

Statistical approaches in the analysis of gene expression data
derived from bone regeneration specific cDNA microarrays

**Jungnam Joo^{1*}, Hongshik Ahn¹, Frank Lombardo², Michael Hadjiargyrou²
and Wei Zhu¹**

¹ Department of Applied Mathematics and Statistics

² Department of Biomedical Engineering

Stony Brook University

Stony Brook, NY 11794

Abstract

Recent advances in molecular biology (e.g. cDNA microarray technology) enables the simultaneous monitoring of the expression level of thousands of genes. Due to the massive amount of complex data generated, sophisticated statistical approaches are necessary in order to properly address the experimental investigation. In this paper, we present statistical analysis of cDNA microarray data derived from bone regeneration experiments. Several interesting features from these data distinguish it from commonly used microarray experiment (i.e., separate hybridization of mRNA samples from reference and experimental tissues, selectively spotted cDNA sequences and 1060 systematically selected blank spots included in each array). Using this data set, we propose new methods for bioinformatic data normalization, as well as the modification and application of various other published methods in order to identify co-regulated gene expression patterns during the healing of a bone fracture. The proposed normalization methods perform effectively to eliminate the variations with a simple algorithm. Results from our cluster analysis revealed several clusters having distinct gene expression patterns during fracture healing. Our simulation study supports the reliability of the proposed methods.

Key words: Bone regeneration, Cluster analysis, Multiple testing, Normalization, Pattern identification

*Current address: Office of Biostatistics Research, NHLBI, Bethesda, MD 20850, jooj@nhlbi.nih.gov

1 Introduction

cDNA microarray technology enables monitoring the expression level of thousands of genes simultaneously. Since microarray experiments generate massive amount of complex data, sophisticated statistical approaches are necessary in order to properly address the problems under investigation.

A typical type of cDNA microarray experiment generates intensities of genes in two different mRNA samples, reference and experimental sample, that are co-hybridized on a single array. Based on the generated microarray data, a number of approaches have been proposed to address several important statistical issues, such as data normalization and identification of genes with different or similar expression levels and patterns.

The reliability of gene expression measurements derived from a microarray is crucial for further analyses. Thus balancing the systematic variations that occur during the microarray experiment is of great practical interest. For normalization of gene intensities in both reference and experimental mRNA samples, the simplest and most widely used method is based on the assumption that there is a constant correction term on the log-scale intensity across the array. That is, the log intensities are corrected by subtracting a constant to obtain normalized values. The global constant is usually estimated from the mean or median over a subset of the genes that are not expected to change. Among the first few suggested normalization procedures, Chen et al. (1997) proposed iterative estimation of the correction term. More recently, Kerr et al. (2000) and Wolfinger et al. (2001) proposed ANOVA models for normalization. Yang et al. (2001) noticed the spatial and intensity dependent biases in numerous experiments and developed methods for estimating the correction term in an intensity dependent manner using the robust-scatterplot smoother loess fit to perform local normalization. Further, Yang and her colleagues extended their method for normalization of location-dependent biases. In addition, the set of genes used for normalization is also very important. The use of all genes is also proposed in the method of Yang et al. (2001) when it is reasonable to assume that the majority of genes are not differentially expressed in the experiment. Otherwise, the set of constantly expressed genes, known as housekeeping or control, are needed to develop a proper method of normalization.

Subsequent to data normalization, statistical methods are applied in order to determine the differential expression of the genes assayed. Cluster analysis was used by Eisen et al. (1998) and

Alizadeh et al. (2000) to identify groups of genes that have similar expression patterns over time. Fraley and Raftery (2002) and Ghosh and Chinnaiyan (2002) introduced the model based clustering with application to gene expression data. Further, identification of differentially expressed genes in different types of cells is another frequently asked question in gene expression research. Dudoit et al. (2002) proposed a method for an adjustment of the p -values for multiple testing procedure that strongly controls the family-wise type I error rate and considers the dependent structure between gene expression levels in a replicated cDNA microarray experiment.

In this paper, we present data from an experiment which was designed to determine the degree of transcriptional complexity of bone regeneration (fracture repair). Several interesting features examined within the bone regeneration data distinguish it from other most commonly used microarray experiments (i.e., separate hybridization of mRNA samples from reference and experimental tissues, selectively spotted cDNA sequences and 1060 systematically selected blank spots included in each array). Using these data, the main purpose of this work was to develop a proper method for normalization and discover the subsets of genes, both known and unknown, that exhibit a distinct expression pattern during fracture repair. Herein we discuss newly proposed statistical methods and a modification of various published works in the analysis of gene expression. Results are provided with a simulation study that evaluates the performance of our utilized methods.

2 Data

Healing of skeletal fractures is essentially a replay of bone development, involving the closely regulated interdependent processes of chondrogenesis and osteogenesis. Using a rat femur model of fracture repair, suppressive subtractive hybridization (SSH; Diatchenko et al., 1996) was performed between RNA isolated from intact bone to that of callus from post-fracture (PF) days 3, 5, 7 and 10 as means of identifying upregulated genes in the regenerative process. These initial time points are selected to represent specific physiological events of the healing callus, including inflammation, chondrogenesis and ossification (Hadjiargyrou et al., 2002). Microarrays were then constructed to confirm the induction of expression and determine the temporal profile of all isolated cDNAs during fracture healing. The following describes these two major experimental stages, SSH and cDNA microarray experiments (Hadjiargyrou et al., 2002).

Suppressive subtractive hybridization

The samples used for this analysis were derived by pooling the RNA from two intact femurs and comparing it to RNA pooled from the fracture calluses from single animals harvested at 3, 5, 7 and 10 days post-fracture (PF). cDNAs derived from the fracture callus material, considered the test pool, and cDNAs from intact femurs, considered the driver pool, were compared by a suppressive subtractive hybridization procedure. This procedure involves subtractive hybridization between the test pool and driver pool, and suppression PCR that selectively amplifies differentially expressed transcripts (upregulated genes) during the fracture healing. Analysis of 3,634 cDNA clones, which were successfully sequenced out of 4,992 cDNAs derived from SSH, revealed numerous known genes (65.8%, 2392 clones) and expressed sequence tags (ESTs; 31%, 1127 clones). The remaining 3.2% (116 clones) yielded no homology and presumably represent novel genes.

