BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(15)

Insights from the clustering of microarray data associated with the heart disease

Venkatesan Perumal* & Vasantha Mahalingam

Department of Statistics, National Institute for Research in Tuberculosis (Formerly Tuberculosis Research Centre), Indian Council of Medical Research, Chennai-31, India; Venkatesan Perumal – Email: venkaticmr@gmail.com; *Corresponding author

Received June 17, 2013; Revised August 08, 2013; Accepted August 09, 2013; Published August 28, 2013

Abstract:
Heart failure (HF) is the major of cause of mortality and morbidity in the developed world. Gene expression profiles of animal model of heart failure have been used in number of studies to understand human cardiac disease. In this study, statistical methods of analysing microarray data on cardiac tissues from dogs with pacing induced HF were used to identify differentially expressed genes between normal and two abnormal tissues. The unsupervised techniques principal component analysis (PCA) and cluster analysis were explored to distinguish between three different groups of 12 arrays and to separate the genes which are up regulated in different conditions among 23912 genes in heart failure canines' microarray data. It was found that out of 23912 genes, 1802 genes were differentially expressed in the three groups at 5% level of significance and 496 genes were differentially expressed at 1% level of significance using one way analysis of variance (ANOVA). The genes clustered using PCA and clustering analysis were explored in the paper to understand HF and a small number of differentially expressed genes related to HF were identified.

Keywords: Microarray data, Cluster analysis, Principal component analysis, Heart failure, R.

Background:
Heart failure still remains as a major public health problem in the industrialised world, despite of significant improvement in the filed of diagnosis and medical therapeutics. Globally, the current prevalence of heart failure is over 23 million **[1].** In cardiovascular research, microarray technologies are in use to test the hypothesis about the molecular mechanisms underlying different pathological conditions and phenotypes and to identify new therapeutic targets. Human samples are subjected to many biological variations due to concomitant etiologies, medications, age, sex and clinical stage. So, the reproducibility is highly affected in case of human samples. There are number of studies on chronically instrumented dogs with high frequency cardiac pacing to study pathophysiological and molecular mechanism related to dilated cadiomyopathy **[2].**

In this paper, the microarray data set on pacing-induced heart failure model for dogs was considered **[2].** An analysis of microarray is a search for genes that have a similar or correlated pattern of expressions. The statistical aspects such as Analysis of Variance (ANOVA), Principal Component Analysis (PCA) and cluster analysis were used in the paper for analysing microarray data in canines (dogs). The main objective of the paper is to identify differentially expressed genes in three groups or classes using one way ANOVA test, to separate the genes which are up regulated in different conditions using principal component analysis and to identify the cluster of samples, cluster of genes, relationships between the samples and genes using cluster analysis.

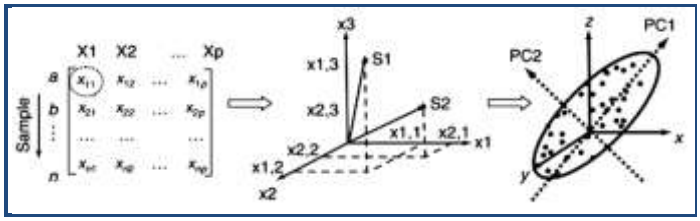


Figure 1: Schematic graph showing three dimensional data represented by two dimensional principal components, where matrix contain n rows and p columns.

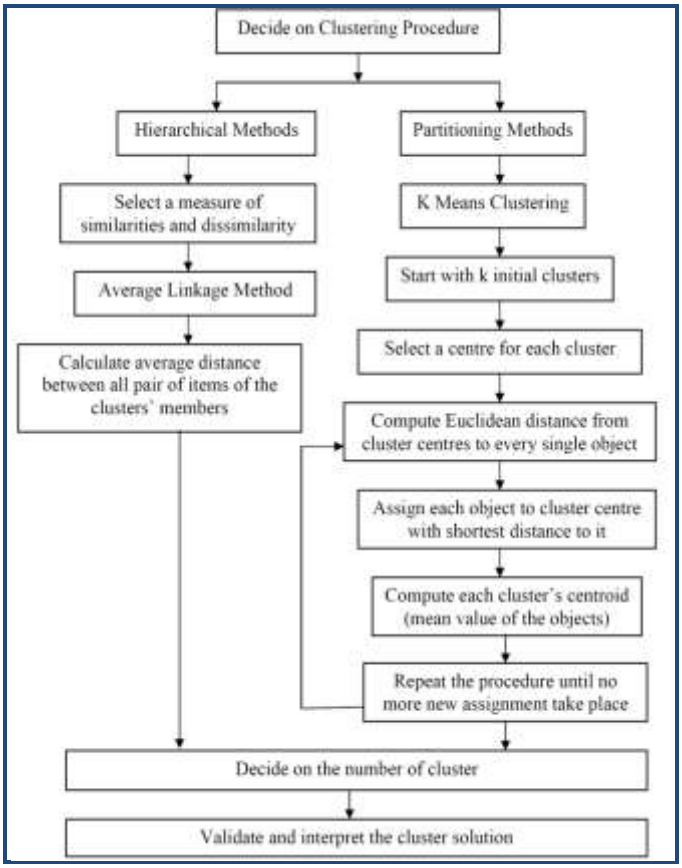


Figure 2: Flow chart explaining steps involved in different types of clustering.

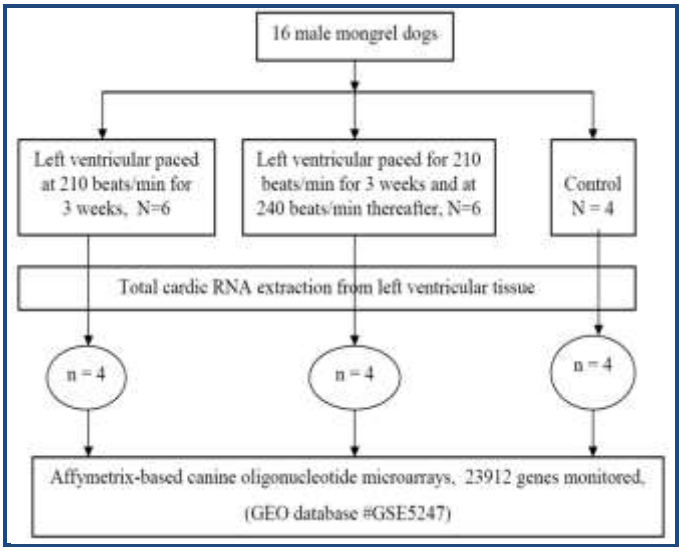


Figure 3: Flow chart describing the microarray experiment conducted on the heart failure model.

Methodology:

Principal Component Analysis and Cluster Analysis

Principal Component Analysis (PCA) is a variable reduction procedure and was derived by Karl Pearson in 1901. It is a classical tool to reduce the dimension of expression data, to visualise the similarities between the biological samples, and to filter noise. PCA is often used as a pre-processing step to clustering **[3].** The basic idea in PCA is to find the components that explain the maximum amount of variance in original variables by few linearly transformed uncorrelated components. **Figure 1** explains schematic diagram of three dimensional data represented by two dimensional principal components **[4].**

Clustering is widely used method in the first step of gene expression data analysis. The aim of cluster analysis is forming groups (clusters) of the objects on the basis of similarity (or distance) between the objects **[5].** It is used for finding correlated and functionally related groups. The most frequently used clustering techniques are Hierarchical clustering and K-means clustering. There are various methods to define the distance between clusters and the most widely used clustering is the average linkage method which works well with standardised microarray data. In Average linkage Clustering, average linkage defines the distance between the two clusters as the average distance between all pairs of items where one member of a pair belong to cluster 1 and other member of pair belongs to cluster 2 **[6].** The k-means clustering algorithm starts with a predefined number of cluster centers (k) specified by the user **[7] (Figure 2).**

Application to heart failure data

The data for the current study were obtained from Gene Expression Omnibus database at the National Centre for Biotechnology Information (GEO: http:// www.ncbi.nlm.nih.gov / geo/ GSE5247). The data consists of sixteen male mongrel dogs divided into three groups: the first group consists of 6 dogs subjected to left ventricular pacing at 210 beats/min for 3 weeks; the second group, 6 others paced for 210 beats/min for 3 weeks and at 240 beats/min thereafter; and the remaining four used as normal controls. Total cardiac RNA was extracted from control (n = 4), 3 wk-paced (n = 4), and decompensated heart failure dogs (n= 4) **[2].** Affymetrix-based canine oligonucleotide microarrays were used in the study to determine the changes in gene expression profile from compensated dysfunction to decompensated heart failure in pacing induced dilated cardiomyopathy. The Data set consists of 23,912 genes and 12 samples (arrays) **(Figure 3).** The open source software R version 2.10.0 is used for the microarray data analysis.

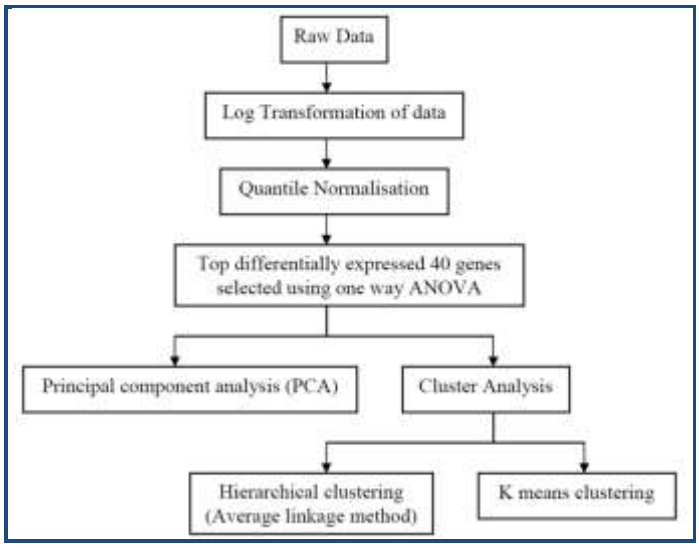
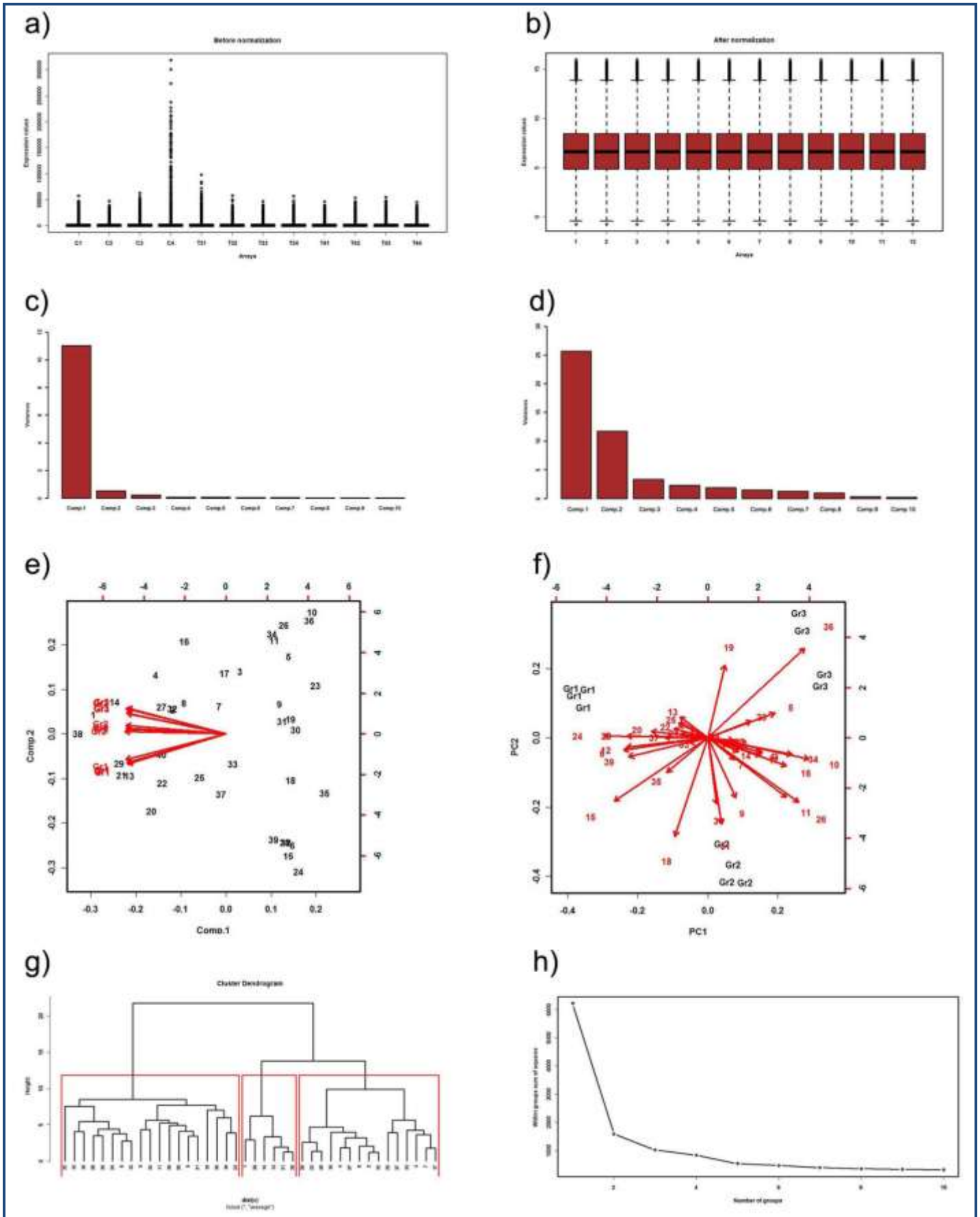


Figure 4: Flow chart describing statistical procedures for microarray data analysis.



Legend: Gr -Group, