Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Exploratory methods for checking quality of microarray data

Eun-Kyung Lee & Taesung Park*

Department of Statistics, Seoul National University, Seoul, Korea;

Taesung Park* - Email: tspark@stats.snu.ac.kr; * Corresponding author

revised January 12, 2007; accepted February 02, 2007; published online April 10, 2007

Abstract:

In microarray experiments many undesirable systematic variations are commonly observed. Often investigators analyzing microarray data need to make subjective decisions about the quality of the experiment, by examining its chip image and a simple scatter plot. Thus, a more rigorous but simple method is desirable to determine the quality of microarray data. We propose two exploratory methods to investigate the quality of microarray experiments with replicated chips. The first method is based on correlations among chips and the second on the actual intensity values for each gene. The proposed methods are illustrated using a real microarray data set. The methods provide an initial estimation for determining the quality of microarray experiments.

Keywords: microarray; quality; checking; exploratory methods

Background:

In microarray experiments different sources of systematic and random errors can arise, which may significantly affect the inference on the measured gene expression patterns. A normalization procedure is regularly employed to remove (or minimize) the artifacts due to such errors. While these normalization approaches are useful for adjusting bias of each individual chip, they do not provide a rigorous statistical criterion to detect chips in poor quality. At an earlier stage of analysis, each microarray slide is often examined graphically using the scatter plot between chips to examine large variability (or low reproducibility) and any unusual patterns. However, such examinations are based on subjective human pattern recognition, and chips in poor quality can frequently enter the subsequent analysis, resulting in unreliable inference on the whole microarray study. Therefore, in this study we are concerned about checking the quality of overall microarray experiments and to identify the outlying chips that have much lower reproducibility than other chips.

There have been several approaches for checking reproducibility in microarray experiments. For example, Parmigiani et al., **[1]** defined integrative correlation between two experiments that are conducted separately to answer the same biological question. This integrative correlation is calculated for each gene and called a gene's reproducibility score. King et al., **[2]** used correlations, the rate of two fold changes, and principal component analysis to check the reproducibility of gene expression measurements. Park et al., **[3]** proposed a diagnostic plots for identifying outlying slides. In this paper, we propose an exploratory method to check the quality of microarray data using two different approaches.

Methodology:

We first describe the approach based on the correlations between chips and then describe the other approach based on the actual intensity values.

Correlation Based Approach

Let Y_{ijg} be the normalized log intensity of the jth replicates of the ith treatment for gene g, where i=1, 2, · · · , I, j = 1, 2, · · · , n_i, and g = 1, 2, · · · ,G. For the ith treatment and the jth chip, denote the correlation coefficients and correlation matrices as follows:

$$R_{ij}^{w} = \{r_{ij,il}, \forall l \neq j\},$$

$$R_{ij}^{b} = \{r_{ij,kl}, \forall k \neq i \quad and \quad l = 1,...,n_k\},$$

$$R_{i}^{w} = \{r_{ij,il}, \forall j \quad and \quad l \quad such \quad that \quad j > l\},$$

$$R_{i}^{b} = \{r_{ij,kl}, \forall k \neq i, j = 1,...,n_i, \quad and \quad l = 1,...,n_k\},$$

Where $r_{ij,kl} = corr(Y_{ij}, Y_{kl})$ and $Y_{ij} = [Y_{ij1}, Y_{ij2}, ..., Y_{ijG}]^T$. R_{ij}^{w} represents the within-group correlations of the jth chip in the ith treatment with the other chips in the treatment i, R_{ij}^{b} and does the between group correlations of the jth chip in the treatment i with the other chips from the different treatments. R_{i}^{w} and R_{i}^{b} represent the collections of all within-group correlations and between group correlations for the ith treatment, respectively. Using these correlation measures, we can check reproducibility.