cDNA microarray experiment

The custom cDNA microarrays were then constructed by spotting the subtracted cDNA clones from the previous experiment as well as 92 controls (31 positive controls) on the nylon filters. The total number of spots in each array is 6144 that are composed of 384 (16×24) units containing 16 (4×4) spots. 1060 spots were selected as blank spots for the purpose of background adjustment (2 to 3 spots per each unit on the average). All the cDNA clones derived from SSH were spotted and this procedure yields various numbers of replication corresponding to each unique gene. Table 1 summarizes the number of genes that have corresponding number of replications. Note that the total of the second column represents the number of unique genes in this study which is 1766. RNA derived from the calluses of three animals was pooled from each separate time point (PF days 3, 5, 7, 10, 14 and 21) and the RNA sample representing intact bone was established by pooling specimens from intact femurs of three different animals. Target RNA samples from each of the 7 different time points (regarding intact as PF day 0) were radioactively labeled and separately hybridized. The detailed procedures for labeling, hybridization and washing are found at GeneFilters from Research Genetics. A raw image file was generated using phosphoimager. Each membrane image was then analyzed using the GenePix Pro (modified version 3.0) microarray software package. To maintain

Table 1: The level of replication

# Rep ¹	# Gene ²	# Rep	# Gene	# Rep	# Gene	# Rep	# Gene
1	1398	9	7	18	1	36	1
2	171	10	5	20	2	45	1
3	69	11	3	22	1	55	1
4	32	12	4	23	1	57	1
5	21	13	3	24	1	95	1
6	11	15	3	26	1	190	1
7	11	16	1	32	1	195	1
8	7	17	4	35	1		

Rep¹: Number of Replication # Gene²: Number of Genes

consistency between spot to spot, we used a fixed circle method that uses a uniform circle to all the spots to define a foreground target area. However, the hybridization signal differs significantly for each gene and sometimes it is even dispersed beyond the signal spot (bleeding). In case of bleeding, a fixed circle method cannot properly define background area for each spot and thus localized background subtraction was not performed in our experiment. Hence, background subtraction using the information from 1060 blank spots is necessary for a further analysis. Note that the reference (intact) and experimental (PF days 3, 5, 7, 10, 14 and 21) samples were separately hybridized. Refer to Hadjiargyrou et al. (2002) for more biological and technical details on the experiment.

3 Methods

3.1 Data adjustment

The data consist of intensities that have been recorded from n arrays with m genes per array including replicated genes. In the bone regeneration data, we have 7 PF time points with 6144 genes, i.e., $n = 7$ and $m = 6144$. Since the intensities exhibit a large multiplicative component and the expression level of each gene significantly differs in magnitude, it is a common practice to evaluate differences among arrays on the logarithmic scale. In order to take into account the fold change, the logarithm with base 2 is used. This leads the following expression

$$d_{ga} = \log_2 I_{ga}, \text{ for } g = 1, \dots, 6144 \text{ and } a = 1, \dots, 7$$

of the intensity on the logarithmic scale from which we start adjustment. We first considered background adjustment and this is regarded as a within array adjustment in the bone regeneration microarray data in which the reference and experimental samples are separately hybridized. Normalization between arrays is then considered that can be made by monitoring the intensities of the 31 positive controls (ribosomal RNA), whose expression levels are supposed to stay the same across all arrays. The proposed methods as well as several published statistical methods with specific modifications are described in this section for both within and between array adjustments.

3.1.1 Within array adjustment - Background correction

Global method

A global method assumes there is a constant factor that can adjust the background in each array. This leads the following adjustment

$$d_{ga} - b(a), \text{ for } a = 1, \dots, n$$

that shifts the center of the distribution of blanks to zero. The mean, median, mode or the trimmed mean of d_{ga} of 1060 blanks in each array can be used to estimate $b(a)$. This method is very popular due to its simplicity, but this kind of global method cannot provide sufficient adjustment especially when the variation of the intensity depends largely on the particular location on the array.

Location dependent method

This method localizes the global background adjustment method by partitioning the spots in each array into blocks with a suitable size. For each partitioned Block B , the adjustment is performed by

$$d_{ga} - b(a, B), \text{ for } a = 1, \dots, n.$$

Similarly, the mean, median, mode or trimmed mean of d_{ga} of blanks included in Block B will be used to estimate $b(a, B)$.

A localized method with partitioned blocks shows a superior performance to the global adjustment in terms of shifting the central measure of blanks to zero. However, it does not reflect the

effect of some neighboring spots in case that the spot is located at the edge of the block. Therefore, we propose a moving block assignment that defines a block by placing each spot into the center of it. For each spot g on array a , $B(g)$ is assigned with its neighboring spots by placing spot g at the center, and the value of background subtraction is calculated by taking the mean, median, mode or the trimmed mean of blanks within that block. As spot g is moving, the block assignment is changed and so is the adjustment factor. This adjustment yields

$$d_{ga} - b(a, B(g)), \text{ for } a = 1, \dots, n.$$

In practice, moving block method was applied for each unit (4×4 spots) rather than each spot since we believe all spots in each unit are close enough to be handled together. After putting a target unit at the center, the units surrounding the center unit (3×3 units including the center) were used to calculate the adjustment factor for the spots belong to the center unit. Choice of the block size is somewhat arbitrary. A simulation study by Delongchamp et al. (2001a and 2001b) indicates that the block sizes that are too small may over-correct array specific effects, while too large block size may result in a within block effect. In this study, a block size of 12×12 (3×3 units) is chosen for the final analysis after several trials which are not too big and not too small. Block size of 8×8 is also tried for only the purpose of comparison with other methods. When the target unit lies at the edge, only the units adjacent to the target unit were used. This is reasonable because the basic motivation of moving block method is to take into account the effect of adjacent spots.

To investigate the performance of three approaches (global, moving block and fixed block methods), the density plots of the blanks (in \log_2 scale) after applying different adjustment methods were compared. The two density plots of blanks after applying each of the location dependent methods (moving block and fixed block) were almost identical while the density plot resulted from global method presented poor concentration around zero. A better performance of a moving block method compared to a fixed block method can be observed near the units that include blank spots with high intensity. The maximum log-intensity of blanks was found at two units in the array representing PF 7 day. These units are located at the 15th row, and 5th and 6th column of the array. The blank spots belong to the neighboring units of these specific units (the units from the

