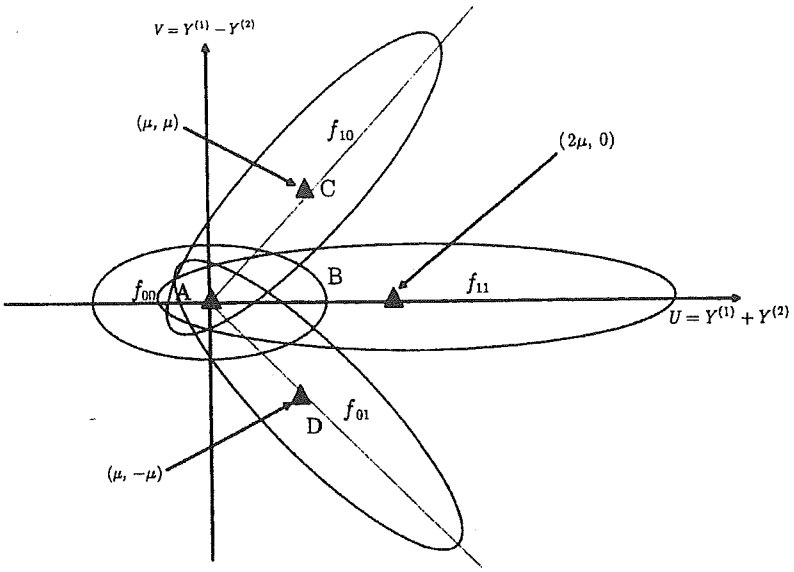

Fig. 2. The two-dimensional mixed normal model is shown schematically. The X-axis indicates the sum of the expression intensities in the query and reference samples; the Y-axis indicates the difference. The states of genes inside the regions A, B, C and D are (OFF, OFF), (ON, ON), (ON, OFF) and (OFF, ON).

where f_{st} denotes the density function under $\tau_1 = s$ and $\tau_2 = t$, p_{st} denotes its mixture rate under $\tau_1 = s$ and $\tau_2 = t$, $\boldsymbol{\theta} = (\mu, \lambda^2, \sigma_\beta^2, \sigma_\epsilon^2)$ and $\boldsymbol{p}$=$(p_{00}, p_{01}, p_{10}, p_{11})$ with $p_{00} + p_{01} + p_{10} + p_{11} = 1$. In this study, each function $f_{st}(u, v)$ is approximated by the two-dimensional normal density function with the moments specified by (4), (5), (6) and (7), respectively. Thus, they are described as

$$f_{00}(u,v|\,\boldsymbol{\theta}) = \phi(u|\,\mathbf{0},\, 2\sigma_\epsilon^2 + 4\sigma_\beta^2)\phi(v|\,\mathbf{0},\, 2\sigma_\epsilon^2),$$
$$f_{11}(u,v|\,\boldsymbol{\theta}) = \phi(u|\,2\mu,\, 4\mu^2(e^{\lambda^2} - 1) + 2\sigma_\epsilon^2)\phi(v|\,0,\, 2\sigma^2),$$

where ϕ denotes a one-dimensional normal density function,

$$f_{10}(u,v|\,\boldsymbol{\theta}) = \phi_2((u,v)|\,\mu(1,1)^{'},\, \Sigma_{10}),$$
$$f_{01}(u,v|\,\boldsymbol{\theta}) = \phi_2((u,v)|\,\mu(1,-1)^{'},\, \Sigma_{01}),$$

where ϕ_2 denotes a two-dimensional normal density function,

$$\Sigma_{10} = \begin{pmatrix} \mu^2(e^{\lambda^2} - 1) + 4\sigma_\beta^2 + 2\sigma_\epsilon^2 & , & \mu^2(e^{\lambda^2} - 1) \\ \mu^2(e^{\lambda^2} - 1) & , & \mu^2(e^{\lambda^2} - 1) + 2\sigma_\epsilon^2 \end{pmatrix}$$

and $\Sigma_{01} = \begin{pmatrix} \mu^2(e^{\lambda^2} - 1) + 4\sigma_\beta^2 + 2\sigma_\epsilon^2 & , & -\mu^2(e^{\lambda^2} - 1) \\ -\mu^2(e^{\lambda^2} - 1) & , & \mu^2(e^{\lambda^2} - 1) + 2\sigma_\epsilon^2 \end{pmatrix}$.

The parameters $\boldsymbol{\theta}$ and $\boldsymbol{p}$ are estimated by maximum likelihood using the Newton-Raphson method and the Spider algorithm (Arcana and Ohtaki, 2005). The flow of estimation is described in Section 3.3.

2.3 Posterior probabilities

Given estimates $\hat{\theta}$ and $\hat{\boldsymbol{p}}$, and normalized data (u_i, v_i), the posterior probabilities with respect to the status of gene expression can be expressed as

$$Pr(\tau^{(1)} = 1 | (u,v), \hat{\theta}) = \frac{\hat{p}_{10} f_{10}(u,v|\hat{\theta}) + \hat{p}_{11} f_{11}(u,v|\hat{\theta})}{f(u,v|\hat{\theta})},$$

$$Pr(\tau^{(2)} = 1 | (u,v), \hat{\theta}) = \frac{\hat{p}_{01} f_{01}(u,v|\hat{\theta}) + \hat{p}_{11} f_{11}(u,v|\hat{\theta})}{f(u,v|\hat{\theta})},$$

$$Pr(\tau^{(1)} \neq \tau^{(2)} | (u,v), \hat{\theta}) = \frac{\hat{p}_{01} f_{01}(u,v|\hat{\theta}) + \hat{p}_{10} f_{10}(u,v|\hat{\theta})}{f(u,v|\hat{\theta})}. \quad (11)$$

If a gene has a relatively large value of $Pr(\tau^{(1)} \neq \tau^{(2)})$, then it is assumed to be expressed differentially between the two samples.