Let $\bar{r}_{ij}^{w}$ be the average of all components of R_{ij}^{w} and $\bar{r}_{ij}^{b}$ be the average of all components of R_{ij}^{b} . If the chips are homogeneous within the same treatment, then R_{ij}^{w} would be

close to 1. Thus, the specificity can be defined in terms of $\bar{r}_{\cdot}^{w}$ and the sensitivity in terms of $\bar{r}_{\cdot}^{w} - \bar{r}_{\cdot}^{b}$, where $\bar{r}_{\cdot}^{w}$ and $\bar{r}_{\cdot}^{b}$ are the overall averages of R_{ij}^{w} and R_{ij}^{b} over i(= 1, · · · , I) and j(= 1, · · · , n_i), that is, they represent the overall means of the within and between correlations, respectively.

Chip-wise correlation plot

For the ith treatment with n_i chips available, there are n_i (n_i -1) / 2 correlation coefficients available. The chip-wise correlation plots show the distribution of these correlations for each chip. The x−axis represents chips and the y-axis represents the distribution of pairwise within correlation coefficients. If a certain chip has a low reproducibility, then it is expected to show a different pattern of correlation coefficient.

Summary correlation plot

For each chip, there are two summary correlation coefficients available: ($\bar{r}_{ij}^{w}$, $\bar{r}_{ij}^{b}$). We propose a summary correlation plot using these measures. The x−axis represents 1− $\bar{r}_{ij}^{w}$ and the y-axis represents 1− $\bar{r}_{ij}^{b}$. Then, each chip can be represented as a point in this plot. If there is an outlying chip, then its point will be located farther from the origin than other treatments. The closer to the origin, the more reproducible is.

Kolmogorov-Smirnov test

To check whether the experiment is reproducible within the same treatment, we compare the distributions of correlations, R_i^w for i = 1, · · · , I. After z-transformation (z = log ((1+r) / (1−r)), we apply the one-sided Kolmogorov-Smirnov test (K-S test) for all pairs of (R_i^w, $R_{i'}^w$) for i ≠ i'. The alternative hypothesis of this test is that the distribution of correlation coefficients derived from the ith treatment is greater than that of the the i'th treatment. The alternative hypothesis is that the distribution function of R_i^w is less than the distribution function of $R_{i'}^w$, that is, the distribution of R_i^w is in the right side of the distribution of $R_{i'}^w$. If the p-value is small, the distribution of Rwi is significantly different from $R_{i'}^w$, that is, the reproducibility differs between two treatments and moreover the i'th treatment is less reproducible than the ith treatment. For the summary of K-S test, we provide I × I matrix $P_{KS} = \{p_{ij}\}$ of p-values, where p_{ij} is the p-value of the K-S test, where the alternative hypothesis is that the distribution of correlation coefficients derived from the ith treatment is greater than that of the jth treatment. Since the p-values of the K-S tests are from the one-sided tests, $p_{ij} \neq p_{ij}$ for all i ≠ j.

Mean test

Alternatively, we can compare the means of two sets of correlations. If the experiment is highly reproducible, the means of R_i^w and $R_{i'}^w$ would be close to each other. To test the differences among I sets of correlations, we use Wilcoxon rank sum test for all pairs of two groups. If the p-value is small, the mean of R_i^w is significantly different from the mean of $R_{i'}^w$. For the summary of mean test, we provide I × I matrix $P_W = \{p_{ij}^*\}$ of p-values, where p_{ij}^* is the p-value of the Wilcoxon rank sum test, where the null hypothesis is that the mean of correlation coefficients derived from the ith treatment is less than that of the jth treatment. Since the p-values of the Wilcoxon rank sum tests are also from the one-sided tests, $p_{ij}^* \neq p_{ji}^*$ for all i ≠ j.

Intensity Based Approach

The correlation based approach checks the treatment-wise and the chip-wise qualities. Therefore, it is not suitable for making any decisions concerning specific genes. In this section, we propose an alternative method to check the gene-wise quality by using the actual intensity values for each gene.

For a specific gene g, we develop test procedures for checking its reproducibility. If the ith treatment group is highly reproducible, the intensity values from the same gene in this group should be similar. For simplicity, we assume that Y_{ijg} has the mean μ_{ijg} with the common variance σ_{ig}^2.