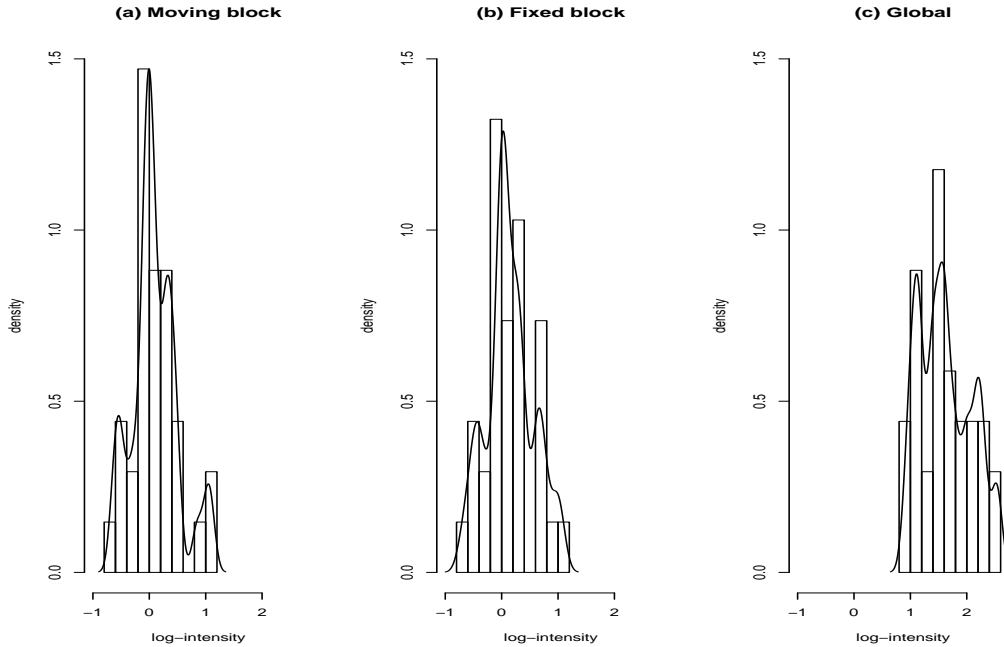


Figure 1: **Comparison between different background adjustment methods:** Density plots of blank spots from a selected region after three different adjustment methods. A moving block method shows superior performance than a fixed block method. Both location dependent methods (a moving and fixed block method) perform better than a global method.

14th to 16th row, and 4th to 7th column) were examined and Figure 1 illustrates the density plots of the blanks on these units after the three approaches. A moving block method resulted in higher peak around zero in the density of blanks compared to the result from a fixed block method. Therefore, we applied a localized background subtraction with the moving block assignment method that subtracts the median of blanks to the bone regeneration data.

3.1.2 Normalization - Between array adjustment

The concept of between array adjustment in the present bone regeneration microarray data, characterized by a separate hybridization of the experimental and reference samples, is identical to that of within array adjustment in the microarray data derived from the experiment that co-hybridizes experimental and reference samples and thus include two data columns (red and green intensity) from each array. To illustrate this procedure for the microarray data, we denote the geometric

mean of the intensities of each gene by $\log_2 \bar{I}_g$, i.e.,

$$\log_2 \bar{I}_g = \left(\sum_{a=1}^n \log_2 I_{ga} \right) / n.$$

Moreover, let G denote the set of all m spots and let H be the set of controls on which there is no treatment effect. This leads the following set:

$$H = \{g \in G : \mu_{g1} = \cdots = \mu_{gn} = \bar{\mu}_g\}.$$

A between array adjustment is made to shift the center of the distribution of controls to their mean intensity ($\bar{\mu}_g$). In case of a global adjustment, the adjustment is made by

$$d_{ga}^* - \{c(a) - M\}, \text{ for } a = 1, \dots, n,$$

where M represents the estimate for the overall mean intensity (mean of the $\bar{\mu}_g$'s) and $c(a)$, again, is estimated by mean, median, mode or trimmed mean of $\{d_{ga}^* : g \in H\}$ for each array a . Note that d_{ga}^* is the background-subtracted log-intensity. The following equation verifies the fact that the mean of controls is shifted to M by subtracting $c(a) - M$ from d_{ga}^* for each array a when $c(a)$ is estimated by the mean of controls:

$$\frac{1}{n_c} \sum_{g \in H} [d_{ga}^* - \{c(a) - M\}] = M + \sum_{g \in H} d_{ga}^* / n_c - c(a) = M, \text{ for } a = 1, \dots, n,$$

where n_c denotes the number of controls. Similarly, when the median of controls is used to estimate $c(a)$, the median is shifted to M .

When the goal is to force the central measures of controls (means or medians) to be the same for all arrays, scaling d_{ga}^* by $c(a)/M$ instead of subtracting $c(a) - M$ can also be applied because of the following relationship

$$\frac{1}{n_c} \sum_{g \in H} \frac{d_{ga}^*}{c(a)/M} = M \cdot \sum_{g \in H} \frac{d_{ga}^*}{n_c} \frac{1}{c(a)} = M, \text{ for } a = 1, \dots, n.$$

Since the scales are different at each array, we applied this method which not only adjusts location

but also stabilizes the scale. The difference of scales were tested using Levene's (1960) test for heteroscedasticity.

3.2 Data filtering

Following the normalization procedure, data reduction was performed by filtering out uninformative genes. In the bone regeneration microarray data, there are many replicated genes. Replication occurs because different segments of the DNA sequence corresponding to a specific gene are spotted several times. Since all replicated spots represent the same gene, we only need a single representative value for each of the specific genes found on the microarray. Besides noise, we expect all the replicated spots for one specific gene to have the same expression pattern over time. Since this does not always happen in practice even after the normalization, it is important to filter out those spots that have inconsistent patterns compared to others. There is no straightforward criterion for this selection, but based on the fact that the spots representing the same gene are expected to possess similar expression patterns over the 7 PF time points, we calculated both Pearson's and Spearman's correlations between these replicated spots over time and used these statistics as measures for filtering. Basically, both the correlation measures are statistics which have a focus on a trend of the measurements rather than magnitude. For this reason, these are most frequently used similarity measures in the cluster analysis of gene expression data. Spearman's correlation is a robust rank correlation which is not affected by outliers, but in biological sense, a rapid increase or decrease of the intensity may need to be captured. Hence, we put an emphasis on Pearson's correlation and used Spearman's correlation as a reference.

For example, gene S-Adenosylmethionine Decarboxylase sequences appear in 10 spots on each of 7 membranes and the Pearson's correlation matrix among these spots over time is given in Table 2. The expression of this gene is quite consistent except on the 9th and 10th replications. These spots have quite low or even negative correlation with most of other replicates and it seems reasonable to filter out these spots before taking mean. The remaining 8 spots are included and the intensities of these spots are averaged into one value. However, there are many cases where the decision is not as simple as in this example. In such cases, we considered the biological aspect of each individual gene and made a decision based on a biological interpretation. 596 spots were identified as outliers based on this criterion out of 3634 successfully sequenced genes. These spots were excluded and

Table 2: Pearson’s correlation matrix between 10 spots of S-Adenosylmethionine Decarboxylase over 7 time points.