3. Implementation of data analysis

In this section we illustrate how to implement the exploration of differentially expressed genes between two cell types based on the proposed model, using a set of real microarray data.

3.1 Background correction

Let $(y_i^{(\ell)}, b_i^{(\ell)})$, $(\ell = 1, 2)$ be a pair of foreground and background intensities measured by both channels and let $y_i^{(\ell)*}$ be a background corrected value for the i-th gene. So far the background correction has commonly been done by $y_i^{(\ell)*} = y_i^{(\ell)} - b_i^{(\ell)}$ (Eisen et al., 2002) and its logarithmic transformed value is taken as a transformed background corrected intensity. When $y_i^{(\ell)} - b_i^{(\ell)}$ is negative, it is truncated and replaced by an appropriate small positive value, which yields frequently that an unreasonably large dispersion appears at low expression region in the logarithmic transformed background corrected intensities. In this study, we alternatively adopt the equation that is expressed by

$$y_i^{(\ell)*} = y_i^{(\ell)} / b_i^{(\ell)} \quad (12)$$

to correct background effects.

3.2 Normalization

In the first row of Figure 3, (a1) and (a2), whole 21168 gene expression levels of the transformed sample data are shown in order of gene ID. The left side is channel 1(the query sample) and the right side is channel 2 (the reference sample). (a2) shows that the intensity level varies as a wave depending on gene ID, suggesting that some spatial dependent difference may exist between the subgrids. Step N1 aims to remove it (see 'Flow of normalization'). The magnifications of the first four subgrids are shown in the second row of Figure 3, (b1) and (b2). A similar periodic pattern appears in every subgrid, implying that the intensity level of gene expression varies depending on location of the spot in the subgrid. If the spots are printed on the array randomly, the profiles of plots should not have such a systematic tendency. The purpose of Step

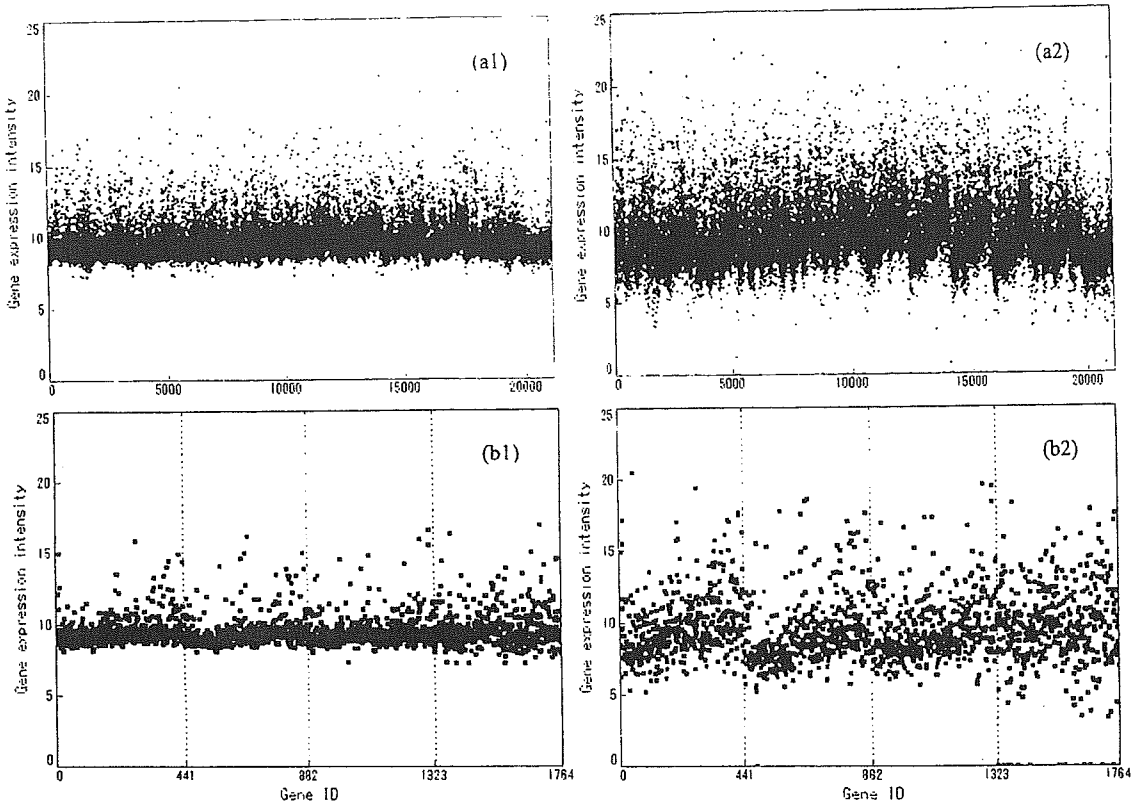

Fig. 3. (a1) and (a2) show the log-transformed gene expression intensities of channels 1 and 2, respectively. (b1) and (b2) are the magnifications of the first four subgrids of both channels.