To check the quality of gene g, we test whether the mean of intensities within a treatment are the same or not. The hypothesis of interest is as follows:

$$H_0 : \mu_{i1g} = \mu_{i2g} = ... = \mu_{in_i g}.$$

For testing this hypothesis, the analysis of variance (ANOVA) model is commonly used. **[4]** In our case, however, there is no replicate data available to calculate the within sums of squares. Thus, a traditional ANOVA model is not applicable. Instead, we use the Local-Pooled-Error

approach (LPE, **[5]**) to estimate σ_{ig}^2 . The LPE is based on the idea that genes with similar intensity values will have similar variabilities within the same treatment. In each treatment, all genes with similar intensities are pooled together to estimate variances.

We apply the following two step procedure for each gene.

Step 1: Estimate σ_{ig}^2 , using LPE.

Step 2: Use the following statistic:

$$X_{i,g}^2 = \sum_j \left(\frac{Y_{ijg} - \overline{Y}_{i.g}}{\sigma_{ig}} \right)^2$$

$X_{i,g}^2$ looks like an F-test statistic, but it approximately follows the x^2 distribution with the degrees of freedom under the assumption that Y_{ijg} are normally distributed and σ_{ig}^2 are known. If $X_{i,g}^2$ is sufficiently large, we can conclude that gene g does not have a high specificity in treatment i. That means there are some chips in which gene g has quite different intensity values in the ith treatment. In that case, gene g is called discordant. Otherwise, gene g is called concordant.

Since we test G genes simultaneously, we may need to consider multiple testing issues. In our procedure, we control the false discovery rate (FDR, **[6, 7]**) using q-values. With the predetermined cutoff value, we decide whether gene g is concordant or not.

After deciding whether each gene is concordant or discordant, we calculate the gamma value as a summary measure of concordance

$$\Gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}, \hat{\Gamma} = \frac{n_c - n_d}{n_c + n_d},$$

where Π_c is the probability of concordance and Π_d the probability of discordance. Here, n_c is the number of concordant genes and n_d is the number of discordant genes. In most cases, n_c + n_d is the total number of genes. The gamma values can be computed for each treatment group $\left(\Gamma_i, \quad for \quad i = 1,..., I \right)$.

Example

In this section, the proposed methods are applied to murine B-cell data. To study gene expression profiles in murine B-cell development, total cellular RNA was extracted from five consecutive B-lymphocyte lineage sub-populations (pre-BI cells, large pre-BII cells, small pre-BII cells, immature B-cells, and mature B-cells), and then, gene expression profiles from the five consecutive stages of mouse B cell development were generated with more than five replicates. **[8]**

Murine B-cell data show lower sensitivity (0.66) and specificity (0.02). For the further exploratory analysis, we apply the proposed methods. In the chip-wise correlation plot (Figure 1), most treatments except small Pre-BII cells (chip 23 - chip 27) show high chip-wise correlations. Chip-wise correlations of the small Pre-BII cell treatment have a highly skewed distribution and the third replicate has very small correlations compared to the other chips in the same group. Therefore, we can conclude that this third replicate is problematic and has to be checked or treated before a further analysis. In the summary correlation plot (Figure 2), Murine B-cell data shows outliers, chip 25. All the chips except chips in Small Pre-BII group are located in the upper triangular and chip 25 is far from the other chips. It supports the result from chip-wise correlation plot (Figure 1).

In Table 1, the last column of P_{KS} and P_W show lower p-values than the others. Therefore, we can conclude that the distribution of within correlation in Small Pre-BII group is greater than the distribution of the other groups. Also the mean of within correlation in small Pre BII group is less than the mean of the other groups.

Next, we apply the test based on intensities within treatment. We assume the FDR as 5%. Table 2 shows the result of the intensity based tests. Murine B-cell data show quite different patterns. Especially, the gamma of small Pre-BII treatment is lowest among five treatments. Therefore we can conclude that Murine B-cell data set is less reproducible.