Spot	1	2	3	4	5	6	7	8	9	10
1	1	0.908	0.769	0.803	0.884	0.900	0.775	0.827	0.027	0.253
2	0.908	1	0.675	0.584	0.618	0.972	0.945	0.925	-0.215	0.494
3	0.769	0.675	1	0.941	0.766	0.784	0.615	0.665	0.364	0.202
4	0.803	0.584	0.941	1	0.916	0.691	0.453	0.592	0.397	0.058
5	0.884	0.618	0.766	0.916	1	0.669	0.453	0.594	0.257	0.048
6	0.900	0.972	0.784	0.691	0.669	1	0.924	0.965	-0.195	0.571
7	0.775	0.945	0.615	0.453	0.453	0.924	1	0.858	-0.221	0.643
8	0.827	0.925	0.665	0.592	0.594	0.965	0.858	1	-0.418	0.673
9	0.027	-0.215	0.364	0.397	0.257	-0.195	-0.221	-0.418	1	-0.675
10	0.253	0.494	0.202	0.058	0.048	0.571	0.643	0.673	-0.675	1

the replicated sequences are averaged to obtain a unique measurement for each specific gene. This procedure resulted in 1685 final distinct genes. The reduction is not as much as it appears to be. As mentioned in Section 2, 1766 unique genes were identified from the 3634 gene spots. 81 genes with irregular expression out of these 1766 unique genes were removed after filtering. This is a 4.6% reduction. It is necessary to eliminate the spots with unusual expression patterns among repeated genes before extracting the representative pattern of a particular gene.

3.3 Pattern identification

Cluster analysis (Eisen et al., 1998) was applied to the normalized and reduced data profile (1685 genes with 7 time points) in order to find patterns and group genes together with similar expression profiles. For our bone regeneration microarray data, our focus was on identifying genes that have different expression profiles in each of the 6 PF time points as compared to those of the intact day. Thus, the normalized log-intensity of each PF time point is subtracted by that of intact. Pearson’s correlation between genes at 6 PF time points was used as a similarity measure. Since we did not have any prior information on clusters, hierarchical clustering techniques using the pairwise average linkage method was selected for a computational algorithm. The pseudo F and pseudo t^2 statistics that performed the best out of thirty methods for estimating the number of clusters in the simulation study by Cooper and Milligan (1984) and Milligan and Cooper (1985) served as criteria for determining the number of clusters. It has been recommended to look for consensus among these statistics. That is, local peaks of the pseudo F statistic combined with a small value of the

pseudo t^2 statistic and a larger pseudo t^2 for the next cluster fusion. Clusters were then compared to examine whether the profiles of each cluster are distinctive. Discriminant analysis that gives the result that shows the distinction between clusters is performed. Clustering algorithm that uses $1 - r$ as distance, where r denotes Pearson's correlation, is identical to the algorithm that uses the Euclidean distance when the measurements are standardized (subtracted by the mean and divided by the standard deviation). We used standardized data as an input for cluster and discriminant analysis for ease of calculation.

4 Results

4.1 Data adjustment

Figure 2 shows the plots of the normalized log-intensity of each PF day versus that of intact day after localized background subtraction with moving block assignment and global between array adjustment that also performs scale adjustment based on the 31 positive controls (both methods are using the median to estimate the adjustment factor). This figure clearly indicates increased gene expression during the progression of the fracture callus through its early stages (PF days 3-10), while the control spots roughly stay on the reference lines ($y = x$) and the blank spots stay around the origin. Relative to the intact bone, the highest level of overall expression is observed at PF day 14 (note the higher intensity and shift of dots) and by PF day 21, expression levels begin to decline towards the control (reference line).

4.2 Cluster analysis

Given the normalized data set, we next utilized clustering to reveal co-regulated gene expression. Table 3 represents the clustering history. The pseudo F statistic peaks at 3, 5, 12 and 19 clusters. The further increase of the pseudo F statistic does not appear to be meaningful. The pseudo t^2 statistic indicates that all these values can be possible candidates. When we choose 5 as the optimal number of clusters and examine the pattern plots, the expression pattern changes are not quite distinct. This means that the pattern plots of 5 clusters are not distinguishable because 5 is a small number to represent more than a thousand subjects. Thus, we excluded 5 and 3. The comparison between the plots of 12 and 19 clusters does not show much difference in the pattern between those

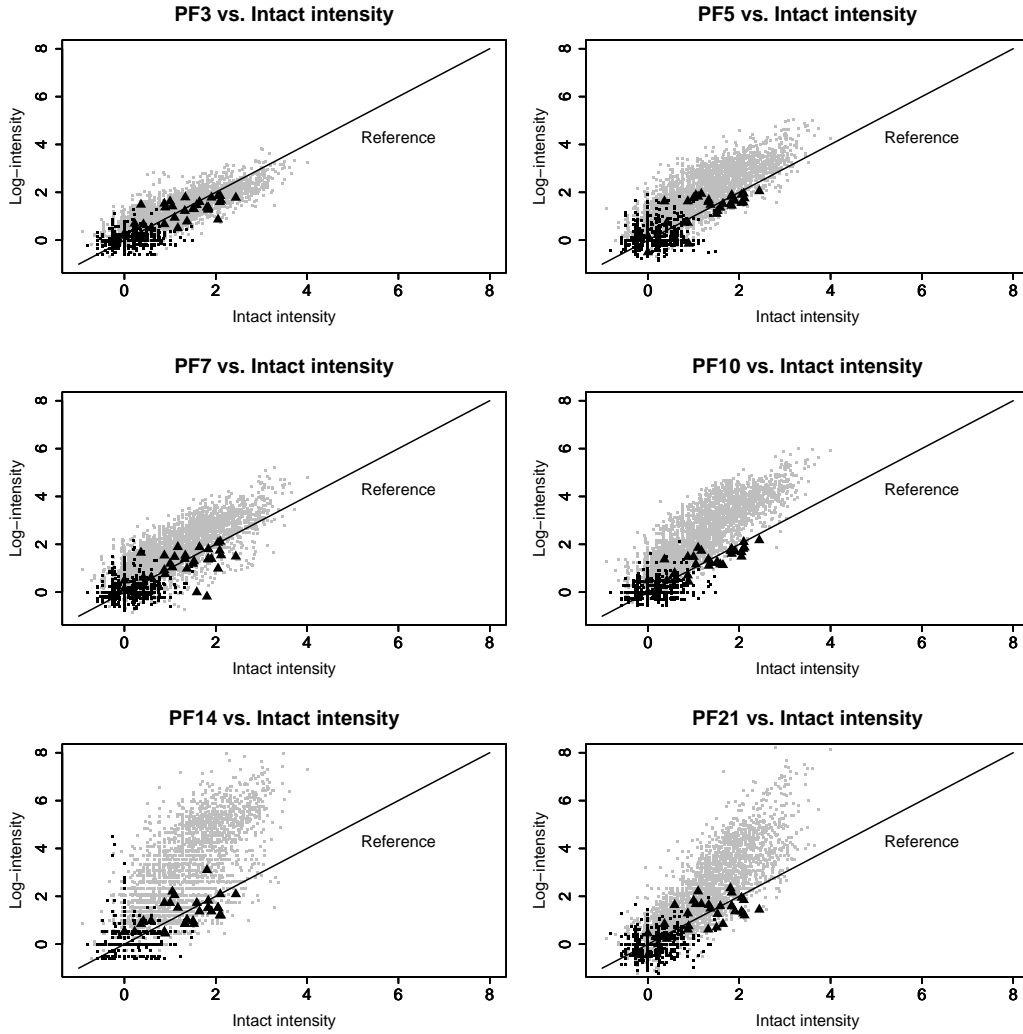


Figure 2: **The normalized intensity plot:** The normalized intensity of each PF day versus that of intact after localized background subtraction with moving block assignment (median correction) and global between array adjustment based on the 31 positive controls. Figures indicate the increased gene expression during the fracture healing, while the control spots (black solid \triangle) roughly stay on the reference line ($y = x$) and the blank spots (blank solid dots) are concentrated around the origin.

Table 3: Clustering history of the bone regeneration microarray data. Pseudo F peaks at 3, 5, 12 and 19 clusters and each of these clusters has relatively small pseudo t^2 that shows large jump at the next cluster fusion.

NCL ¹	PSF ²	PST2 ³	Dist ⁴
30	128	13.1	0.7125
29	131	15.3	0.7204
28	135	5.8	0.7223
27	125	136	0.7258
26	129	12.7	0.7334
25	133	8.9	0.7375
24	138	9.3	0.7435
23	141	42.0	0.7547
22	147	2.8	0.7568
21	152	18.6	0.7662
20	155	41.5	0.7742
19	161	15.2	0.7815
18	157	106	0.7863
17	159	52.8	0.8021
16	165	36.3	0.8306
15	174	17.8	0.8423
14	187	10.0	0.8531
13	200	15.1	0.8593
12	201	82.9	0.8646
11	178	229	0.8822
10	180	90.5	0.9000
9	189	59.1	0.9061
8	212	21.5	0.9242
7	224	82.9	0.9639
6	234	100	0.9670
5	273	51.3	0.9826
4	235	263	1.0318
3	330	27.7	1.0842
2	43.4	598	1.1132
1	.	43.4	1.1976

¹NCL : Number of clusters in each step

²PSF : Pseudo F Statistic

⁴Dist: Normalized root mean square distance

³PST2: Pseudo t^2 Statistic

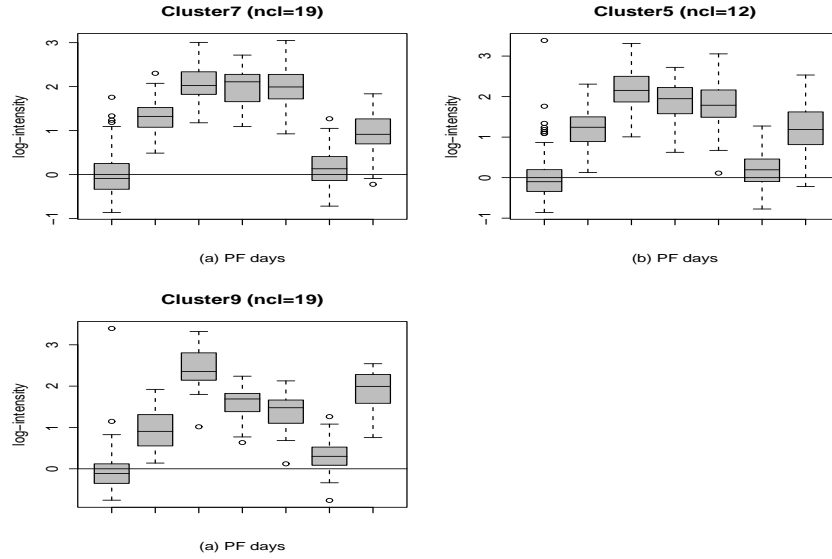


Figure 3: **Comparison between 12 clusters and 19 clusters:** Clusters in the first column (Part (a)) are combined to the cluster in the second column (Part (b)) when 19 clusters are reduced to 12. In each plot, the box plot is given for each time point.

clusters that are separated by the increase of the number of clusters from 12 to 19. Figure 3 shows one example. Part (a) represents the pattern plots of Clusters 7 and 9 when 19 is selected as the optimal number of clusters. These clusters are combined to one cluster (Cluster 5) when 19 clusters are reduced to 12. The pattern plot of the combined cluster is given in part (b). The plot in part (b) appears to have a similar pattern as both plots shown in part (a). Hence, we chose 12 as a possible optimal number of clusters for the analysis of the bone regeneration microarray data.

Figures 4 shows the pattern plots of each cluster with the number of genes in each cluster. The clusters show distinct patterns of expression. Some clusters have genes whose expression peaks early and then declines (Clusters 2, 3, 6, 11), while other clusters show a continuing increase of gene expression through the healing process (Clusters 8, 9). A number of other clusters show a pattern of successive increased and decreased expression (Clusters 1, 4, 5, 7, 10, 12), indicating a fluctuation in general metabolic activity.

Application of the discriminant analysis yields the classification results shown in Table 4. The last column shows the percentage of the correctly classified objects in each cluster which is determined by the discriminant function. The high percentage (all greater than 79%) indicates a quite successful classification of clusters.

