A Novel Approach to Determine Normal Variation in Gene Expression Data*

Vinay Nadimpally and Mohammed J. Zaki
Computer Science Department
Rensselaer Polytechnic Institute, Troy, NY 12180
{nadimv, zaki}@cs.rpi.edu

ABSTRACT

Animal models for human diseases are of crucial importance for studying gene expression and regulation. In the last decade the development of mouse models for cancer, diabetes, neuro-degenerative and many other diseases has been on steady rise. Microarray analysis of patterns of gene expression in mouse models of various pathological types and the study of molecular level changes as a result of interventions, holds lot of promise to the understanding of biological processes involved. The genes which show normal variance across genetically identical mice are of particular interest because they could serve as a databank for possible false positives in gene expression studies in similar kind of mice. Also they could provide useful insights into the biological processes behind the differential expression patterns in otherwise similar mice. Our approach systematically removes variance due to experimental noise in each of the mice and then mines for normal variance among the identical mice.This analysis carried over six tissues sampled from mice, resulted in several genes which showed variations among identical mice, thus enabling a comprehensive database of normal variations in gene expression for mouse models. A large number of these genes are known to be related to stress response, hypertension and heat shock. Also Principal Component Analysis was done to visualize similarity among the mice models and within the replicates. These studies help in the design of gene expression studies in mouse models and help in validation of the results.

Keywords: gene expression, replicates, mouse models, normal variance, immune response, hypertension, principal component analysis

1. INTRODUCTION

High-throughput gene expression has become an important tool to study transcriptional activity in a variety of biological samples. Mouse models are playing an important role in the study of developmental biology, genetics, behavior, and disease. To interpret experimental data, the extent and diversity of gene expression for the system under study should be well characterized. An ubiquitous and under-appreciated problem in microarray analysis is the incidence of microarrays reporting non-equivalent levels of an mRNA or the expression of a gene for a system under replicate experimental conditions. This is due to various sources of noise contributed by the technicalities of the experiment and also biological variation. Statistical methods have been developed to determine the significance of the fold changes observed between the test sample and normal sample, accounting for the noise. Unfortunately little is known about the normal variance of mouse gene expression in vivo or the noise in the microarray data due to biological variation.

Levels of several genes vary significantly in more than one tissue. These variances in gene expression might be one of the reason for observing different phenotypes of transgenic mice based on a different genetic background. These studies will help to define the baseline level variability in mouse gene expression and accentuate the importance of replicated microarray experiments. Furthermore, the analysis is expected to expose some genes which have been reported previously as differentially expressed because of pathological state or experimental variation. Also further study of these genes could lead us to useful insights into the biology behind the normal gene expression variation, independently of genetics.

The use of replicates in microarray experiments is not very widely recognized because of the prohibitive costs and the time involved. However, the use of replicates is very important to develop more reliable and consistent findings. A single microarray experiment would contain significant inherent noise. By pooling data from replicates, we can provide a more reliable classification of gene expression [11]. Statistical methods have been used for analyzing replicated cDNA microarray experiments. The authors stress that replication in microarray studies is not equivalent to duplication and hence is not a waste of scientific resources. Experimental replication is essential to reliable scientific discovery in genetic research. [20] developed a mathematical model to estimate the number of replicates needed to provide data within acceptable error limits. [11] have modeled the effect of number of cultures, arrays, etc. on the variability in the data (from preliminary experiments); target fold change in data and costs of experimental components.

Although wide range of study has gone into the use of replicates to minimize the effect of noise on the resultant findings, little research has gone into explaining the normal variation of some genes in genetically similar species. [4] analyzed the expression profile of normal male C57Bl6, C3H, Balb/c, DBA/2, and FVB mice, respectively, to study the normal variance. In total, they profiled gene expression of

*This work was supported in part by NSF CAREER Award IIS-0092978, DOE Early Career Award DE-FG02-02ER25538 and NSF grant EIA-0103708

75 mouse samples of 15 different individuals. They used a U74A mouse Gene array from Affymetrix to quantify transcript levels in brain, liver, heart, spleen, and kidney in the five genetic backgrounds. [16] assess natural differences in murine (i.e., mice/rats) gene expression, they used a 5406-clone spotted cDNA microarray to quantify transcript levels in the kidney, liver, and testis from each of 6 normal male C57BL6 mice. They use ANOVA to compare the variance across the six mice, to the variance among four replicate experiments performed for each mouse tissue. For the 6 kidney samples, 102 of 3,088 genes (3.3%) exhibited a statistically significant mouse variance at a level of 0.05. In the testis, 62 of 3,252 genes (1.9%) showed statistically significant variance, and in the liver, there were 21 of 2,514 (0.8%) genes with significantly variable expression. Immune-modulated, stress-induced, and hormonally regulated genes were highly represented among the transcripts that were most variable. The expression levels of several genes varied significantly in more than one tissue. Natural variability is also interesting from a biological standpoint. Inbred mouse populations allow us to study how gene expression varies without the effect of genotypic variation.

The analysis of the gene expression assays from three tissues in mouse (kidney, testis, and liver) have resulted in a set of genes which have been found to have normal physiological variance across the six mice. A large set of these genes are missing from a previous analysis carried out on the same datasets [16]. Also our method was tested on a dataset [12], where similar analysis was conducted using Affymetrix oligo arrays. Our approach resulted in finding of many genes which exhibited normal variance among three identical mice. This analysis was done over four mouse tissues (brain, heart, liver, lung). We found a good number of genes which are known to be involved in stress response and hypertension. Some of the genes have been reported in previous studies as significantly differentially expressed, under different pathological conditions. This calls for further analysis of these genes, before reporting them as differentially expressed genes in all further studies, involving study of gene expression in mice using microarrays. Principal Component Analysis has been done to capture the variance in the data and provide a powerful tool to visualize the similarity among the mice and within the replicates.

2. METHODOLOGY

We used three datasets of kidney, liver and testis available at `http://www.pedb.org/MSM/NORMAL` provided by [16]. Six genetically identical male C57BL6 mice were used to compare the expression values of 5,406 unique mouse genes. Four separate microarray assays were conducted for each organ from each animal, for a total of 24 arrays per organ. For half of the replicate arrays, the experimental RNA was labeled with the Cy3 dye and the reference RNA with the Cy5 dye; for the other half, the labeling scheme was reversed to control for any dye-based bias. Also the dataset of [12] was used which contained the expression values for three mice across the four organs of liver, heart, lung and brain. Each experiment was replicated three times. Affymetrix oligo chips were used in these experiments.

In ideal conditions, the gene expression values for each gene should be the same across all array experiments. But due to the technical limitations the data contains lot of inherent noise, which could also be due to normal variation in expression of the genes across the genetically identical male mice. Our goal is to extract those genes which are contributing to the noise due to their biological variance.

In our paper, we try to capture the genes which show variance among the identical mice, by trying to eliminate the variations which come in due to experimental errors and fluctuations. We use a very robust method to exclude genes, which would eliminate any considerable variance in the replicates. Our approach is based on the following steps: 1) Calculation of fold-change ratio and discretization of expression levels for each gene, 2) Elimination of experimental noise, 3) Constructing an expression profile for each gene, and 4) Calculating and raking by gene variability via entropy calculation. We describe the steps in more detail below.

2.1 Fold-Change Ratio

We assume that we have n genes, in m mice, with r replicates for each mice, for a given tissue. We denote gene i as g^i. Let S_t^i denote the expression level for gene g^i in the test sample and S_r^i the expression of g^i in the reference microarray samples. We define *fold-change ratio* as the log-odds ratio of the expression intensities of the test sample over the reference sample, given as $\log_2(\frac{S_t^i}{S_r^i})$. To analyze the variability, we discretize the fold-change into k bins ranging from very low expression levels to very high expression levels. The data is normalized in such a way that the median of the deviation from the median was set to the same value for the distribution of all the log-ratios on each array [16]. Similar analysis was done for the Affymetrix data. The raw data containing the intensities was median centered and scaled by the standard deviation. This normalization technique was chosen after experimenting with other methods like linear regression and mean centering. Though none of these methods yielded a normal distribution for the histogram plot of the gene expression values of all clones in a sample, the median centered normalization technique performed the best and also provided a uniform distribution for our binning method.

If the number of bins for expression level discretization is too small or too high, then it leads to problems in analysis. In coarse binning, the information about the values is ignored, and in a very fine binning, the patterns are lost. We tried several values of k and found that $k = 5$ works well. The bin intervals are determined using the uniform frequency binning method. Other popular methods like discriminant discretization, boolean reasoning based and entropy based discretization can be considered [13]. In frequency binning method we discretized the relative expression (fold-change) into 5 levels depending on their expression value. The values of -1.5, -0.5, 0, 0.5, and 1.5 for the fold change ratio were taken as thresholds for very low (VL), low (L), normal (N), high (H), and very high (VH) expression, respectively. That is, $VL \in (-\infty, -1.5]$, $V \in (-1.5, -0.5]$, $N \in (-0.5, 0.5)$, $H \in [0.5, 1.5)$, and $VH \in [1.5, +\infty)$. We use the notation g_e^i to denote the expression level for gene g^i in a given replicate, where $e \in \{VL, L, N, H, VH\}$. The binning strategy used in the case of Affymetrix data was slightly different. Since the distribution was not symmetric the intervals too were optimized such that the number of genes falling within each bin was roughly the same. Histogram plots were constructed for all the samples to analyze for the distribution patterns.

	Rep1	Rep2	Rep3	Rep4
Mouse1	$\{g^1_{VH}, g^2_{VL}, g^3_{VH}, g^4_L\}$	$\{g^1_{VH}, g^2_{VL}, g^3_{VH}, g^4_N\}$	$\{g^1_{VH}, g^2_{VL}, g^3_{VH}, g^4_N\}$	$\{g^1_{VH}, g^2_{VL}, g^3_{VH}, g^4_N\}$
Mouse2	$\{g^1_{VH}, g^2_{VL}, g^3_L, g^4_N\}$	$\{g^1_{VH}, g^2_{VL}, g^3_L, g^4_N\}$	$\{g^1_{VH}, g^2_{VL}, g^3_L, g^4_N\}$	$\{g^1_{VH}, g^2_{VL}, g^3_H, g^4_N\}$
Mouse3	$\{g^1_{VH}, g^2_N, g^3_{VH}, g^4_N\}$	$\{g^1_{VH}, g^2_N, g^3_{VH}, g^4_N\}$	$\{g^1_{VH}, g^2_N, g^3_{VH}, g^4_N\}$	$\{g^1_{VH}, g^2_N, g^3_{VH}, g^4_L\}$
Mouse4	$\{g^1_{VH}, g^2_N, g^3_{VL}, g^4_L\}$	$\{g^1_{VH}, g^2_N, g^3_{VL}, g^4_N\}$	$\{g^1_{VH}, g^2_N, g^3_{VL}, g^4_N\}$	$\{g^1_{VH}, g^2_N, g^3_{VL}, g^4_N\}$
Mouse5	$\{g^1_{VH}, g^2_H, g^3_L, g^4_L\}$	$\{g^1_{VH}, g^2_H, g^3_L, g^4_L\}$	$\{g^1_{VH}, g^2_H, g^3_L, g^4_L\}$	$\{g^1_{VH}, g^2_H, g^3_L, g^4_L\}$
Mouse6	$\{g^1_{VH}, g^2_H, g^3_{VH}, g^4_L\}$	$\{g^1_{VH}, g^2_H, g^3_{VH}, g^4_L\}$	$\{g^1_{VH}, g^2_H, g^3_{VH}, g^4_N\}$	$\{g^1_{VH}, g^2_H, g^3_{VH}, g^4_N\}$

Table 1: The gene expression states of 4 genes in 24 (6 mice, 4 replicates) assays, with five possible levels: Very High (VH), High (H), Very Low (VL), Low (L) or Normal (N)

The expression levels were binned in the following 4 intervals $[-\infty, -0.1)$, $[-0.1, 0)$, $[0, 0.3)$, $[0.3, \infty]$.
Explaining the method for the first dataset with the help of an example, let us consider the expression of 4 genes in six mice with 4 array replicates in each, as shown in Table 1. The analysis of the second dataset also employs the same methodology, only that the initial binning step is slightly variant.

2.2 Elimination of Experimental Noise

In order to eliminate the noise due to experimental fluctuations, we process the data taking one mice at a time. For every mice the genetic expression signatures are obtained and compared across all r replicates. Only those genes which show consistent expression signature in all r replicates are chosen and the ones which show even a slight deviation in any of the replicates are eliminated. This methodology takes a very stringent approach towards eliminating even the slightest errors due to technical noise. One shortcoming of this approach is that it would not eliminate any genes which show high fluctuations in the range $(-0.5, 0.5)$. In our study of normal variance to identify genes which have been falsely reported as differentially expressed, the genes which we might fail to eliminate do not contribute to the databank anyway, because they lie in the normal expression range. So, our approach would eliminate most of the noise which comes due to technical/experimental issues. This operation is done on all m mice, as a result of which we have gene expression signatures in all the mice with minimal experimental noise.

	Gene Expression
Mouse1 (F_1)	$\{g^1_{VH}, g^2_{VL}, g^3_{VH}\}$
Mouse2 (F_2)	$\{g^1_{VH}, g^2_{VL}, g^4_N\}$
Mouse3 (F_3)	$\{g^1_{VH}, g^2_N, g^3_{VH}\}$
Mouse4 (F_4)	$\{g^1_{VH}, g^2_N, g^3_{VL}\}$
Mouse5 (F_5)	$\{g^1_{VH}, g^2_H, g^3_L, g^4_L\}$
Mouse6 (F_6)	$\{g^1_{VH}, g^2_H, g^3_{VH}\}$

Table 2: Gene expressions after elimination of experimental noise

Table 2, illustrates this process on our example data. For example consider Mouse1. Since gene g^4 is differentially expressed as L in replicate 1, but as N in the other three replicates, we eliminate g^4 from further consideration. The resulting expression signatures for Mouse1 and other mice from our example are shown in Table 2.

2.3 Gene Expression Profile

Let F_j represent the gene expressions of the j-th mice after the elimination of experimental noise. The F_j's contain the expression level (very high, high, normal, low, very low) information of each gene in each of the m mice in our example. Some values could be missing due to elimination in the first stage. The F_j values, for our example of six mice, are shown in Table 2.
From the F_j values we construct a frequency table, which contains the number of occurrences of each gene for each discretized expression level (VH, H, N, L, VL). The frequency of every distinct (gene g^i, expression level e) pair across all F_j, is used to populate the frequency table. The frequency for gene g^i and expression level e is given as $f^i_e = \sum_{j=1}^{m} \delta^i_e(j)$, where m is the number of mice, and $\delta^i_e(j)$ is a characteristic function that notes the presence/absence of gene g^i at level e in mouse j, defined as: $\delta^i_e(j) = 1$, if $g^i_e \in F_j$, and $\delta^i_e(j) = 0$, if $g^i_e \notin F_j$. The frequency table obtained for our example is shown in Table 3. As an example, g^2, has expression level VL in mice 1 and 2, level N in mice 3 and 4, and level H in mice 5 and 6. Thus the expression profile for g^2 is given by the vector $(0, 2, 2, 0, 2)$, as shown in the table.

	f^i_{VH}	f^i_H	f^i_N	f^i_L	f^i_{VL}
Gene g^1	6	0	0	0	0
Gene g^2	0	2	2	0	2
Gene g^3	3	0	0	1	1
Gene g^4	0	0	1	1	0

Table 3: Expression Profile: Frequency table for the four genes

2.4 Entropy-based Variability Ranking

The genes that show presence in more than one discrete level are of interest to us. The frequency table is analyzed further to identify those genes which show considerable variance by their presence in more than one state. To capture the variance in a gene's expression level, the entropy measure was used. Entropy gives us the amount of disorder in the expression values of a gene, and thus is a measure of the normal variance, since the noise due to experimental variation is eliminated prior to this step. The entropy measure for a gene g^i is given as follows, $E(g^i) = -\sum_{e=1}^{k} p^i_e \log_2(p^i_e)$, where k is the number of discrete expression levels, and p^i_e is the probability of gene g^i having expression level e, which is given as $p^i_e = \frac{f^i_e}{\sum_{j=1}^{k} f^i_j}$.
By definition of entropy, if a gene has only one expression level (say j), then $p^i_j = 1$ and $E(g^i) = 0$. On the other hand, if a gene has the most variance (i.e., equal occurrence at each expression level), then $P^i_j = 1/k$ for all expression levels j, and $E(g^i) = -\sum_{j=1}^{k} 1/k \log_2(1/k) = -\log_2(1/k) = \log_2(k)$. In our approach genes with entropy 0, i.e., those

having no variance in expression across the mice, are discarded, and the remaining genes are ranked in descending order of their entropy (and thus variance). The entropy ranking for the four genes (along with the probability of each expression level) in our example are shown in Table 4. Gene g^2 and g^3 are of most interest to us because they show variation in expression states across the six mice. On the other hand gene g^1 is always high in all six mice, showing no variance.

	p^i_{VH}	p^i_H	p^i_N	p^i_L	p^i_{VL}	Entropy
Gene g^2	0	0.33	0.33	0	0.33	1.59
Gene g^3	0.6	0	0	0.2	0.2	1.37
Gene g^4	0	0	0.5	0.5	0	1
Gene g^1	1.0	0	0	0	0	0

Table 4: Entropy-based gene variability ranking

2.5 Weighted Expression Profiles

In our approach to experimental noise elimination, any gene with varying expression level among the replicates is considered experimental noise, and eliminated. Instead of such a stringent approach, we can choose to retain a gene provided it has the same expression level in a given fraction of the replicates. For instance, gene g^4 has expression level N in three out of the four replicates for Mouse1. If we set our threshold to 75%, then we would retain g^4_N in the gene expression for Mouse1 in Table 2.

Another approach is to construct a weighted expression signature, as follows: For every gene we record the fraction of the replicates in which it takes a particular value. For instance, for Mouse1, gene g^1 always takes the value VH, so its weighted expression is $g^1_{VH(1.0)}$. On the other hand, gene g^4 is N in three and L in one out of the four replicates; we record its weighted expression as $g^4_{N(0.75),L(0.25)}$. We denote by $w^i_e(j)$ the weight of gene g^i at expression level e in Mouse j. Table 5 shows the weighted expression signatures for all the six mice (note: if the weight is 1.0 we omit the weight; we write g^1_{VH} instead of $g^1_{VH(1.0)}$).

	Gene Expression
Mouse1 (F_1)	$\{g^1_{VH}, g^2_{VL}, g^3_{VH}, g^4_{N(0.75),L(0.25)}\}$
Mouse2 (F_2)	$\{g^1_{VH}, g^2_{VL}, g^3_{H(0.25),L(0.75)}, g^4_N\}$
Mouse3 (F_3)	$\{g^1_{VH}, g^2_N, g^3_{VH}, g^4_{N(0.25),L(0.75)}\}$
Mouse4 (F_4)	$\{g^1_{VH}, g^2_N, g^3_{VL}, g^4_{N(0.75),L(0.25)}\}$
Mouse5 (F_5)	$\{g^1_{VH}, g^2_H, g^3_L, g^4_L\}$
Mouse6 (F_6)	$\{g^1_{VH}, g^2_H, g^3_{VH}, g^4_{N(0.5),L(0.5)}\}$

Table 5: Weighted gene expressions

From the weighted gene expressions, we can construct a weighted profile using the approach in Section 2.3. The weighted frequency for gene g^i and expression level e is given as $f^i_e = \sum_{j=1}^{m} w^i_e(j)$, where m is the number of mice. The weighted frequency table obtained for our example is shown in Table 6. As an example, g^4, has expression levels $N(0.75)$ in Mouse1, $N(1.0)$ in Mouse2, $N(0.75)$ in Mouse3 and Mouse4, and $N(0.5)$ in Mouse6. Thus $f^4_N = 0.75 + 1.0 + 2 \times 0.75 + 0.5 = 3.75$, and similarly $f^4_L = 2.25$. Thus the weighted expression profile for g^4 is given by the vector $(0, 0, 3.75, 2.25, 0)$, as shown in the table.

	f^i_{VH}	f^i_H	f^i_N	f^i_L	f^i_{VL}
Gene g^1	6	0	0	0	0
Gene g^2	0	2	2	0	2
Gene g^3	3	0.25	0	1.75	1
Gene g^4	0	0	3.75	2.25	0

Table 6: Weighted Expression Profile

From the weighted expression profile, we can derive the entropy-based variability ranking for each gene as shown in Table 7. Comparing with Table 4, we find that g^3 is ranked higher in terms of variability than g^2, but the overall trend is similar.

	p^i_{VH}	p^i_H	p^i_N	p^i_L	p^i_{VL}	Entropy
Gene g^3	0.5	0.04	0	0.29	0.17	1.64
Gene g^2	0	0.33	0.33	0	0.33	1.59
Gene g^4	0	0	0.62	0.38	0	0.95
Gene g^1	1.0	0	0	0	0	0

Table 7: Entropy-based gene variability ranking

2.6 Principal Component Analysis (PCA)

PCA [7] is a classical technique to reduce the dimensionality of the data set by transforming to a new set of variables (the principal components). It has been used in the analysis of gene expression studies. Principal components (PC's) are uncorrelated and ordered such that the k-th PC has the k-th largest variance among all PC's. The k-th PC can be interpreted as the direction that maximizes the variation of the projections of the data points such that it is orthogonal to the first $k-1$ PC's. PCA is sometimes applied to reduce the dimensionality of the data set prior to clustering. Using PCA prior to cluster analysis may aid better extraction of the cluster structure in the data set. Since PC's are uncorrelated and ordered, the first few PC's, which contain most of the variations in the data, are usually used in cluster analysis. Unless external information is available, [19] recommend cautious interpretation of any cluster structure observed in the reduced dimensional subspace of the PC's. They observe no clear trend between the number of principal components chosen and the cluster quality. [6] use PCA analysis for extracting tissue specific signatures.

In this paper we use PCA to analyze how well the genes we have extracted capture the normal variance between the mice, eliminating the variance due to any other sources to the maximum possible extent. Projection on to a 3-dimension space (the top three PCs) allows for better visualization of the entire data set. Figure 1 shows the arrangement of the samples by plotting them on the principal components derived from: 1) PCA analysis of all the genes in the dataset, and 2) PCA analysis of only those genes which were found to have normal variance across mice. These plots are shown for all the three tissues under study. Kidney and testis show non-random arrangement of the assay points while liver has less discernible patterns. As discussed in the next section, we observed that 1) the experimental replicates belonging to any single mice cluster close to each other, and 2) the mice (biological replicates) are also grouped into visible clusters. Pathologically similar mice are clustered together.

3. RESULTS

We applied our entropy-based method to detect normal variance in gene expression for the two datasets taken from [16] and [12]. Due to lack of space we give a detailed account only for the first dataset, and show some results on the latter in the Appendix.

3.1 Kidney Tissue

For the kidney tissue, Table 9 shows the genes that were previously unreported as having differential expression, Table 10 shows the list of genes that were previously reported as varying by [16] and also confirmed by our entropy-based approach, and Table 8 shows the genes that were reported as varying by [16] but not found by our analysis. We missed some previously known varying genes due to our stringent elimination of any gene that shows variable expression in different experiments. We verified that all these genes were dropped in our experimental noise elimination step.

UnigeneID	Name	Relative Expression
Mm.28100	1300002P22Rik	0.76
Mm.13859	1810055P16Rik	-1.42
Mm.159813	2210418O10Rik	0.31
Mm.157135	2610023M21Rik	-0.05
Mm.28386	3110001D19Rik	-1.04
Mm.23682	4930453N24Rik	-4.63
Mm.194940	AI194696	-2.62
Mm.22459	AI265322	0.89
Mm.447	AI447904	-1.94
Mm.163	B2m	-2.07
Mm.88793	Fga	-5.50
Mm.26991	H13	0.12
Mm.1192	Igj	1.00
Mm.20354	Klc1	0.13
Mm.18651	Lorsdh	0.63
Mm.1499	Pkib	-1.00
Mm.1779	Scp2	-1.02
Mm.46401	Son	-0.23
Mm.196591	Spi1-3	-5.52
Mm.29758	Tcp11	-5.15
Mm.1948	Tctex1	-3.87
Mm.14418	Tesp2	-3.35
Mm.18847	Wrnip	0.08

Table 8: Unconfirmed previously reported varying genes (Kidney)

In kidney tissue around 3.5% of the 3088 genes where found to be showing considerable variance across the six mice. As reported by [16] we found several immune modulated and stress responsive genes. Apart from Cish (reported previously), other hypertension related genes like Asah1 and Adrb2 were found to show differential expression across the six mice. Adrb2 and Asah1 are well known rat hypertension-associated homologue also known to be play a role in hypertension in humans. Also the gene Ezh2 (enhancer of zeste homologue 2 (Drosophila)) was found to be differentially expressed. Small interfering RNA (siRNA) duplexes targeted against Ezh2 reduce the amounts of Ezh2 protein present in prostate cells and also inhibit cell proliferation in vitro. Ectopic expression of Ezh2 in prostate cells induces transcriptional repression of a specific cohort of genes [14]. A gene from the solute carrier family Slc12a2, was found to be having high normal variance among the six mice. Many of the solute carrier genes are known to be involved in hypertension. Shoc2, involved in shock response and Itga3, a immune response related gene were also found to be differentially expressed. This could be due to the varying physiological state of the mice at the time of killing. These genes were also reported by [16]. Also Psmd7, a programmed cell death protein, was found to be variably expressed among the six mice.

3.2 Liver Tissue

In liver tissue 23 out of 2513 genes, show significant variation in their expression levels among the six mice. CisH and cyp4a14 were also found showing high variance in the liver tissues as we had observed in the kidney. CXADR, a target for experimental gene therapies for cancer [9], has been found to be considerably varying in gene expression among the mice. NeuroD2 was found to exhibit differential expression levels among the six mice. Its associated protein is necessary for development and survival of central nervous system neurons [15]. Over expression of Hoxb6 [8] produces early postnatal death with craniofacial and axial anomalies, especially in the cervical region. Hoxb6 is expressed in lung of 14 day old mouse embryos, declining in level with gestational age but still present at birth [3]. Hoxb6 is involved in transcription factor complexes and has found to be normally varying. Also Ncor2, a nuclear receptor co-repressor 2 was found to be variably expressing, It performs combinatorial roles of the nuclear receptor co-repressor in transcription and development. A heat shock protein, Hspa9a, was also found to be variably expressing. The liver was found to have the least amount of genes whose expressions levels varied considerably.

3.3 Testis Tissue

Out of the 3252 genes analyzed in the testis tissue, 63 were found to be showing differential expression levels across the six mice. R1p2, a normally varying gene in our analysis is known to be involved in the developmental stages of mouse [10]. Nr4a2 a nuclear receptor and RAC3 a co-activator for nuclear/steroid receptors have been found to be differentially expressing. [18] show that mouse RAC3 is expressed in a tissue-specific fashion and distributed mainly in the oocytes; they found that the steroid receptor co-activator SRC-3 (p/ CIP/ RAC3/ AIB1/ ACTR/ TRAM-1) is required for normal growth, puberty, female reproductive function, and mammary gland development. We hypothesize that RAC3 could play an important role in the development of testis as well. Map/Erk kinase 1 (MEK1) and MEK2 activate the Erk/MAP kinases and have been implicated in cell growth and differentiation. [1] have observed that MEK2 RNA message is expressed at high levels in all embryonic tissues examined, including all neural tissues, and liver. MEK1, on the other hand, is expressed at very low levels in most embryonic murine tissue but can be detected in developing skeletal muscle. RELNOC was another protein kinase which was found to be differentially expressing among the six mice. Cox8a is another important and well studied gene involved in electron transport which was found to be differentially expressing among the mice.

Importantly, many of the genes that we found to vary normally have been reported previously to be differentially expressed because of a pathological process or experimental

UnigeneID	Name	Expression	Entropy
Mm.5598	Adrb2	0.30	0.28
Mm.22747	AI415104	0.40	0.30
Mm.13148	AI450991	-0.49	0.30
Mm.22183	AI785303	0.39	0.24
Mm.14189	AI841487	-0.34	0.28
Mm.22547	Asah1	0.53	0.30
Mm.20722	AU022875	0.44	0.24
Mm.28639	AU041016	0.43	0.28
Mm.29987	B230114J08Rik	-0.56	0.28
Mm.24143	Bbp	0.47	0.24
Mm.25848	Bckdha	0.50	0.28
Mm.157101	C330007P06Rik	0.28	0.24
Mm.354	Calb1	1.46	0.28
Mm.34246	Calm	-0.05	0.29
Mm.2018	Cbfb	-0.55	0.30
Mm.10124	Chetk	-0.51	0.28
Mm.25836	Cldn8	1.43	0.28
Mm.22409	Clic4	0.54	0.30
Mm.2735	Cubn	1.48	0.28
Mm.21965	D5Ertd593e	-0.56	0.28
Mm.21103	D7Wsu105e	0.45	0.29
Mm.19726	Dnahc11	-0.35	0.30
Mm.56	Dscr1	0.37	0.22
Mm.140186	Eif4ebp2	0.42	0.28
Mm.5027	Ezh1	0.56	0.30
Mm.57075	Fau-ps3	0.20	0.30
Mm.28480	Fkbp3	-0.27	0.28
Mm.193539	H1f2	1.33	0.30
Mm.156892	Hnrpd	-0.50	0.30
Mm.2849	Hspa9a	0.38	0.24
Mm.14099	Hzf-pending	-1.55	0.28
Mm.28223	Idb4	0.50	0.24
Mm.29590	Idh3b	0.41	0.22
Mm.57035	Itga3	0.54	0.24
Mm.7362	Lmnb2	-1.47	0.30
Mm.4088	Ltc4s	0.48	0.30
Mm.2395	Mea1	0.50	0.28
Mm.8866	Mllt10	-0.46	0.28
Mm.19170	Mnpep-pending	-0.47	0.30
Mm.16366	Mtcp1	0.44	0.28
Mm.22508	Mtx	-1.40	0.30
Mm.57230	Neurod3	-0.10	0.22
Mm.1131	Npdc1	0.32	0.28
Mm.4918	Nr3c1	0.55	0.24
Mm.56948	Nt5	1.35	0.22
Mm.7952	Peg3	0.56	0.30
Mm.826	Pigf	0.99	0.30
Mm.18347	Psmd7	-0.42	0.22
Mm.2404	Ptpn16	0.92	0.24
Mm.180561	Rbpsuh	0.31	0.28
Mm.33376	Shoc2-pending	0.49	0.28
Mm.4168	Slc12a2	0.78	0.22
Mm.27330	Smarce1	0.37	0.22
Mm.10704	Snx12	0.60	0.24
Mm.154045	Tacstd2	1.43	0.28
Mm.112	Tcea3	0.56	0.30
Mm.2215	Tcof1	0.43	0.28
Mm.88645	Tes	-1.50	0.30
Mm.24096	Thbd	0.60	0.24
Mm.23959	Trim13	-0.66	0.28
Mm.10153	Twg-pending	0.54	0.24
Mm.21846	Ubl3	0.54	0.30
Mm.1298	Zfp36	0.63	0.28

Table 9: Previously unreported genes with normal variance (Kidney)

UnigeneID	Name	Expression	Entropy
Mm.101274	2010008E23Rik	-0.47	0.29
Mm.10826	Umod	1.35	0.22
Mm.1339	chgb	-0.40	0.28
Mm.13445	261000803Rik	1.40	0.24
Mm.14097	Tapbp	-0.48	0.30
Mm.1541	Snta1	0.28	0.28
Mm.15811	Bcl6	-0.18	0.30
Mm.17224	-	0.70	0.30
Mm.17353	-	-1.53	0.22
Mm.17974	-	0.29	0.22
Mm.18535	-	-0.70	0.30
Mm.18571	Brf1	0.25	0.22
Mm.19187	Ptma	0.62	0.28
Mm.19310	Cors-pending	1.46	0.28
Mm.19316	-	-1.84	0.28
Mm.200415	-	0.39	0.30
Mm.20046	Epb4.9	1.54	0.30
Mm.205791	-	1.35	0.30
Mm.208	1110060D06Rik	-0.10	0.22
Mm.21228	2610101J03Rik	-0.67	0.22
Mm.22513	kifc3	0.36	0.24
Mm.23452	-	-1.79	0.22
Mm.23473	-	0.64	0.30
Mm.23565	-	-1.97	0.24
Mm.23689	-	0.58	0.28
Mm.23853	-	0.69	0.24
Mm.24044	-	0.71	0.22
Mm.24108	-	0.58	0.30
Mm.24192	-	-0.34	0.29
Mm.24395	-	0.38	0.22
Mm.24529	1100001F19Rik	0.27	0.30
Mm.25120	-	-0.86	0.22
Mm.25497	-	-2.36	0.22
Mm.27302	-	0.41	0.22
Mm.27311	AI463227	0.42	0.22
Mm.27725	A2m	-3.72	0.24
Mm.27797	-	0.26	0.24
Mm.29381	Au021460	-1.42	0.28
Mm.29595	-	0.76	0.30
Mm.29932	-	-1.91	0.28
Mm.30227	-	-3.04	0.22
Mm.30266	Cnot7	0.03	0.28
Mm.30266	-	-0.06	0.20
Mm.30605	-	0.65	0.22
Mm.31748	-	1.24	0.24
Mm.31764	-	-2.62	0.22
Mm.31773	-	-1.53	0.22
Mm.31992	-	-1.91	0.24
Mm.32508	-	-2.08	0.22
Mm.32758	-	0.78	0.24
Mm.334	Solt	-1.08	0.30
Mm.34248	Dab2	1.55	0.29
Mm.4407	And	0.31	0.20
Mm.44199	-	0.32	0.24
Mm.4592	Cish	1.45	0.28
Mm.54120	-	-0.61	0.22
Mm.604	-	-1.69	0.24
Mm.6407	-	-0.75	0.22
Mm.75983	-	-1.95	0.22
Mm.87470	-	-2.31	0.22
Mm.9199	Pole2	-0.30	0.24

Table 10: Previously reported (common) varying genes (Kidney)

intervention. One recent study used microarrays to investigate the differential gene expression patterns during preimplantation mouse development [10]. Rpl12 was reported to be differentially expressed while we found it to be normally varying in the testis tissue. PUFA (polyunsaturated fatty acids) feeding can influence Protein Kinase C (PKC) activity [2]. Itrp1 is another gene which has been reported as differentially expressed [5] in papillary thyroid carcinoma, while we found this gene to be normally varying in kidney tissues. Another study investigated the effects of acetaminophen on gene expression in the mouse liver [17]. Eight of the genes reported to differ in response to acetaminophen, including CisH2, and Hsp40, were genes we found to vary normally.

3.4 PCA Analysis

Examining the PCA plots visually gives us a good idea of how our approach has faired (see Figure 1). In the case of capturing normal variance in kidney tissue, our method was able to separate the noisy data and group the assays into distinguishable clusters. By eliminating the experimental noise we expect the replicates to form a close group for each of the mice. In the case of the kidney tissue for genes with normal variance, the assays arrange into two clusters. One of the clusters has assays which include the replicates from four mice (M1, M2, M5, M6), while the other cluster has mice M3 and M4. This indicates that there is a high similarity among these mice in kidney tissue. Similar results were obtained after the analysis of testis tissue. However the same mice did not cluster together as those from the kidney analysis. This shows that the normal expression patterns are tissue specific and cannot be generalized on the whole. In the case of liver tissue no clear patterns were visible on performing the PCA analysis. To summarize, kidney gene expression patterns from the third and the fourth mice are fundamentally different from the other ones. In the testis, the first two mice are systematically different from the last four mice. No pattern was observed in liver. PCA can be additionally used as a platform to compare the performance of different methodologies to determine normal variance. The performance can be judged visually on the basis how well the replicates cluster together or measure the goodness of the clusters.

3.5 Other Datasets

We applied our entropy-based methodology on the Affymetrix dataset from [12]. Some of the genes showing normal variance for Lung tissue are shown in Table 11. Similar results were obtained for other tissues (Brain, Heart, Liver); we omit the results due to lack of space.

4. CONCLUSION

The methodology followed here has significant advantages over the ANOVA analysis [16] used on the same datasets. One of the main advantages being that this method is very robust to outliers. By binning the expression values we are minimizing the effect of a few outliers as far as possible. The possible use of even a single outlier skews the F-statistic thus giving a false positive for differentially expressing genes in normal identical mice. But the binning methodology seems to give an unfair advantage to those genes which fall in the center of the bins (more variance allowed), over the ones which lie close to the interval boundaries. The interval boundaries on inspection, revealed very few genes existing near the bin interval cuts. We found many more genes which showed higher variability than those reported in the existing database. Also it has been observed that the genes which were missing from our results were eliminated during the first step of pruning where genes showing the slightest experimental variation were eliminated. If the data is extremely noisy, containing high levels of both experimental and biological variance, or if the number of experiments is large, this pruning step might eliminate most of the genes. To avoid this, the weighted profile method can be used to characterize the variance. Also the underlying assumptions about the distributions of the populations and the independence of the samples, required by the ANOVA analysis, are not necessary for our methodology. Since our methodology makes use of the frequency distributions in the bins, the entropy measure calculated can be compared across various experimental setups. The dependence on the size of the experiments (number of experimental and biological replicates) is less compared to the F-statistic measure.

Our approach to identify genes which show significant variation in their expression levels among the six mice has resulted in a set of stress response and hypertension genes which where not found in a earlier work [16] . Sah, Cox, Cyp4a series of genes are well studied hypertension related genes. This has been observed in spite of stringent criteria to eliminate experimental noise. Also the percentage of the genes which show normal variance are found to be exactly the same as the ones obtained in the earlier work even though the approaches followed by the two authors is totally different. This was observed at significance levels of two fold for the log2 based ratios. Similar analysis was done with higher significance values which resulted in fewer numbers of genes to be normally varying. These set of genes represent stronger candidates showing normal variance in their expression levels. The approach though simple in nature seems to perform considerably well and there is around 65% overlap in the results of this approach and the earlier one which is considered as a benchmark for studying normal variance in mice. Also a vast number of normally varying genes genes where found from the [12] dataset in brain, heart and lung tissues. These are ranked on the basis of the entropy measure and depending on the user needs the database can be referred for top x% of the normally varying genes found. Approximately 4% of the 15,000 clone id's present on the chip seem to exhibit normal variance among the identical mice in all of these tissues. Quite a few of these genes have been reported as differentially expressed in gene expression studies under various pathological conditions. The authors suggest caution to investigators in the case where they observe these genes to be varying in their analysis. Further quantification studies need to be done before reporting them. Also the analysis can be conveniently modified to find the control genes which show negligible variance, which could be useful for normalization techniques which make use of the control genes. Also PCA could offer a powerful visualization tool to see which of the mice or the replicates are systematically similar to each other. We observed that the third and the fourth mice were very different from the other four in the kidney tissue. In the testis the first two mice were systematically different from the other four. In liver no clear pattern was discernible. This leads us to believe that normal variance is tissue specific.

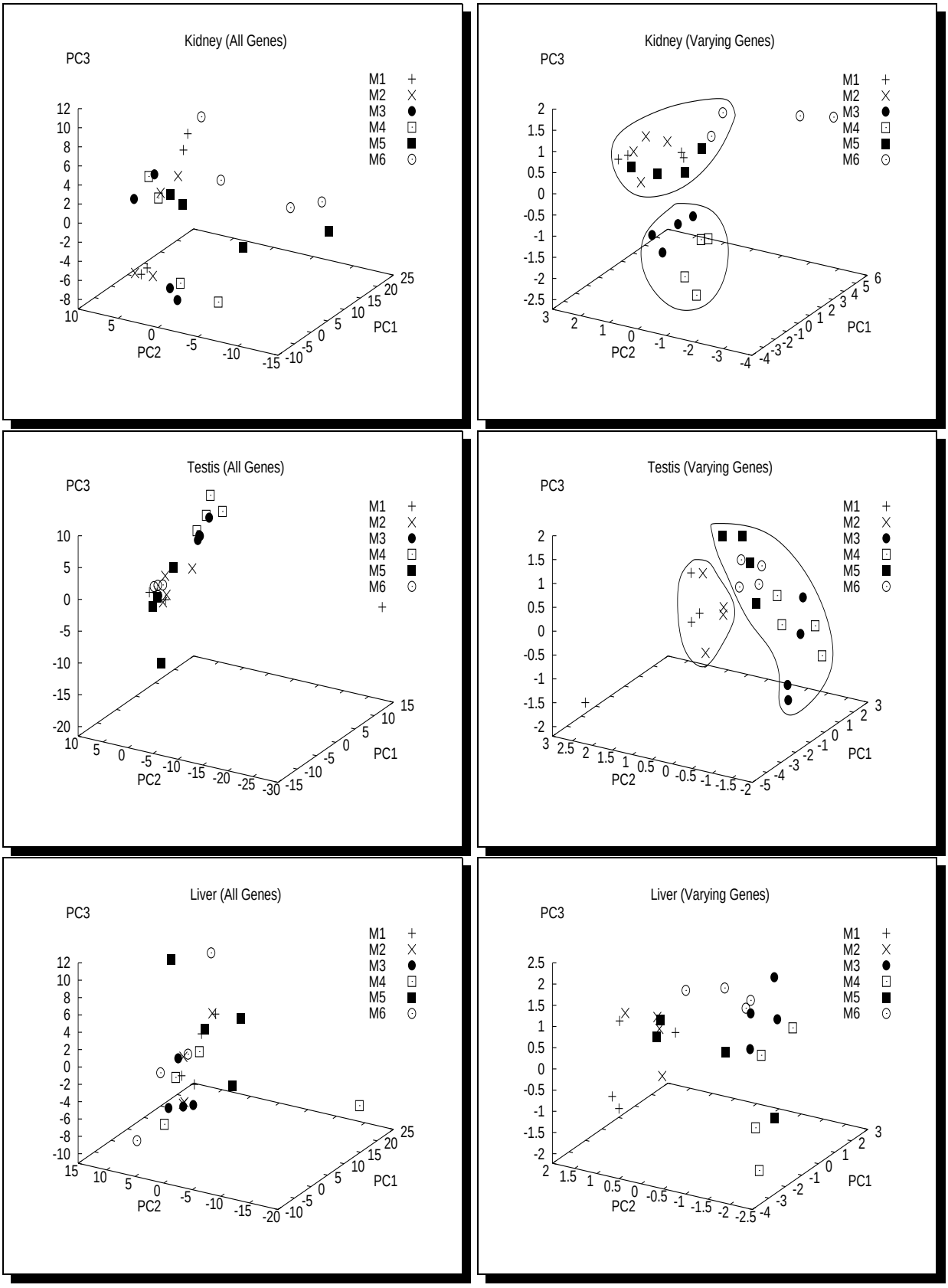

Figure 1: Principal component analysis of all genes (left column) and genes with normal variance (right column). Three tissues were studied: Kidney (top row), Testis (middle row), and Liver (bottom row). Results for all 6 mice and 4 replicates are shown

gene name	unigene id	biological process	molecular function	entropy
Bcl7c	Mm.2898			0.3
Cbx1	Mm.29055	chromatin assembly/disassembly	chromatin binding	0.3
Igfbp3	Mm.29254	regulation of cell growth	insulin-like growth factor binding	0.3
Eno3	Mm.29994	glycolysis	phosphopyruvate hydratase	0.3
Ptger3	Mm.30424	G-protein signaling, coupled to IP3 second messenger (phospholipase C activating)	protein binding	0.3
Scn8a	Mm.3076	adult walking behavior		0.3
Entpd2	Mm.31308	purine ribonucleoside diphosphate catabolism	apyrase	0.3
Itih1	Mm.3227		serine protease inhibitor	0.3
Madh5	Mm.33951	common-partner SMAD protein phosphorylation		0.3
Eln	Mm.34170			0.3
Rxra	Mm.3470	transcription regulation	ligand-dependent nuclear receptor	0.3
DXErtd242e	Mm.34942			0.3
Krtap6-3	Mm.3504			0.3
Gdf8	Mm.3514		cytokine	0.3
Il11	Mm.35814		cytokine	0.3
Rad54l	Mm.3655	DNA repair	DNA helicase	0.3
Pl2	Mm.37203		hormone	0.3
Slc2a3	Mm.3726	carbohydrate transport		0.3
Piga	Mm.3781	post-translational membrane targeting		0.3
Pou2f2	Mm.37811	transcription regulation	transcription factor	0.3
Elavl2	Mm.3823		nucleic acid binding	0.3
Prrx1	Mm.3869	developmental processes	DNA binding	0.3
Tcf15	Mm.3881	transcription regulation	transcription factor	0.3
Fgf15	Mm.3904	signal transduction	growth factor	0.3
Csnk	Mm.3975			0.3
Tbxas1	Mm.4054	prostaglandin metabolism	thromboxane-A synthase	0.3
4933406G12Rik	Mm.409			0.3
Xist	Mm.4095			0.3
Csnb	Mm.4105			0.3
Mc2r	Mm.41498	G-protein coupled receptor protein signaling pathway		0.3
Oprm	Mm.4191	G-protein signaling, adenylate cyclase inhibiting pathway		0.3
Mcpt8	Mm.41979	proteolysis and peptidolysis	serine-type endopeptidase	0.3
Myo1f	Mm.42019	cytoskeleton organization and biogenesis	calmodulin binding	0.3
Slc8a1	Mm.4211	calcium ion transport	calmodulin binding	0.3
Phka1	Mm.42254	glycogen metabolism	phosphorylase kinase	0.3
AA536748	Mm.4328			0.3
Nxph2	Mm.44246		receptor binding	0.3
Itgav	Mm.4427	integrin-mediated signaling pathway	cell adhesion receptor	0.3
Epha5	Mm.4466	transmembrane receptor protein tyrosine kinase signaling pathway	ephrin receptor	0.3
Impdh1	Mm.45234	purine nucleotide biosynthesis	IMP dehydrogenase	0.3
Tcf1	Mm.455	transcription regulation	transcription factor	0.3
2810417H13Rik	Mm.45765			0.3
Foxa1	Mm.4578	transcription regulation	transcription factor	0.3
Pvt1	Mm.4608			0.3
Chrna7	Mm.4611	synaptic transmission	GABA-A receptor	0.3
Snrpa	Mm.4633			0.3
Cdh3	Mm.4658	homophilic cell adhesion	calcium-dependent cell adhesion molecule	0.3
Tnfsf8	Mm.4664	immune response	cytokine	0.3
Irebf1-pending	Mm.470	transcription regulation	DNA binding	0.3
Htr1a	Mm.4716	G-protein coupled receptor protein signaling pathway		0.3

Table 11: Genes with normal variance in Lung tissue

5. REFERENCES

[1] A. Alessandrini, B.K. Brott, and R.L. Erikson. Differential expression of mek1 and mek2 during mouse development. *Cell Growth Differ*, 8:505–11, 1997.

[2] A. Berger, D.M. Mutch, J. Bruce, G. Matthew, and A. Roberts. Dietary effects of arachidonate-rich fungal oil and fish oil on murine hepatic and hippocampal gene expression. *Lipids in Health and Disease*, 1:2–10, 2002.

[3] C.W. Bogue, I. Gross, H. Vasavada, D.W. Dynia, C.M. Wilson, and C. Jacobs. Identification of hox genes in newborn lung and effects of gestational age and retinoic acid on their expression. *Am J Physiol*, 266:448–54, 1994.

[4] M. Bonin, S. Poth, D. Bhugon, and O. Riess. Dna-microarray technology defining the normal variance in mouse gene expression genome meeting. In *Genome Meeting*, July 2002.

[5] Y. Huang, M. Prasad, W.J.. Lemon, H. Hampel, F.A. Wright, K. Kornacker, V. LiVolsi, W. Frankel, R.T. Kloos, C. Eng, N.S. Pellegata, and A. de la Chapelle. Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *PNAS*, 98:15044–15049, October 2001.

[6] M. Jatin, W. Schmitt, D. Hwang, L.-L. Hsiao, S. Gullans, G. Stephanopoulos, and G. Stephanopoulos. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res*, 12:1112–1120, 2002.

[7] I.T. Jolliffe. *Principal Component Analysis, Springer Series in Statistics*. Springer Verlag, New York, 1986.

[8] S. Kaur, G. Singh, J.L. Stock, C.M. Schreiner, A.B. Kier, K.L. Yager, M.L. Mucenski, W.J. Scott, and S.S. Potter. Dominant mutation of the murine hox-2.2 gene results in developmental abnormalities. *J Exp Zool*, 264:323–36, 1997.

[9] M. Kim, K.R. Zinn, B.G. Barnett, L.A. Sumerel, V. Krasnykh, D.T. Curiel, and J.T. Douglas. The therapeutic efficacy of adenoviral vectors for cancer gene therapy is limited by a low level of primary adenovirus receptors on tumour cells. *Eur J Cancer*, 14:1917–26, 2002.

[10] M.S.H Ko, J.R. Kitchen, X. Wang, T.A. Threat, A. Hasegawa, T. Sun, M.J. Kargul, M.K. Lim, Y. Cui, Y. Sano, T. Tanaka, Y. Liang Y, S. Mason, P.D. Paonessa, A.D. Sauls, G.E. DePalma, R. Sharara, L.B. Rowe, J. Eppig, C. Morrell, and H. Doi. Large-scale cdna analysis reveals phased gene expression patterns during preimplantation mouse development. *Development*, 127:1737–1749, 2000.

[11] M.L.T. Lee, F.C. Kuo, G.A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cdna hybridizations. *PNAS*, 97:9834–9839, 2000.

[12] P.D. Lee, R. Sladek, C. Greenwood, and T. Hudson. Control genes and variability: Absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Research*, 12(2):292–297, February 2002.

[13] H. Midelfart, J. Komorowski, K. Norsett, F. Yadetie, A.K. Sandvik, and A. Legreid. Learning rough set classifiers from gene expression and clinical data. *Fundamenta Informaticae*, 53:155–183, November 2002.

[14] D. O'Carroll, S. Erhardt, Pagani, M. Barton, S.C. Surani, and T. Jenuwein. The polycomb-group gene ezh2 is required for early mouse development. *Mol Cell Biol*, 21:4330–6, 1997.

[15] J.M. Olson, A. Asakura, L. Snider, R. Hawkes, A. strand, J. Stoeck, A. Hallahan, J. Pritchard, and S.J. Tapscott. Neurod2 is necessary for development and survival of central nervous system neurons. *Developmental Biology*, 234:174–87, 2001.

[16] C.C. Pritchard, L. Hsu, J. Delrow, and P.S. Nelson. Project normal: Defining normal variance in mouse gene expression. *PNAS*, 98:13266–13271, 2001.

[17] T.P. Reilly, M. Bourdi, J.N. Brady, C.A. Pise-Masison, M.F. Radonovich, J.W. George, and L.R. Pohl. Expression profiling of acetaminophen liver toxicity in mice using microarray technology. *Biochem. Biophys. Res. Commun*, 282:321–328, 2001.

[18] J. Xu, L. Liao, G. Ning, H. Yoshida-Komiya, C. Deng, and B.W O'Malley. The coactivator src-3 (p/cip/rac3/aib1/actr/tram-1) is required for normal growth, puberty, reproductive function and mammary gland development. *PNAS*, 97:6379–6384, 2002.

[19] Yeung and Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763–774, 2002.

[20] A. Zien, J. Fluck, R. Zimmer, and T. Lengauer. Microarrays: How many do you need? In *RECOMB*, April 2002.