	30
	AI890849	Aldehyde dehydrogenase 2 family (mitochondrial)	-0.22	2×10^{-4}	[-0.423,-0.010]	5
	AI432671	Unknown	0.13	4×10^{-4}	[0.002,0.264]	2
	AA434102	Lectin, galactoside-binding soluble, 9 (galectin 9)	-0.28	5×10^{-4}	[-0.553,0.002]	12
	AI262115	Polymerase (RNA) II (DNA directed) polypeptide A, 220kDa	-0.42	5×10^{-4}	[-0.843,0.003]	6
	AA485353	Lectin, galactoside-binding, soluble, 3 binding protein	-0.21	6×10^{-4}	[-0.431,0.003]	4
	R87642	LOC440135	0.15	7×10^{-4}	[-0.004,0.302]	14
	AI475738	Likely ortholog of	-0.16	7×10^{-4}	[-0.333,0.005]	1
	AI017808	Hydroxyacid oxidase (long chain)	-0.13	7×10^{-4}	[-0.268,0.004]	3
	AI251639	Unknown	0.15	8×10^{-4}	[-0.006,0.310]	16

<Table 4> Top 10 genes, which were selected by Bayesian multilevel linear mixed effect model, that best discriminate between CEA$\geq$5 and CEA<5, and the gene ontology of top 10 genes. Z_i^S is the Sth MCMC samples of Z_i, $S=1,\ldots,10000$. $z_{\frac{10^{-4}}{2}}$ is the $\frac{10^{-4}}{2}$th upper quantile of standard normal distribution. "Unknown" means that the gene name or ontology is unknown.

Subtype	Gene Id	Gene name	Z_i	Prop. of $\frac{Z_i^S}{SE(Z_i)} \notin (-Z_{\frac{10^{-4}}{2}}, Z_{\frac{10^{-4}}{2}})$	$(1-10^{-4})$% HPD	Rank of t-stat
CEA$\leq$5 vs CEA>5	H44953	Caspase 4, apoptorelated cysteine protease	-0.18	2×10^{-4}	[-0.357,-0.012]	378
	AI309037	Unknown	0.16	4×10^{-4}	[0.004,0.322]	1
	AI380209	Hypothetical protein MGC52423	0.22	5×10^{-4}	[-0.003,0.397]	154
	AA457543	Unknown	0.20	7×10^{-4}	[-0.3011,0.002]	498
	AA485431	Guanine nucleotide binding protein (G protein) alpha 12	-0.15	7×10^{-4}	[-0.01,0.002]	82
	AA991196	Leukocyte-associated Ig-like receptor 1	-0.15	1×10^{-3}	[-0.315,0.009]	228
	AI003792	Ethanolamine kinase 1	0.23	1×10^{-3}	[-0.017,0.477]	42
	AI393078	Homeo box B6 Complement	0.30	1×10^{-3}	[-0.022,0.619]	223
	T68274	component 8 beta polypeptide	-0.24	1×10^{-3}	[-0.504,0.018]	46
	AA701476	CD14 antigen	-0.16	1×10^{-3}	[-0.325,0.014]	39

4.3 Classifying test sets using the support vector machine (SVM)

In order to measure the predictive accuracy based on the selected genes the data set was randomly divided into training and test sets with a ratio of 2:1. We generated 10 pairs of training and test sets. We used SVM for the classification of the test set. We compared the prediction rates based on the top 10 genes selected by our Bayesian approach with the rates based on the top 10 genes in terms of the t-statistic using SVM. Using gene selection by our Bayesian approach we observed 0.54, 0.66, 0.54, 0.74, and 0.65, for the prediction rates of colon versus rectal cancers, CEA≥5 versus CEA<5, stages B versus C, stages B versus D, and stages C versus D, respectively. The prediction rates of two subtypes in the aforementioned classes based on genes selected by t-statistic were 0.52, 0.63, 0.51, 0.71 and 0.63. The result shows that our Bayesian approach makes mild improvements in prediction rates in comparison with the gene selection approach using the t-statistic. This improvement was achieved by identifying a new source of variation, experimenter effect, and using clinical information such as CEA and stages.