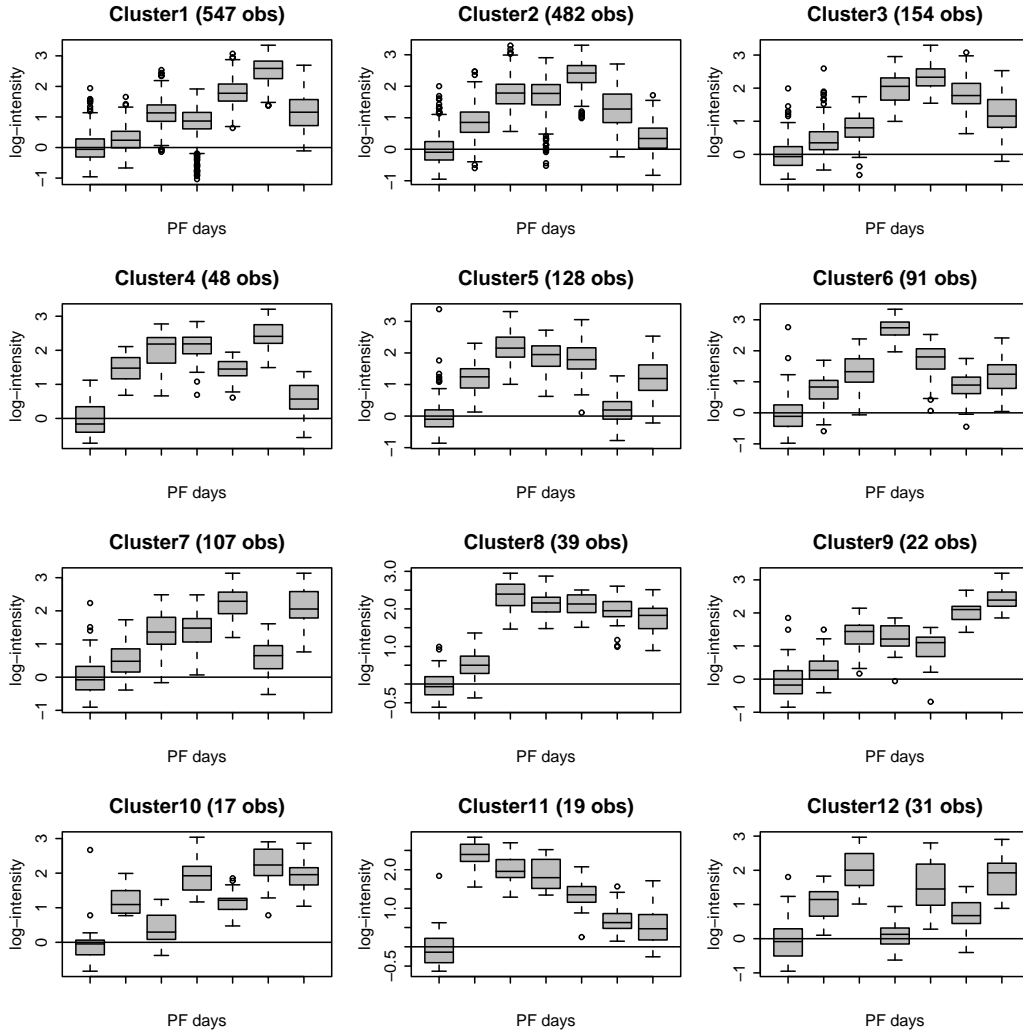


Figure 4: **Pattern plots** of the 12 clusters obtained from the bone regeneration data.

Table 4: **Discriminant analysis result:** Percentage of the correct classification.

Cluster	Number of observations	Predicted value	Percentage
1	547	475	86.8
2	482	385	79.9
3	154	133	86.4
4	48	47	97.9
5	128	113	88.3
6	91	84	92.3
7	107	94	87.9
8	39	38	97.4
9	22	20	90.9
10	17	15	88.2
11	19	19	100.0
12	31	30	96.8

Table 5: Mean profile and the percentage of each cluster in the simulation study.

Cluster	Percentage	Mean profile						
		Intact	PF 3	PF 5	PF 7	PF 10	PF 14	PF 21
1	15%	-0.12	1.70	1.24	-0.05	1.78	1.20	1.73
2	35%	-0.79	0.53	1.03	1.14	2.42	1.37	0.30
3	50%	-0.95	0.84	1.06	1.61	1.26	0.43	1.20

4.3 Simulation study

4.3.1 Design

A simulation study was carried out to evaluate the performance of our methods for analyzing the microarray data. As numbers of columns and rows of an array were reduced to half, the final size of array was reduced to one fourth. Five thousand sets of 1356 genes including 24 controls and 264 blank spots were generated. Each of the 1356 genes was randomly assigned to one of three clusters with distinct patterns. Table 5 displays the mean profile of each cluster and the percentage of genes assigned to each of these clusters. For each gene, the location in the array was also randomly assigned. We calculated the percentage of blocks with significant block effects in the real example for each array and randomly selected k_a blocks that will have significant block effects in each array based on this percentage. The simulation data were generated based on the following model

$$\frac{\log_2 I_{ga} - b(a, B(g))}{c(a)/M} \approx \log_2 \mu_{ga},$$

where $b(a, B(g))$ is a location dependent background effect and $c(a)/M$ is a global control effect of Array a . $\log_2 \mu_{ga}$ represents the true log-intensity of Gene g on Array a . For genes which are neither controls nor blanks, the multivariate normal distribution was chosen as a distribution of the mean profile for each cluster with standard deviation of $0.5 \cdot I(\log_2 \mu_{ga})$. $\log_2 \mu_{ga}$ was set to be zero for blanks and set to be a constant (ranging from 0.7 to 0.9) for controls. Global control effect of each array, $c(a)/M$, was generated from a uniform distribution and multiplied by $\log_2 \mu_{ga}$. For those genes belonging to the randomly selected k_a blocks, block effect (selected from the true value) and background effect (generated from the uniform distribution) were combined to form the location dependent background effect $b(a, B(g))$. Otherwise, $b(a, B(g))$ was generated from the uniform distribution that represents simple background effect without block effect.

Table 6: The average proportion that the genes in the same original cluster are clustered together.

Cluster	Proportion
1	0.841
2	0.891
3	0.745

Table 7: Frequency (%) of each number of clusters selected out of 5000 simulation trials.

Number of clusters	Percentage	Cumulative percentage
3	0.4226	0.4226
4	0.3792	0.8018
5	0.1040	0.9058
6	0.0366	0.9424
7	0.0204	0.9628
8	0.0114	0.9742
9	0.0102	0.9844
10	0.0060	0.9904
11	0.0030	0.9934
12	0.0022	0.9956
13	0.0010	0.9966
14	0.0012	0.9978
15	0.0010	0.9988
16	0.0004	0.9992
17	0.0006	0.9998
18	0.0000	0.9998
19	0.0002	1.0000

4.3.2 Results from the simulation study

For each of the five thousand simulated data sets, we performed the normalization (location dependent background adjustment with global within array adjustment) and average linkage cluster analysis. The proportion of genes that are clustered together among the genes originally generated from the same cluster was calculated. This proportion was then averaged for each run of the simulation. The result is given in Table 6. Moreover, the frequency of selecting the correct number of clusters out of five thousand runs was counted to examine the accuracy of the criterion for selecting the number of clusters. Table 7 shows the frequency of the chosen number of clusters.