N2 is to minimize biases associated with location of spots in a subgrid (see 'Flow of normalization'). The S-D plot of the sample data is shown in the left of Figure 5. If the efficiency of the two dyes were the same, the points should distribute almost symmetrically around the X-axis. This dependent dye bias can be removed by Step N3 (see 'Flow of normalization').

Such aberrant trends in the original observations should be removed by applying a combination of global normalization using all of the data and a local one using physical subsets of the data (Quackenbush, 2002). We propose the normalization procedure which is applied to the S-D transformed variables U and V rather than to the original variables $Y^{(1)}$ and $Y^{(2)}$. The procedure is based on the following two assumptions derived from the proposed model. Firstly, most of the u_i are presumed to be generated from the normal mixture density function with two components. Secondly, most of the v_i show only measurement error. Details are described in the Discussion. We can describe the flow of this normalization briefly as follows.

[Flow of normalization]

Step N1. [Adjustment among the subgrids]

Let the values $u_k(i, j)$ and $v_k(i, j)$ be those of the gene located at coordinates (i,j) in the k-th subgrid ($i = 1, \cdots, 21$, $j = 1, \cdots, 21$, $k = 1, \cdots, 48$, $\ell = 1, 2$). Then the data are updated

as follows:

$$u_k(i,j) := u_k(i,j) - Q_k^{(u)}(35) + Q_*^{(u)}(35),$$
$$v_k(i,j) := v_k(i,j) - Q_k^{(r)}(50),$$

where $Q_k^{(u)}(35)$ and $Q_k^{(r)}(50)$ indicate the 35% point of u and 50% point of v in the k-th subgrid, respectively, and $Q_*^{(u)}(35)$ indicates the 35% point of u using all spots on the slide.

Step N2. [Adjustment among spots in a subgrid]

Let $u^*(i,j)$ and $v^*(i,j)$ be the leveled $u(i,j)$ and $v(i,j)$ using all subgrids, which are defined by

$$\begin{cases} u^*(i,j) = \frac{1}{48}\sum_{k=1}^{48} u^{(k)}(i,j), \\ v^*(i,j) = \frac{1}{48}\sum_{k=1}^{48} v^{(k)}(i,j), \end{cases}$$

respectively. Assume that

$$\begin{cases} u^*(i,j) - \bar{u} = a_u(i) + b_u(j) + c_u((i-m_r)(j-m_c)) + \varepsilon_u(s,t), \\ v^*(i,j) - \bar{v} = a_r(i) + b_r(j) + c_r((i-m_r)(j-m_c)) + \varepsilon_r(s,t), \end{cases}$$

where

$$\bar{u} = \frac{1}{21\times 21}\sum_i\sum_j u^*(i\,j), \quad \bar{v} = \frac{1}{21\times 21}\sum_i\sum_j v^*(i\,j),$$

and $a_u(i)$, $b_u(j)$, $c_u((i-m_r)(j-m_c))$, $a_r(i)$, $b_r(j)$ and $c_r((i-m_r)(j-m_c))$ are fixed parameters satisfying

$$\sum_i a_u(i) = 0, \ \sum_j b_u(j) = 0, \quad \sum_i\sum_j c_u(i-m_r)(j-m_c)) = 0.$$

Then estimate the functions a_u, b_u, c_u, a_r, b_r and c_r nonparametrically through the ACE algorithm (Breiman and Friedman, 1985). Given the estimates $\hat{a_u}$, $\hat{b_u}$, $\hat{c_u}$, $\hat{a_r}$, $\hat{b_r}$ and $\hat{c_r}$, update $u(i,j)$ and $v(i,j)$ using the following formula:

$$\begin{cases} u(i,j) := u(i,j) - \hat{a}_u(i) + \hat{b}_u(j) + \hat{c}_u((i-m_r)(j-m_c)), \\ v(i,j) := v(i,j) - \hat{a}_r(i) + \hat{b}_r(j) + \hat{c}_r((i-m_r)(j-m_c)). \end{cases}$$

Step N3. [Removal of dye dependent bias]

Apply the non-parametric regression model,

$$v_i = \phi(u_i) + \varepsilon_i, \quad E(\varepsilon_i) = 0,$$

to the S-D plots to obtain the trend of v_i depending on u_i, and estimate the trend function ϕ by using a moving average method. Then update v_i by $v_i - \hat{\phi}(u_i)$. The normalized channel specific gene expression intensity is given by

$$\begin{cases} \hat{y}_i^{(1)} = \frac{1}{2}\{u_i + v_i - \hat{\phi}(u_i)\}. \\ \hat{y}_i^{(2)} = \frac{1}{2}\{u_i - v_i + \hat{\phi}(u_i)\}. \end{cases}$$

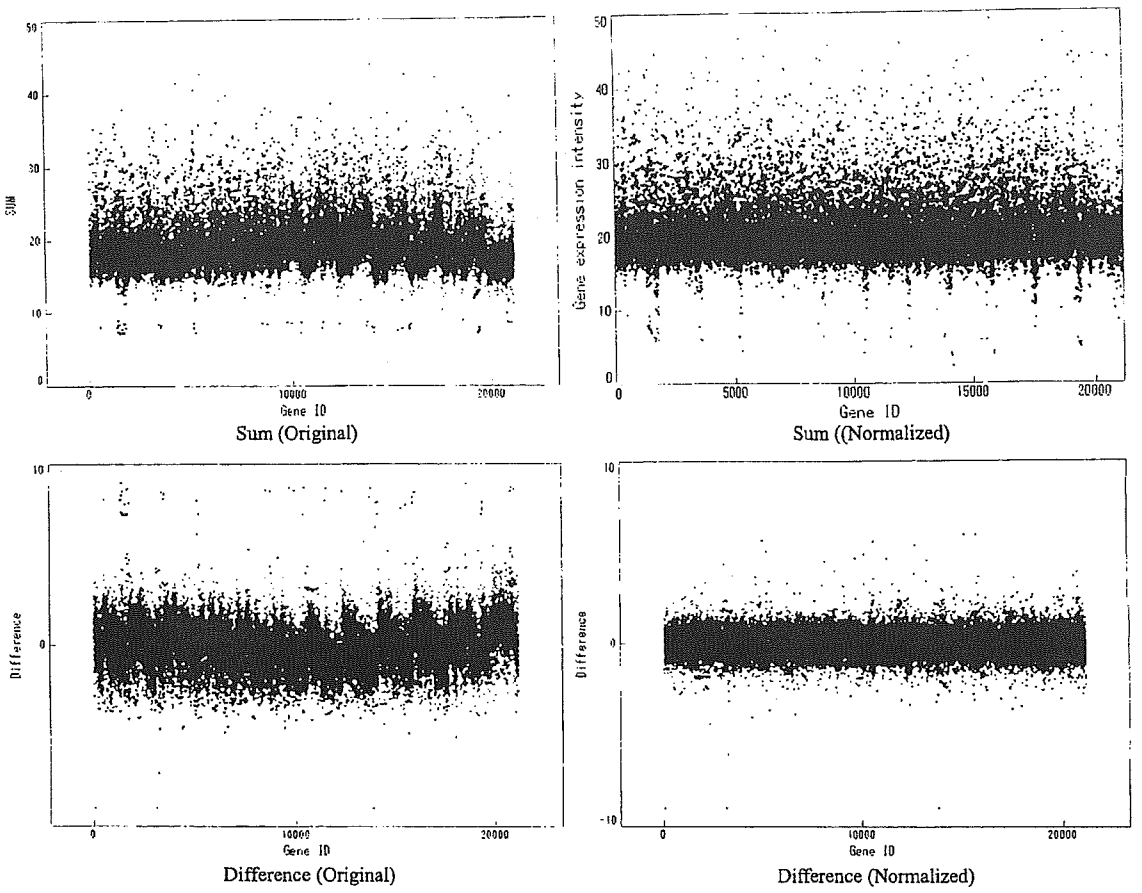

Fig. 4. The scatter plot of the sum of the two channels is shown above and that of the difference is shown below. The scatter plots of original data are shown on the left, and those of normalized data on the right.

Repeat the normalization procedures from Step N1 to Step N3 until the systematic error is removed. Figures 4 and 5 show the efficiency of normalization graphically by comparing the original data (left side) with the normalized data (right side). The Y-axis shows the "sum" of the two channels or the "difference" of the two cannels. Figure 5 shows the S-D plot.

3.3 Parameter estimation

In the ordinary microarray examination, a very low frequency, such as less than three percent, is expected for the heterogeneous components. Therefore they can be negligible at the estimation of the initial value of μ. The model with four components is used in the step E3. The flow of parameter estimation can be described by the following steps.

Step E1. Fit the following mixed normal model with two components to data along the U-axis,

$$(1-\xi)\phi(u-\mu_0|\sigma_0^2)+\xi\phi(u-\mu_1|\sigma_1^2),$$

where μ_0 and μ_1 denote means, σ_0^2 and σ_1^2 denote variances, ϕ is the normal density function given by $\phi(t|\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{t^2}{2\sigma^2}}$, and ξ denotes mixing proportion. We assume here that

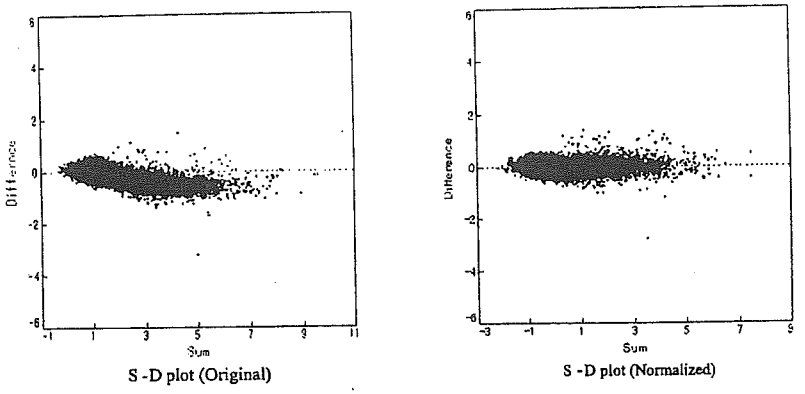

Fig. 5. The S-D plot of original data is shown on the left, and that of normalized data on the right.

$\mu_0 < \mu_1$ and $0 < \xi < 1$.

Figure 6(a) shows the comparison of the fitted mixed normal distribution function with two components (pink line) and the empirical cumulative distribution (brown line). The red line shows the distribution of expression intensities of (ON, ON) genes, the blue line the distribution of expression intensities of (OFF, OFF) genes, and the lime line the distribution of expression intensities of (ON, OFF) or (OFF, ON) genes. Figure 6(b) shows the corresponding density functions.

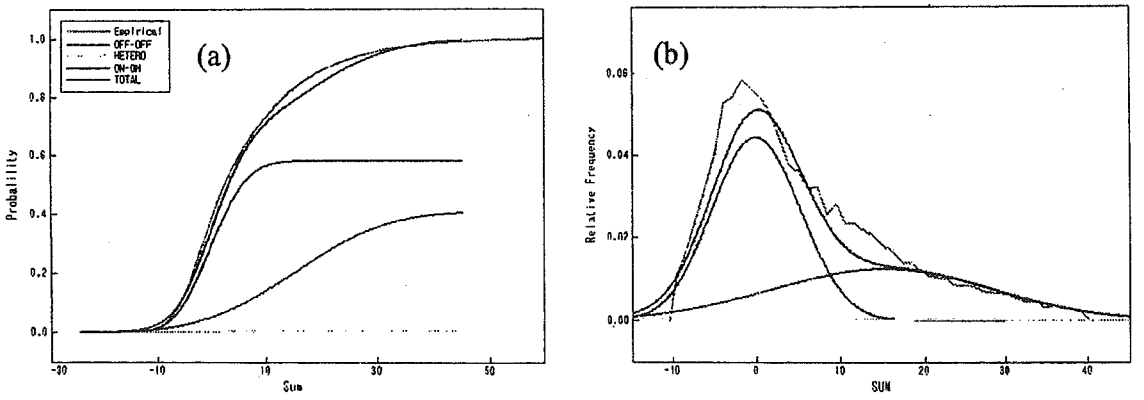

Fig. 6. (a) The estimated mixed normal distribution with two components is shown by the pink line. The distribution of expression intensities of (ON, ON) genes is shown by the red line, that of (OFF, OFF) genes by the blue line and that of (ON, OFF) or (OFF, ON) genes by the lime line. The cumulative distribution of empirical data is shown by the brown line. (b) The density functions corresponding to each distribution function. (See the colored figure given later)

Step E2. Estimate the unknown parameters μ, λ^2, σ_β^2 and σ_ϵ^2 using the following formula:

$$\hat{\mu} = \frac{\mu_1 - \mu_0}{2}, \quad \hat{\sigma_\varepsilon^2} = \frac{1}{2|\{i|\ u_i < \hat{\mu_0}\}|} \sum_{i \in \{i|\ u_i < \hat{\mu_0}\}} v_i^2,$$

$$\hat{\sigma_\beta^2} = \frac{1}{4}\hat{\sigma_0^2} - \frac{1}{2}\hat{\sigma_\varepsilon^2}, \quad \text{and} \quad \hat{\lambda^2} = \log\left(1 + \frac{\hat{\sigma_1^2} - \hat{\sigma_0^2}}{4\hat{\mu}^2}\right).$$

Step E3. Given $\hat{\theta} = (\hat{\mu}, \hat{\lambda^2}, \hat{\sigma_\beta^2}, \hat{\sigma_\varepsilon^2})$, fit the following mixed normal model with four components to the data $\{(u_i, v_i),\ i = 1, \cdots, n\}$:

$$\begin{aligned}
&p_{00}\phi(u|4\hat{\sigma_\beta^2} + 2\hat{\sigma_\varepsilon^2})\phi(v|2\hat{\sigma_\varepsilon^2}) \\
&+ p_{10}\phi_2(u - \hat{\mu}, v - \hat{\mu}|\hat{\Sigma_{10}}) + p_{01}\phi_2(u - \hat{\mu}, v + \hat{\mu}|\hat{\Sigma_{01}}) \\
&+ (1 - p_{00} - p_{10} - p_{01})\phi(u - 2\mu|4\hat{\mu}^2(e^{\hat{\sigma_\alpha^2}} - 1) + 4\hat{\sigma_\beta^2} + 2\hat{\sigma_\varepsilon^2})\phi(v|2\hat{\sigma_\varepsilon^2}).
\end{aligned}$$

Then calculate the tentative maximum likelihood estimate $\hat{\boldsymbol{p}} = (\hat{p}_{00}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{11})$ of $\boldsymbol{p}$.