The prediction rates for colon versus rectal cancers and stages B versus C were relatively low among other classes of subtypes because of their biological similarities and complex structures. On the other hand, the prediction rates for stages B versus D was relatively large because of their biological heterogeneosity.

<Table 5> Top 10 genes, which were selected by Bayesian multilevel linear mixed effect model, that best discriminate between stages B and C, and the gene ontology of top 10 genes. Z_i^S is the Sth MCMC samples of Z_i, $S=1,\ldots,10000$. $z_{\frac{10^{-4}}{2}}$ is the $\frac{10^{-4}}{2}$th upper quantile of standard normal distribution. "Unknown" means that the gene name or ontology is unknown.

Subtype	Gene Id	Gene name	Z_i	Prop. of $\frac{Z_i^S}{SE(Z_i)} \not\in (-Z_{\frac{10^{-4}}{2}}, Z_{\frac{10^{-4}}{2}})$	$(1-10^{-4})\%$ HPD	Rank of t-stat
B vs C	AA887401	Mitochondrial ribosomal protein S16	0.21	5×10^{-4}	[0.003,0.409]	2
	H62527	Tyrosine 3 5 monooxygenase/tryptophan 5, monooxygenase activation protein, beta polypeptide	-0.16	5×10^{-4}	[-0.328,-0.001]	3
	AA504131	MADS box transcription enhancer factor2 polypeptide C (myocyte enhancer factor 2C)	-0.21	9×10^{-4}	[-0.429,0.009]	5
	AI371315	F-box protein 16 U2(RNU2)	-0.16	1×10^{-3}	[-0.333,0.017]	6

<Table 5> Continued

Subtype	Gene Id	Gene name	Z_i	Prop. of $\frac{Z_i^S}{SE(Z_i)} \not\in (-z_{\frac{10^{-4}}{2}}, z_{\frac{10^{-4}}{2}})$	$(1-10^{-4})\%$ HPD	Rank of t-stat
B vs C	AA456085	small nuclear RNA&-0.15 auxiliary factor 1-like 3	-0.15	1×10^{-3}	[-0.314,0.018]	1
	T81764	Cell division cycle 27	-0.19	2×10^{-3}	[-0.411,0.024]	36
	AA455111	Heterogeneous nuclear ribonucleoprotein C (C1/C2)	-0.22	2×10^{-3}	[-0.461,0.027]	15
	AA489055	Dihydrofolate reductase	-0.12	2×10^{-3}	[-0.253,0.015]	14
	AA490894	Calpastatin	-0.15	2×10^{-3}	[-0.317,0.021]	7
	AA948055	Chromobox homolog 5 (HP1 alpha homolog,Drosophila)	-0.21	2×10^{-3}	[-0.460,0.031]	38

<Table 6> Top 10 genes, which were selected by Bayesian multilevel linear mixed effect model, that best discriminate between stages B and D, and the gene ontology of top 10 genes. Z_i^S is the Sth MCMC samples of Z_i, $S=1,\ldots,10000$. $z_{\frac{10^{-4}}{2}}$ is the $\frac{10^{-4}}{2}$th upper quantile of standard normal distribution. "Unknown" means that the gene name or ontology is unknown.