Pattern plots of a sample simulated data set resulting in three clusters by our analysis are given in Figure 5 and compared with the original mean profiles. This figure shows that the normalization is properly performed to remove block and control effects and the clustered sets represent the original patterns quite accurately. The percentage of correct selection of the number of clusters is 42.3% out of 5000 simulation trials when we consider exactly 3 clusters only, but it increases to

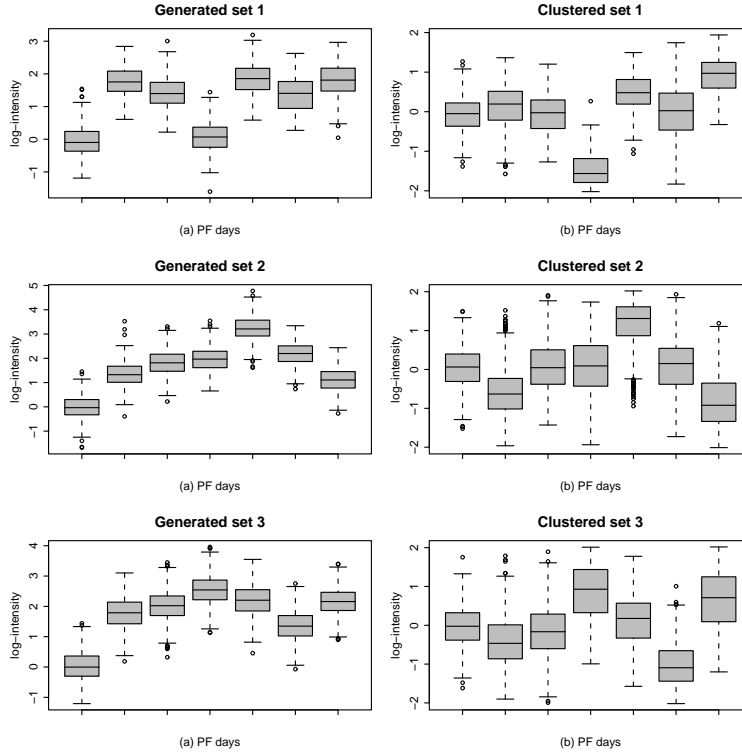


Figure 5: Comparison between the original profiles and clustered profiles from a sample simulation data set resulting in 3 clusters. Part (a) shows the original profiles corresponding to each cluster. Plots in part (b) show 3 clusters obtained after applying the proposed normalization and cluster analysis.

90.6% if we include 4 or 5 clusters selected by our procedure. Even when four or five is selected as the number of clusters by the criterion we used, usually the fourth and fifth clusters contain very small number of genes, while the other three clusters contain the genes representing the main expression patterns. Tables 8 and 9 summarize the results from two sample simulation data sets in which four and five clusters are selected, respectively. Figure 6 provides the pattern plots corresponding to the sample simulation data set summarized in Table 8.

5 Discussion

The underlying objective of the cDNA microarray experiment of bone regeneration presented here was to verify the upregulation of genes derived from SSH. Further investigation of gene expression and clustering will enable us to explore future genomic research of skeletal fracture repair. In a statistical sense, the main focus of this paper was to develop a proper method for normalization

Table 8: Result from the simulation - summary of a sample simulation data set in which four is selected as the number of clusters.

Generated set		Clustered set		Proportion ¹
Cluster	Number of genes	Cluster	Number of genes	
1	178	1	211	89.3%
2	439	2	502	90.7%
3	631	3	515	78.0%
		4	20	

¹ Proportion of genes in the second column which are clustered together in the fourth column.

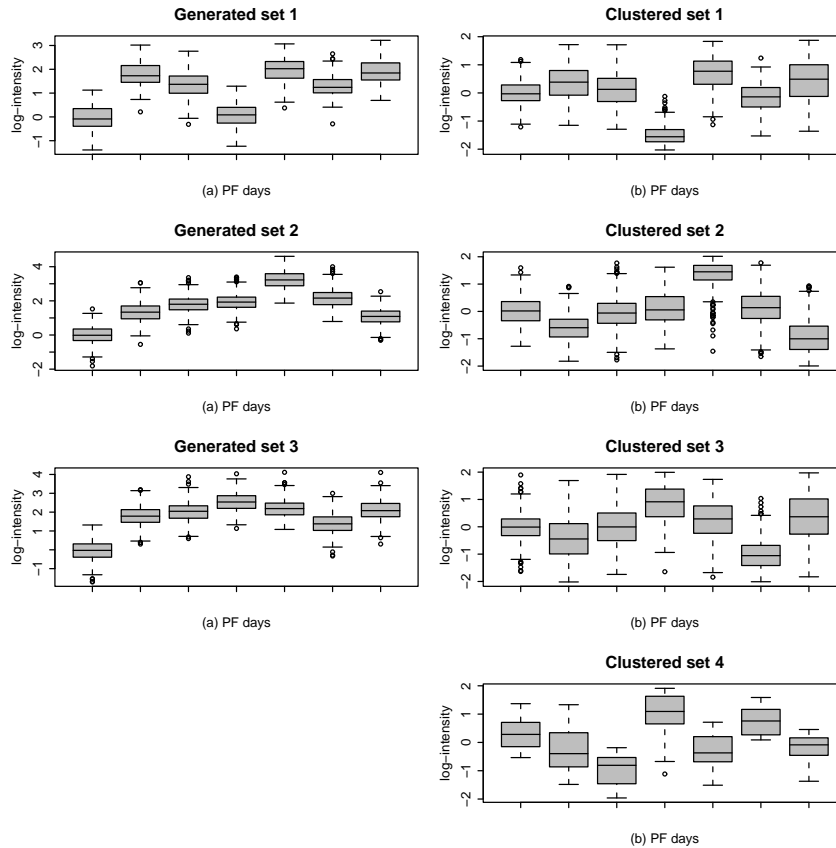


Figure 6: Pattern plots from a sample simulation data set in which four is selected as the number of clusters: Part (a) shows the original profile of the three generated sets corresponding to each cluster. The pattern plots of four clusters after applying normalization cluster analysis are given in part (b).

Table 9: Result from simulation - summary of a sample simulation data set in which five is selected as the number of clusters.

Generated set		Clustered set		Proportion ¹
Cluster	Number of genes	Cluster	Number of genes	
1	199	1	176	75.9%
2	434	2	543	94.2%
3	615	3	502	76.3%
		4	8	
		5	20	

¹ Proportion of genes in the second column which are clustered together in the fourth column.

and statistical analysis for the gene expression data generated from the custom bone regeneration microarray experiments.

Normalization as an attempt to reduce the systematic biases is an important pre-process of the gene expression data. There are several interesting aspects in the present cDNA microarrays constructed for the bone regeneration experiment which distinguish our data from more typical gene expression data. Each of the following aspects is not very unusual in microarray technology these days, but we found that not many statistical methods are developed for this type of data. Therefore, it is important to point out these differences and develop a statistical methodology suitable to the given data.