Step E4. Given $\hat{\mathbf{p}} = (\hat{p}_{00}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{11})$, fit the following mixed normal model with four components to the data $\{(u_i, v_i),\ i = 1, \cdots, n\}$,

$$\begin{aligned}
&\hat{p}_{00}\phi(u|4\sigma_\beta^2 + 2\sigma_\varepsilon^2)\phi(v|2\sigma_\varepsilon^2) \\
&+ \hat{p}_{10}\phi(u - \mu, v - \mu|\Sigma_{10}) + \hat{p}_{01}\phi_2(u - \mu, v + \mu|\Sigma_{01}) \\
&+ \hat{p}_{11}\phi_2(u - 2\mu|4\mu^2(e^{\sigma_\alpha^2} - 1) + 4\sigma_\beta^2 + 2\sigma_\varepsilon^2)\phi(v|2\sigma_\varepsilon^2).
\end{aligned}$$

Then calculate the tentative maximum likelihood estimate $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\lambda^2}, \hat{\sigma_\beta^2}, \hat{\sigma_\varepsilon^2})$ of $\boldsymbol{\theta}$. Iterate step E3 and E4 to convergence.

We obtain the estimates $\hat{\boldsymbol{\theta}} = (7.55,\ 0.64,\ 6.35,\ 0.96)$ and $\hat{\boldsymbol{p}} = (0.582,\ 0.004, 0.004,\ 0.410)$ with the normalized sample data. Given $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{\theta}}$ and the normalized sample data, the degree to which a gene is expressed differentially in the two samples is quantified by the posterior probabilities in (11). The colored S-D plot for normalized sample data is shown in Figure 7. Each color indicates the magnitude of probability: blue represents 0.0 to 0.2, light blue 0.2 to 0.4, yellow 0.4 to 0.6, magenta 0.6 to 0.8 and red 0.8 to 1.0.

4. Discussion

So far several statistical approaches have been made to identify the important genes amongst the many that are measured. Inference about differential gene expression between two cell types is typically based on the difference of measurements between channels 1 and 2 (or the ratio if the logarithmic transformation is used). We adopted two key distinctions of the present approach from earlier efforts. First we introduced the concept of the binary states of gene expressions. "ON" and "OFF", which yields the simple mathematical modeling. The model suggests that the large difference of gene expression intensities between channels 1 and 2 does not always gives

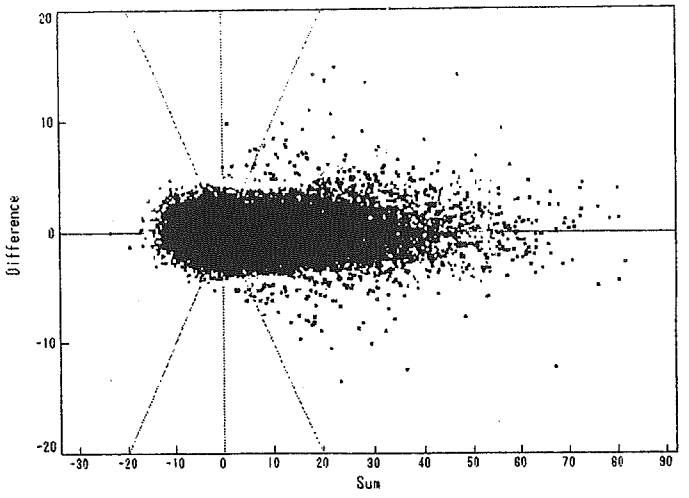

Fig. 7. Each color indicates the magnitude of the posterior probability of a gene expressed differentially in the two samples. Dark blue represents 0.0 to 0.2, light blue 0.2 to 0.4, green 0.4 to 0.6, yellow 0.6 to 0.8 and red 0.8 to 1.0. (See the colored figure given later)

a significant difference, because when its status is (ON, ON), the difference has no meaning. Actually it may be over-simplification to quantify gene expression level using a binary latent variable. A more detailed model may be needed for analyzing the gene expression data in future. However, no high-throughput method is available for discriminating multi-levels more than two for positive expression of gene. Therefore we use the binary latent variable model in microarray data analysis for the present.

The other key distinction of our approach is using the "sum" and "difference" of gene expression intensities of the two samples simultaneously. Several reports have indicated that the magnitude of the difference can have a systematic dependence on overall intensity. Dudoit et al. (2002) described that the plot of $M = \log_2(CH1/CH2)$ vs. $A = \log_2 \sqrt{CH1 * CH2}$, called an MA-plot, clearly showed dependence of the log ratio M on overall spot intensity A. Quackenbush (2002) used a plot similar to the MA-plot. He proposed that the easiest way to visualize the intensity dependent effect is to plot $R = \log_2(CH1/CH2)$ vs. $I = \log_{10}(CH1 * CH2)$, and used this plot for removing the intensity dependent error. In this study, we characterized a gene by the joint "sum" and "difference" of its expression intensities in two samples and applied a two-dimensional mixed normal model with four components. Figure 2 illustrates our model. The genes of interest to us belong to regions C and D, where V reflects the true difference of expression intensity between the two samples. On the contrary, genes belonging to regions A and B are not informative, because V reflects only measurement error. If there exists no common variation between the two cells, then the variances of "sum" and "difference" are expected to be same for the component (OFF, OFF), yielding a spherical shape of scatter plot should appear at the region A in SD-plot after normalization, which makes us easy to check visually degree of the performance of normalization.