We can conclude that murine B-cell data show lower reproducibility, sensitivity and specificity. Therefore, it is not clear whether or not a further statistical test procedure can detect true differences successfully among the five consecutive stages, especially with small pre-BII cells. It is mainly due to one outlying chip (chip 25), as shown in Figure 3. Therefore, the analyst should check the experimental procedure and tissues used for this chip before a further statistical analysis.

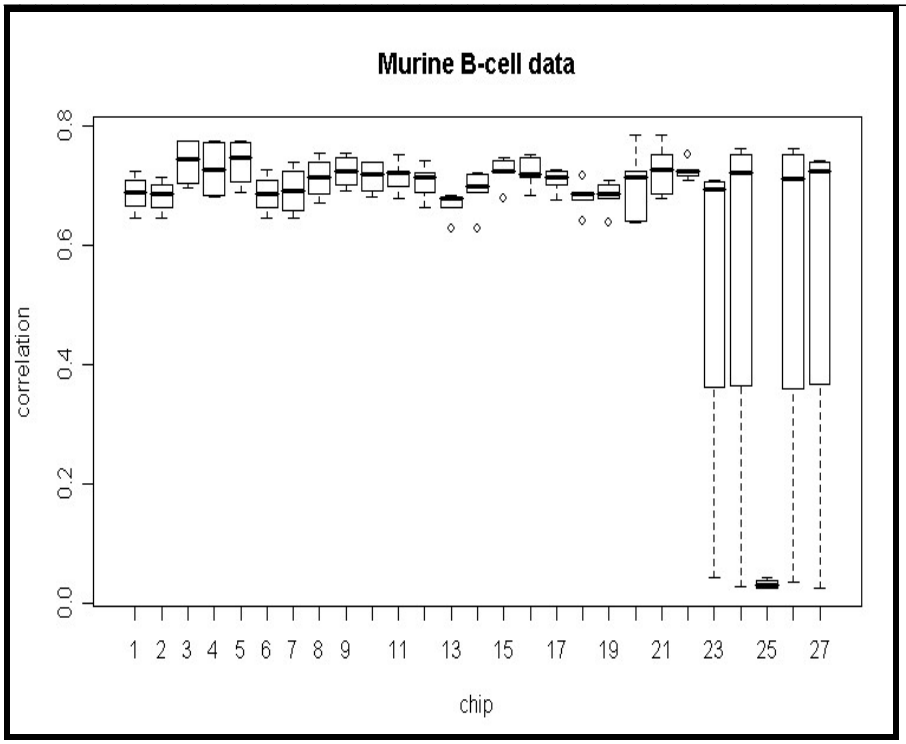

Figure 1: Chip-wise correlation plot: Murine B-cell data. The plots are for the five treatments: Immature B (1, 2, 3, 4, 5), Large Pre-BII (6, 7, 8, 9, 10), Mature B (11, 12, 13, 14, 15, 16), Pre-BI (17, 18, 19, 20, 21, 22), and Small Pre-BII (23, 24, 25, 26, 27)