First, the background subtraction was not performed during image analysis. Instead, 1060 blank spots were included and systematically located in each array. Thus, a background subtraction based on these blank spots was essential before further analysis. Ideally, the intensity of these spots should be shifted to zero. On the other hand, the intensity levels of the control (housekeeping) genes are expected to be constant across arrays. Therefore, we can treat these in a similar way since the difference is merely the center which should be zero for blanks, and a certain constant for controls. Another important feature of the present data is that the reference (intact mRNA) and the experimental (mRNA from PF days 3, 5, 7, 10, 14 and 21) samples were separately hybridized. Note that a separate hybridization prevents several complex adverse effects that may arise during the experiment such as cross-hybridization between two mRNA samples and dye biases, while it causes confounded real treatment effect with the array effects. Thus, controls (a set of genes that do not possess a treatment effect) were used as reference for a between array normalization rather than a within array normalization.

The number of blank spots is 1060 which is roughly 18% of the data. Even after partitioning, each block includes a sufficient number of blanks. Hence, it was possible to obtain a satisfactory result by performing the localized background subtraction both with fixed block and moving block assignments.

Intensity dependent method (Yang et al., 2001) is well known for between array adjustment (normalization). However, several problems arise in applying this method to the bone regeneration microarray data. First, due to the small number of controls, it is difficult to conclude that we observe intensity based biases. Moreover, when the mean intensities of controls are not spread around all the ranges of the mean intensity, which is the case in the bone regeneration microarray data, it is not possible to obtain an adjustment factor for those genes that lie out of the range of the mean intensities of controls. A slight modification of intensity dependent adjustment can be carried out by partitioning genes based on their location instead of mean intensity. Ideally, the trimmed mean or median (mean or mode can also be used) of control genes in each partition becomes the estimate of the adjustment term. However, there are many blocks that do not include any control spots after partitioning due to the small number of control genes (31). Even if a block contains a control, the number of control spots included is only one or two. If we use all the genes, the real treatment effect should be estimated and added back to the adjusted data, since gene sequences spotted on the array are selected by SSH that we expect high expression. However, estimation of the real treatment effect is a difficult task. Since the result from this method showed a poor performance than that from the global method (in terms of shifting the center of controls to the reference point), the final method applied in this study was a global between array adjustment.

Given the normalized data set followed by data reduction by filtering, the statistical method that can answer the question of grouping genes with a similar pattern of expression is relatively straightforward using a clustering algorithm with Pearson's correlation as a similarity measure. Assessing the clustering result is the most critical part in applying cluster analysis, while it is complicated. Pseudo F and pseudo t^2 statistics were used to select the number of clusters for the present microarray data in order to provide possible candidates. In deciding the final number of clusters among these candidates, we tried to distinguish the distinct patterns effectively, but at the same time, we tried to avoid unnecessary increase in the number of clusters. This final decision was made to provide a rough guideline to the future research. A biological inputs will be valuable in

making the decision. The distinction of clusters was examined by comparing pattern plots between clusters and by comparing the distributions of within and between cluster correlations.

The result from this cluster analysis revealed that majority of these functionally unknown genes (734 genes) are grouped within clusters 1, 2 and 3 that also contain 449 known genes. From Figure 4, the transcriptional profile patterns of these three clusters are marked by a sharp increase at PF day 3 which remain relatively high throughout the first two weeks, and then start decreasing at the end of the second week (Clusters 2 and 3) or decreasing at week three (Cluster 1). This high level of initial activity correlates well with the temporal sequence of biological events occurring during the early stages of fracture repair. However, given the large number of known genes presented in these three clusters, as well as the diverse functional families they present, it would be premature and inaccurate to assign a possible function to the unknown genes. Clearly, a further biological examination is required to help define the nature of these novel genes.

Our simulation study supports the accuracy of the criteria we used to select the number of clusters. More detailed procedure to determine the number of clusters can be considered using a model based clustering algorithm. However, most of the test procedures are based on the mean profile while our concern is on the difference in the pattern rather than the difference in the magnitude. Therefore, an adjustment of the data by standardization is necessary.

Some of the methods utilized here are quite subjective in order to answer the questions raised in this particular study, in particular data filtering and the decision of the number of clusters, and requires intensive interdisciplinary work. The idea of moving block method proposed in this study, however, can easily be adapted in various normalization methods of microarray data.

The methods proposed in this paper was motivated from the given bone regeneration data. Our methods have been developed for answering the questions on gene activities during bone healing process. Thus some complicated techniques have been involved in the proposed methods. In making a decision on things such as the size of moving blocks or the number of distinct clusters, the conclusions may vary. Further inputs from biologists may also change the conclusion. We do not expect the identical results by other investigators, but the variations in conclusion would not reverse our findings on overall patterns of the gene activities on bone healing process.

References

- Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, E., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, T. O., Warnker, T., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000), Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.
- Chen, Y., Dougherty, R. and Bittner, M. L. (1997), Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, **2**, 364-374.
- Cooper, M. C. and Milligan, G. W. (1984), The effect of error on determining the number of clusters. *College of Administrative Science Working Paper Series*, **84-2**. The Ohio State University, Columbus, OH.
- Delongchamp, R. R., Velasco, C., Evans, R., Harris, A. and Casciano, D. (2001a), A median estimate that adjusts cDNA array data from nuisance effects. Unpublished manuscript.
- Delongchamp, R. R., Velasco, C. and Razzaghi, M. (2001b), Computing normalized intensities using standard statistical packages. Unpublished manuscript.
- Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchick, A., Moqudam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. and Siebert, P. D. (1996), Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences*, **93**, 6025-6030.
- 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. *Statistica Sinica*, **12**, 111-139.
- Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998), Cluster analysis and

- display of genome-wide expression patterns. *Proceedings of National Academy of Sciences*, **95**, 14863-14868.
- Fraley, C. and Raftery, A. E. (2002), Model based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, **97**, 611-631.
- Ghosh, D. and Chinnaiyan, A. M. (2002), Mixture modeling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275-286.
- Hadjiargyrou, M., Lombardo, F., Zhao, S., Ahrens, W., Joo, J., Ahn, H., White, D. W. and Rubin, C. T. (2002), Transcriptional profiling of bone regeneration: Insight into the molecular complexity of wound repair. *Journal of Biological Chemistry*, **277**, 30177-30182.
- Kerr, M. K., Martin, M. and Churchill, G. A. (2000), Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, **7**, 819-837.
- Levene, H. (1960), Robust Tests for Equality of Variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann.(Eds.), *Contributions to Probability and Statistics*, Stanford: Stanford University Press, 278-292.
- Milligan, G. W. and Cooper, M. C. (1985), An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159-179.
- Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R. S. (2001), Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, **8**, 625-638.
- Yang, Y. H., Dudoit, S., Luu, P. and Speed, T. P. (2001), Normalization for cDNA microarray data. In Bittner, M. L., Chen, Y., Dorsel, A. N. and Dougherty, E. R. (Eds.), *Microarrays: Optical Technologies and Informatics*, **4266** of *Proceedings of SPIE*.