One major source of variation is the background intensities. Yoon et al. (2004) stressed the importance of adjusting the effect of the background intensities in the normalization process and concluded that the background measure $\log y_i^{(\ell)} - \log b_i^{(\ell)}$ performs well in their normalization process. As for Affymetrix GeneChip probe level data, Irizarry et al. (2003) stated that the subtraction $PM - MM$ as a way of correcting for non-specific binding is not appropriate and proposed the expression measures based on a log scale linear additive model (RMA). Our approach is similar to theirs. To avoid the quantity $y_i^{(\ell)} - b_i^{(\ell)}$ taking on negative values, we adopt the equation (12) to correct background effects. The nonparametric regression model $\log y_i^{(\ell)} = \phi(\log b_i^{(\ell)}) + \varepsilon_i$, where $E(\varepsilon_i) = 0$, may be applicable to the background correction, but it usually does not work well because of over-fit the data according to our experience. Figure 8 is the scatter diagram of $(\log y_i, \log b_i)$, in which the regression coefficient is close to 1.0 (0.77). The assumption that the foreground intensity is in proportion to the background intensity seems to be appropriate.

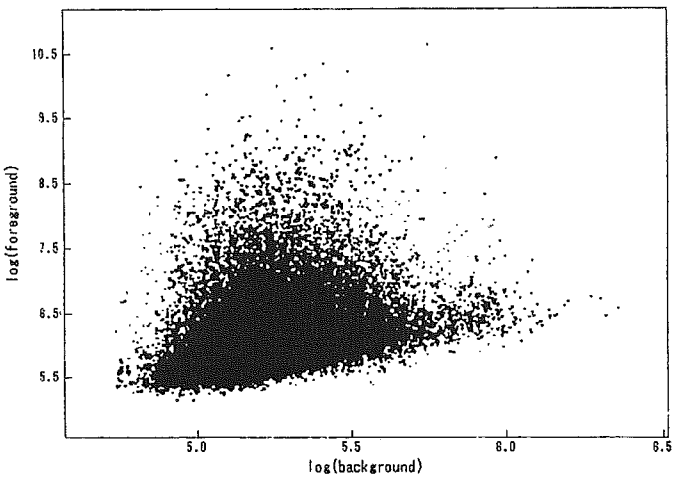

Fig. 8. Scatter diagram of pairs of background and foreground intensities. It shows a tendency towards stronger background intensity with stronger foreground intensity. The calculated regression coefficient is 0.77.

There are many approaches to normalizing expression levels. Dudoi et al. (2002) stated that the usual global normalization approaches are not adequate to remove intensity -or spatially-dependent dye biases. They applied the lowess smoother to each subgrid to remove them. Our approach, though similar to theirs, is not quite the same. The main difference is that we consider the location of a subgrid on a slide and the location of a spot on a subgrid. By Step N1, biases associated with the location of subgrids on a slide can be accommodated. We give a detailed explanation of this in 'Adjustment among subgrids'. Step N2 accounts for the removal of the biases associated with a two-dimensional location of a spot on a subgrid. We try to minimize them using the multiple nonparametric regression model with coordinates of the spot as explanatory variables. Biases linked to the different efficiencies of the two dyes are minimized by Step N3.

[Adjustment among subgrids]

It is natural to assume that most of the genes in a microarray are similarly expressed in the query and reference samples ((OFF, OFF) or (ON, ON)), whereas only a small fraction of genes are differentially expressed ((ON, OFF) or (OFF, ON)). Most of the pairs (u_i, v_i) are presumed to be generated from the two-dimensional normal mixture distribution having components ((ON, ON) and (OFF, OFF)), where the values v_i show only measurement error. Based on this assumption, we adjust the (u_i, v_i)'s among subgrids. Assuming the DNA probes are spotted randomly on the glass slide, we can regard the data set $\{u_k(i.j)\}$ of an arbitrary subgrid k as random samples from the normal mixture distribution with components ((ON, ON) and (OFF, OFF)). If there is no bias depending on location of the subgrid, the values $Q_k^{(u)}(35)$ (35% point of $\{u_{(k)}(i, j)|i = 1,\cdots,12, j = 1,\cdots,12\}$) should have approximately the same value regardless of k. This percentile point depends on the mixture proportion of the two components. If, for example, the mixture proportion is one or zero, namely a single component, the median value (50% point) is the most stable point for the subgrids. The following simulation study supports empirically that the 35% point of $\{u_{(k)}(i, j)|i = 1,\cdots,12, j = 1,\cdots,12\}$ can be regarded as the most stable point.

Thirty thousand simulated data sets $\{u_i|i = 1,\cdots,441\}$ were generated based on model (1), i.e. a gene being "ON" was generated from the log normal density function with mean $\log\mu - \frac{\lambda^2}{2}$ and variance λ^2 and a gene being "OFF" from the normal density function with mean 0 and variance σ_ε^2. The parameters $(\mu, \lambda, \sigma_\beta^2, \sigma_\varepsilon^2)$ were set to (0.5, 0.8, 0.2, 0.1). The mixture proportion of the two components was varied from 0.1 to 0.5 in intervals of 0.1. The standard deviations of each percentile point for each mixture proportion are shown in Figure 9. It demonstrates the