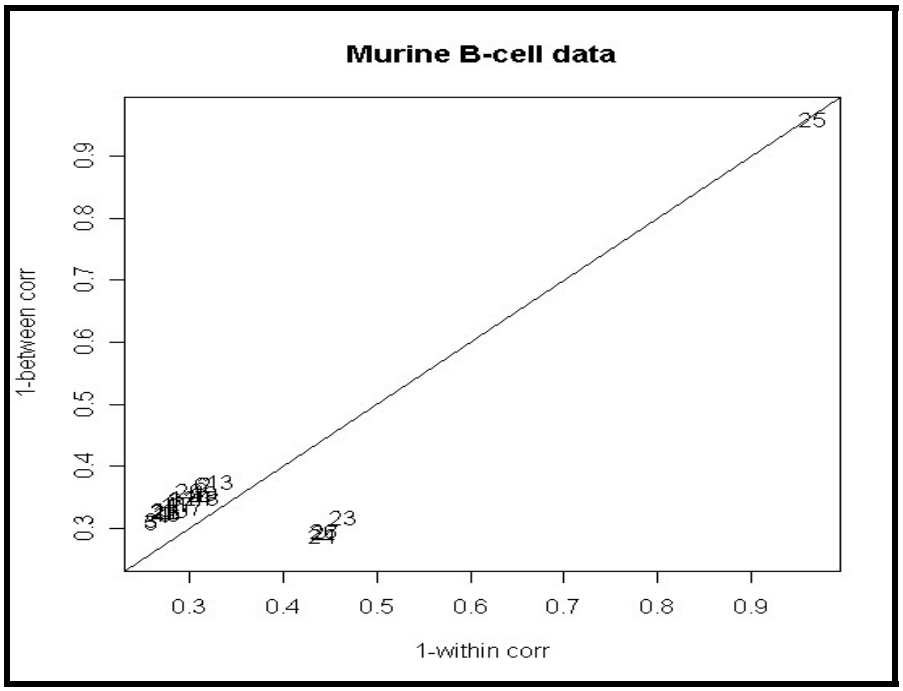

Figure 2: The summary correlation plot. The solid line across the plot is the reference line for specificity. The chips lower than this line represent low specificity and the chips upper than this line represent high specificity

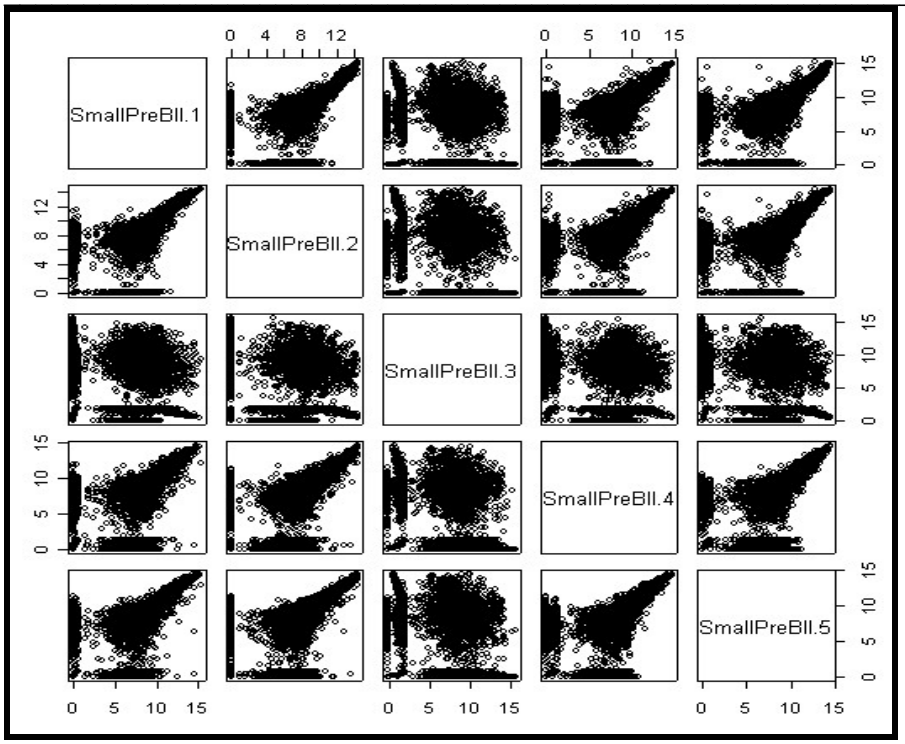

Figure 3: The scatter plot matrix of five replicates for Small Pre BII treatment in Murine B-cell data

P_{KS}	Imm. B	Large BII	Mat. B	Pre BI	Small BII
Immature B	1.00	0.41	0.34	0.52	0.20
Large Pre BII	0.90	1.00	0.62	0.62	0.20
Mature B	0.89	0.81	1.00	0.77	0.15
Pre BI	0.89	0.81	0.94	1.00	0.15
Small Pre BII	1.00	0.90	0.89	0.72	1.00
P_W	Imm. B	Large BII	Mat. B	Pre BI	Small BII
Immature B	1.00	0.37	0.30	0.34	0.11
Large Pre BII	0.66	1.00	0.45	0.42	0.24
Mature B	0.72	0.58	1.00	0.47	0.17
Pre BI	0.68	0.60	0.55	1.00	0.18
Small Pre BII	0.90	0.78	0.84	0.83	1.00

Table 1: P_{KS} and P_W matrices of Murine B-cell data

Treatment	Conc/disc	Γ
Murine B-cell (27)		
Immature B(5)	1086/5509	0.6707
Large Pre BII (5)	1079/5516	0.6728
Mature B(6)	1145/5450	0.6528
Pre BI (6)	1095/5500	0.6679
Small Pre BII (5)	1320/5275	0.5997

Table 2: Summary table for the within test based on intensities

Discussion:

At the initial stage of the microarray data analysis, the exploratory data analysis (EDA) provides the first contact with data. The techniques of EDA consist of a number of informal steps such as checking the quality of the data, calculating simple summary statistics, and constructing appropriate graphs.

The proposed method is a more formal way of checking quality than simple EDA plots. Thus, at an initial stage of the microarray data analysis, the proposed method provides useful information regarding the quality of microarray experiments. The correlation based approaches check the treatment-wise quality, while the test based on the actual intensity values checks the gene-wise quality for each gene.

The proposed method is quite effective in detecting some outlying chips. It is much easier to apply than a traditional method of checking outlying chips either by the principal component analysis or the quality control plot. **[3]**

There are some statistical issues to be taken into consideration, however. First, the log intensities may not have an approximate normal distribution. For simplicity, we have assumed the normal distribution for testing all hypotheses. However extensions to other distributional assumptions are certainly possible. For example, the other distributions such as log-normal and gamma distributions can be easily handled. Second, we did not use a stringent criterion for identifying the concordant/discordant genes. All these genes should be checked by using a analysis such as SAM **[9]** or t-test **[10]** during a later stage of analysis. Third, the correlation coefficients derived from all possible pairs of chips may not be independent. We did not consider these correlations in the current analysis. A more sophisticated approach based on the bootstrapping method is under development which considers possible correlations among the correlation coefficients.

We would like to emphasize that the proposed method is an exploratory analysis. We believe the proposed method to be practically useful, simple and easy to implement that will provide a more rigorous approach in a preliminary overview regarding the quality of microarray experiments. Most proposed methods are implemented in the software arrayQCplot **[11]** and can be downloaded from Bioconductor(www.bioconductor.org).

Acknowledgement:
The authors would like to thank to anonymous referees and the editor whose comments were extremely helpful. This study was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126) and the Brain Korea 21 Project of the Ministry of Education.

References:

[01] G. Parmigiani, *et al., Clin Cancer Res.,* 10:2922 (2004) [PMID: 15131026]
[02] C. King, *et al., J Mol Diagn.,* 7:57 (2005) [PMID:15681475]
[03] T. Park, *et al., Biotechniques,* 38:463 (2005) [PMID:15786811]
[04] P. Pavlidis & W. S. Noble, *Genome Biol.,* 2:RESEARCH0042 (2001) [PMID:11597334]
[05] N. Jain, *et al., Bioinformatics,* 19:1945 (2003) [PMID:14555628]
[06] J. D. Storey & R. Tibshirani, *Proc Natl Acad Sci*., 100:9440 (2003) [PMID: 12883005]
[07] S. Pounds, *Brief Bioinform.,* 38:463 (2005) [PMID:16761362]
[08] R. Hoffmann, *et al., Genome Res.,* 12:98 (2002) [PMID: 11779835]
[09] V. G. Tusher, *et al., Proc Natl Acad Sci*., 98:5116 (2001) [PMID: 11309499]
[10] S. E. Choe, *et al., Genome Biol.,* 6:R16 (2005) [PMID: 15693945]
[11] E. K. Lee, *et al*., *Bioinformatics,* 22:2305 (2006) [PMID: 16864592]

Edited by Susmita Datta
Citation: Lee & Park, Bioinformation 1(10): 423-428 (2007)