Subtype	Gene Id	Gene name	Z_i	Prop. of $\frac{Z_i^S}{SE(Z_i)} \not\in (-z_{\frac{10^{-4}}{2}}, z_{\frac{10^{-4}}{2}})$	$(1-10^{-4})\%$ HPD	Rank of t-stat
B vs D	AA915975	Unknown	0.31	4×10^{-4}	[0.014,0.595]	138
	AA970766	Selenoprotein S	0.30	5×10^{-4}	[0.009,0.589]	112
	AI251322	Unknown	-0.24	5×10^{-4}	[-0.480,-0.006]	1
	AA477428	Polymerase (RNA) II (DNA directed) polypeptide G	-0.54	8×10^{-3}	[-1.252,0.173]	2
	AI241088	Solute carrier family 6 (neurotransmitter transporter, GABA) member 13	0.29	1×10^{-3}	[-0.010,0.593]	57
	AI361422	Hypothetical protein FLJ14721	-0.23	1×10^{-3}	[-0.479,0.016]	357
	AA449459	Sulfotransferase family 1E, estrogen-preferring, member 1	0.27	1×10^{-3}	[-0.025,0.568]	59
	AI3717701	Unknown	0.25	2×10^{-3}	[-0.026,0.534]	10
	AI095114	Hypothetical protein FLJ14451	0.22	2×10^{-3}	[-0.024,0.459]	444
	AA970729	Hypothetical protein LOC283852	0.12	2×10^{-3}	[-0.014,0.263]	368

5. Discussion

We have proposed the Bayesian multilevel mixed effect model to assess and examine the experimenter effect or patient effect which is nested in each experimenter as well as to utilize clinical information for identifying a set of DE genes in various classes of subtypes. The model can identify the experimenter effect or patient effect which is nested in each experimenter as a new source of the variation. This model can be used to analyze the inter-laboratory microarray data. In our cDNA microarray experiment of colorectal cancers, we noticed that the experimenter effect was a major source of the variation but the patient which was nested in the experimenter was not. We also observed that our Bayesian approach could make mild improvement over the permutation test in prediction rate. The improvement was achieved due to identifying another source of variation, namely, the experimenter effect, and utilizing clinical information in the Bayesian multilevel mixed effect model. We note that the hierarchical Baysеain model proposed in the paper is not novel. However, as far as we know, it is the first report which explicitly sort out the experimenter effect as a source of a systematic effect in the analysis of gene expression data.

However, the small sample size of stage D and low prediction rates indicate that the current sample size 68, is not large enough to detect the multiple mechanisms that underlie each class of subtypes. One possible reason of the low prediction rates may be that CEA and stage do not contain sufficient information to identify a set of DE genes for the class of subtype, e.g. colon versus rectal cancers or stages B versus C.

Our Bayesian method was developed under the normality assumption in terms of error. It is more useful that we extend the method without this assumption, which is left for a further research. It is also interesting to detect both parametric and nonparametric components of covariates by adding the nonparametric function, f, in the model (3.1), $\sum_{m=1}^{M} \lambda'_{km} x^{m}_{jkl} + f(x_{jkl})$. We also note in this study that different gene sets provide more or less the same prediction rate. This aspect may be viewed from the multiplicity of model and integrating these different models would be valuable in the future research.

<Table 7> Top 10 genes, which were selected by Bayesian multilevel linear mixed effect model, that best discriminate between C and D, and the gene ontology of top 10 genes. Z_i^S is the Sth MCMC samples of Z_i, $S=1,\ldots,10000$. $z_{\frac{10^{-4}}{2}}$ is the $\frac{10^{-4}}{2}$th upper quantile of standard normal distribution. "Unknown" means that the gene name or ontology is unknown.

Subtype	Gene Id	Gene name	Z_i	Prop. of $\frac{Z_i^S}{SE(Z_i)} \notin (-Z_{\frac{10^{-4}}{2}}, Z_{\frac{10^{-4}}{2}})$	$(1-10^{-4})\%$ HPD	Rank of t-stat
C vs D	AA679278	Nuclear factor of activated, T-cells, cytoplasmic calcineurin-dependent 1	0.35	2×10^{-4}	[0.051,0.652]	30
	AI372014	LOC439987	0.17	8×10^{-4}	[0.004,0.342]	7
	AA700883	Step II splicing factor SLU7	-0.16	8×10^{-4}	[-0.320,-0.003]	51
	N93686	Unknown	0.24	1×10^{-3}	[0.001,0.482]	319
	T63324	Unknown	0.25	1×10^{-3}	[0.001,0.499]	106
	N66957	Cytochrome P450, family 27, subfamily A, polypeptide 1	-0.24	1×10^{-3}	[-0.482,000]	2
	AI216166	Chromosome 6 open reading frame 11	-0.24	1×10^{-3}	[-0.481,-0.001]	20
	AI341422	Unknown	0.21	1×10^{-3}	[-0.004,0.422]	99
	AI361167	Unknown	-0.20	1×10^{-3}	[-0.405,0.004]	40
	AI950855	Splicing factor 1	-0.24	1×10^{-3}	[-0.484,0.011]	74

Appendix

Samples from the joint posterior distribution $p(\theta|D)$ are drawn using Markov Chain Monte Carlo (MCMC) techniques, and in particular, Gibbs sampling, which iteratively samples from the collection of full conditionals and produces draws from the joint posterior once convergence is attained. We define $n_{..} = \sum_{k=1}^{K}\sum_{l=1}^{L} n_{kl}$, $n_{k.} = \sum_{l=1}^{L} n_{kl}$, $n_{.l} = \sum_{k=1}^{K} n_{kl}$. The full conditionals for the parameters and hyperparameters of the model are given by

$$p(\alpha_i|rest) \propto \exp\left[-\frac{1}{2}\sum_l\sum_k\sum_j\left(\frac{y_{ijkl}-(\alpha_i+\beta_k+\gamma_{ik}+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}+\tau_l)}{\sigma}\right)^2+\left(\frac{\alpha_i-\alpha_{0i}}{\sigma_\alpha}\right)^2\right]$$

$$\propto N\left(\frac{\alpha_0\sigma^2-\sigma^2_\alpha\sum_l\sum_k\sum_j(\beta_k+\gamma_{ik}+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}+\tau_l-y_{ijkl})}{\sigma^2+n_{..}\sigma^2_\alpha}, \frac{\sigma^2_\alpha\sigma^2}{\sigma^2+n_{..}\sigma^2_\alpha}\right),$$

$$p(\beta_k|rest) \propto \exp\left[-\frac{1}{2}\sum_l\sum_j\sum_i\left(\frac{y_{ijkl}-(\alpha_i+\beta_k+\gamma_{ik}+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}+\tau_l)}{\sigma}\right)^2+\left(\frac{\beta_k-\beta_{0k}}{\sigma_\beta}\right)^2\right]$$

$$\propto N\left(\frac{\beta_{0k}\sigma^2-\sigma^2_\beta\sum_l\sum_j\sum_i(\alpha_i+\gamma_{ik}+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}+\tau_l-y_{ijkl})}{\sigma^2+n_{k.}p\sigma^2_\beta}, \frac{\sigma^2_\beta\sigma^2}{\sigma^2+n_{k.}p\sigma^2_\beta}\right),$$

$$p(\gamma_{ik}|rest) \propto \exp\left[-\frac{1}{2}\sum_l\sum_j\left(\frac{y_{ijkl}-(\alpha_i+\beta_k+\gamma_{ik}+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}+\tau_l)}{\sigma}\right)^2+\left(\frac{\gamma_k-\gamma_{0k}}{\sigma_\gamma}\right)^2\right]$$

$$\propto N\left(\frac{\gamma_{0i}\sigma^2-\sigma^2_\gamma\sum_l\sum_j(\alpha_i+\beta_k+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}+\tau_l-y_{ijkl})}{\sigma^2+n_{k.}\sigma^2_\gamma},\frac{\sigma^2_\gamma\sigma^2}{\sigma^2+n_{k.}\sigma^2_\gamma}\right),$$

$$p(\lambda_m|rest) \propto \exp\left[-\frac{1}{2}\sum_l\sum_k\sum_j\left(\frac{y_{ijkl}-(\alpha_i+\beta_k+\gamma_{ik}+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}+\tau_l)}{\sigma}\right)^2+\left(\frac{\lambda_{im}-\lambda_{0k}}{\sigma_\lambda}\right)^2\right]$$

$$\propto N\left(\frac{\lambda_{0k}\sigma^2-\sigma^2_\lambda\sum_l\sum_k\sum_j x_{jkl}(\beta_k+\alpha_i+\gamma_{ki}+s_{j(l)}+\tau_l-y_{ijkl})}{\sigma^2+\sigma^2_\lambda\sum_l\sum_k\sum_j x^2_{jkl}},\frac{\sigma^2_\lambda\sigma^2}{\sigma^2+\sigma^2_\lambda\sum_l\sum_k\sum_j x^2_{jkl}}\right),$$

$$p(\tau_l|rest)\propto\exp\left[-\frac{1}{2}\sum_k\sum_j\sum_i\left(\frac{y_{ijkl}-(\alpha_i+\beta_k+\gamma_{ik}+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}+\tau_l)}{\sigma}\right)^2+\left(\frac{\tau_l}{\sigma_\tau}\right)^2\right]$$

$$p(s_{j(l)}|rest) \propto \exp\left[-\frac{1}{2}\sum_k\sum_i\left(\frac{y_{ijkl}-(\alpha_i+\beta_k+\gamma_{ik}+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}+\tau_l)}{\sigma}\right)^2+\left(\frac{s_{j(l)}}{\sigma_s}\right)^2\right]$$

$$\propto N\left(\frac{\sigma^2_s\sum_k\sum_j(\alpha_i+\beta_k+\gamma_{ik}+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}-y_{ijkl})}{\sigma^2+Kp\sigma^2_s},\frac{\sigma^2_\tau\sigma^2}{\sigma^2+Kp\sigma^2_s}\right),$$

$$p(\sigma^2|rest)\propto IG\left(A+\frac{n.p}{2},B+\frac{\sum_l\sum_k\sum_j\sum_i[y_{ijkl}-(\alpha_i+\beta_k+\gamma_{ik}+\sum_{m=1}^{M}\lambda'_{im}x^m_{jkl}+s_{j(l)}+\tau_l)]^2}{2}\right),$$

$$p(\sigma^2_\tau|rest) \propto IG\left(A_\tau+\frac{L}{2},B_\tau+\frac{\sum_l\tau^2_l}{2}\right),$$

$$p(\sigma^2_s|rest) \propto IG\left(A_s+\frac{Lp}{2},B_s+\frac{\sum_l\sum_i s^2_{j(l)}}{2}\right),$$

$$p(\alpha_0|rest)\propto\exp\left[-\frac{1}{2}\left(\frac{\alpha_i-\alpha_0}{\sigma_\alpha}\right)^2+\left(\frac{\alpha_0-a_0}{v_\alpha}\right)^2\right]\propto N\left(\frac{\sum_i\alpha_i v^2_\alpha+a_0\sigma^2_\alpha}{pv^2_\alpha+\sigma^2_\alpha},\frac{v^2_\alpha\sigma^2_\alpha}{pv^2_\alpha+\sigma^2_\alpha}\right),$$

$$p(\beta_{0k}|rest) \propto \exp\left(-\frac{1}{2}\left(\frac{\beta_k-\beta_{0k}}{\sigma_\beta}\right)^2+\left(\frac{\beta_{0k}-b_{0k}}{v_\beta}\right)^2\right) \propto N\left(\frac{\beta_k v^2_\beta+b_{0k}\sigma^2_\beta}{v^2_\beta+\sigma^2_\beta},\frac{v^2_\beta\sigma^2_\beta}{v^2_\beta+\sigma^2_\beta}\right),$$

$$p(\gamma_{i0}|rest) \propto \exp\left(-\frac{1}{2}\left(\frac{\gamma_{ik}-\gamma_{i0}}{\sigma_\gamma}\right)^2+\left(\frac{\gamma_{i0}-d_{i0}}{v_\gamma}\right)^2\right) \propto N\left(\frac{\sum_l\tau_l v^2_\tau+c_0\sigma^2_\tau}{Lv^2_\tau+\sigma^2_\tau},\frac{v^2_\tau\sigma^2_\tau}{Lv^2_\tau+\sigma^2_\tau}\right),$$

$$p(\lambda_{i0}|rest) \propto N\left(\frac{\lambda_i v^2_\lambda+e_{i0}\sigma^2_\lambda}{v^2_\lambda+\sigma^2_\lambda},\frac{v^2_\lambda\sigma^2_\lambda}{v^2_\lambda+\sigma^2_\lambda}\right).$$

Acknowledgments

Byung Soo Kim's research was supported by grant No. R01-2004-000-10057-0 from the Basic Research Program of the Korea Science & Engineering Foundation. Sun Young Rha's research was supported by the Korea Science and Engineering Foundation through the Cancer Metastasis Research Center at Yonsei University, Republic of Korea.

References

[1] Bryk, A. and Raudenhush, S. (1992). *Hierarchical Linear Models for Social and Behavioral Research*, Sages, Newbury Park, CA.

[2] Chu, T.M., Weir, B. and Wolfinger, R. (2002). A systematic statistical linear modelling approach to oligonucleotide array experiments. *Mathematical Biosciences,* Vol. 176, 35-51.

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

[4] Hsieh, W.P., Chu, T.M., Weir, B., Wolfinger, R. and Gibson, G. (2003). Mixed model reanalysis of primate data suggests tissue and species biases in aligonucleotide-based gene expression profiles. *Genetics,* Vol. 165, 747 -757.

[5] Ibrahim, J., Chen, M.H. and Gray, R. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of American Statistical Association,* Vol. 97, 88-99.

[6] Jörnsten, R., Wang, H.Y., Welsh, W.J. (2005). DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics,* Vol. 2, 4155-4161.

[7] Kerr, M.K. and Churchill, G.A. (2001). Experimental design for gene expression microarrays, *Biostatistics,* Vol. 2, 183-201.

[8] Kerr, Martin, M. and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data.. *Journal of Computational Biology,* Vol. 7, 819-837.

[9] Kim, B.S., Kim, I., Lee, S., Kim, S., Rha, S.Y. and Chung, H.C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics,* Vol. 21, 517 -528.

[10] Kim, H., Goulb G.,H., Park, H. (2005). Missing value estimation for DNA

microarray gene expression data: local least squares imputation. *Bioinformatics*, Vol. 2, 187-198.

[11] Park, C.H., Jeong, H.J., Jung, J.J., Lee, G.Y., Kim, T.S., Yang, S.H., Chung, H. C. and Rha, S.Y. (2004). Fabrication of high quality cDNA microarray using a small amount of cDNA. *International Journal of Molecular Medicine*, Vol. 13, 675-679.

[12] Rosenwald, A., Wright, G., Chan, W., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Konrad Muller-Hermelink, H., Smeland, E.B., Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell Lymphoma. *New England Journal of Medicine*, Vol. 346. 1937-1947.

[13] Scheel, I., Aldrin, M., Glad, I.K., Sorum, R., Lyng. H., Frigessi, A. (2005). The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics*, Vol. 21, 4277-4279.

[14] Tadesse, M.G., Ibrahim, J.G. and Mutter, G.L. (2003). Identification of differentially expressed genes in high-density oligonucleotide arrays accounting for the quantification limits of the technology. *Biometirics*, Vol. 59, 542-554.

[15] Townsend, J.P. and Hartl, D.L. (2002). Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. *Genome Biology*, Vol. 3(12), research0071.1 -0071.16.

[16] Tusher, V., Tibshirani, R. and Chu, C. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proceedings of the National Academy of Sciences, U.S.A,* Vol. 98, 5116-5121.

[17] van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.M., voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrich, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., Bernards, R. (2002). A Gene-Expression Signature as a predictor of Survival in Breast Cancer. *New England Journal of Medicine,* Vol. 347. 1999-2009.

[18] Wolginger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari C., Paules, R.S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, Vol. 8, 625-637.

[19] Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V. , Ngai, J., Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite

method addressing single and slide scale adjustment systematic variation. *Nucleic Acids Research*, Vol. 30, e15.

[Received June 2006, Accepted September 2006]