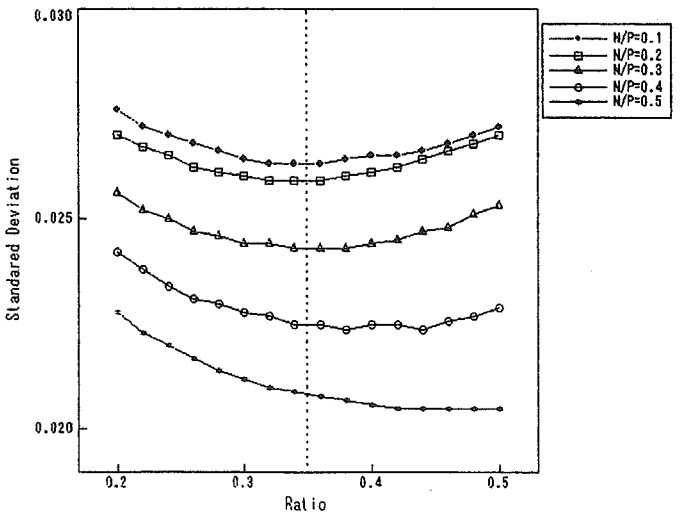

Fig. 9. The standard deviations of the percentile points from thousands of simulated data sets are shown. Blue, light blue, green, yellow, and red show the values for mixture proportions of 0.1, 0.2, 0.3, 0.4 and 0.5.

validity of taking the 35 % point, as it is the most stable.

Acknowledgements

The present study was supported by grants from the New Energy and Industrial Technology Development Organization and the Ministry of Education, Culture, Sports, Science and Technology. We thank Dr. Okazaki and Dr. Hayashizaki, Riken Yakohama Institute, for providing their microarray data. We are grateful to Dr. Cologne, Radiation Effects Research Foundation, for his reading of the manuscript. We thank the editor and two unnamed referees for helpful comments on a previous draft that led to substantial improvement of the manuscript.

REFERENCES

Arcana, IM. and Ohtaki, M.(2005). Multi-target models and their application to data analysis of cellular mortality due to radiation exposure. *Hiroshima Journal of Medical Science* **9(20)**

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

Box, GPE. and Cox DR. (1964). An analysis of transformation (with discussion). *Journal of the Royal Statistical Society* **B26**, 211-252.

Breiman, L. and Friedman, JH. (1985). Estimating optimal transformations for multiple regression and correlation, (with discussion). *Journal of American Statistical Association* **80**, 580-598.

Butte, A. (2002). The use and analysis of microarray data. *Nature Publishing Group* **1**, 951-960.

Churchill, GA. (2002). Fundamentals of experimental design for cDNA microarray. *Nature Genetics Supplement* **32**, 490-495.

Dudoit, S., Yang, YH., Callow, MJ.and Speed, TP. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Technical Report* #578

Eisen, M. (1999). Scan Analyze User Manual.
http//rana.lbl.gov/manuals/ScanAnalyzeDoc.pdf

Fan, J., Tam, P., Woude, GV. and Ren, Y.(2004). Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proceedings of the National Academy of Sciences* **101**, 1135-1140.

Gerhold, DL., Jensen, RV. and Gullans, SR. (2002). Better therapeutics through microarrays. *Nature Genetics. Supplement* **32**, 547-552.

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

Irizarry, RA., Bolstad, BM., Collin F., Cope, LM., Hobbs, B., Speed, TP. (2003). Summaries of Affymetrix GeneChips probe level data. *Nucleic Acids Research* **31**(4), e15.

Kerr, MK., Martin, M. and Churchill, GA. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7-6**, 819-837.

Kendziorski, CM., Newton, MA., Lan, H.and Gould, MN. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899-3914.

Lee, MLT., Kuo, FC., Whitmore, GA. and Sklar, J.(2000). Importance of replication in microarray gene expression studies: Statistical method and evidence from repetitive cDNA hybridization. *Proceedings of the National Academy of Sciences* **97**, 9834-9839.

Li, C.and Wong, WH. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences* **98(1)**, 31-36.

Long, AD., Mangalam, HJ., Chan, BYP., Tolleri, L., Hatfield, GW. and Baldi, P. (2001). Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia Coli K12. *The Journal of Biological Chemistry* **276-23**, 19937-19944.

Newton, MA., Kendziorski, CM., Richmond, CS. and Blatner FR. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8(1)**, 37-52.

Quackenbush, J.(2002). Microarray data normalization and transformation. *Nature Genetics Supplememt* **32**, 496-501.

Saviozzi, S.and Caogero, RA. (2003). Microarray probe expression measures, data normalization and statistical validation. *Comparative and functional genomics* **4**, 442-446.

Schena, M., Shalon, D., Davis, RW. and Brown, PO.(1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470.

Schudhhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H. and Herzel, H. (2002). Normalization strategies for cDNA microarrays. *Nucleic Acid Research* **28(10)**, e47.

Yang, YH., Dudoit, S., Luu, P., Lin, DM., Peng, V., Nagai, J. and Speed, TP. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acid Research* **30(4)**, e15.

Yoon, D., Yi, SD., Kim. JH., Park, T. (2004). Two-stage normalization using background intensities in cDNA microarray data. *BMC Bioinformatics* **5(1)**, 97.

Wu, W., Wildsmith, SE., Winkley, AJ., Yallop, R., Elock, FJ. and Bugelski, PJ. (2001). Chemometric strategies for normalization of gene expression data obtained from cDNA microarrays. *Analytica Chimica Acta* **466**, 451-466.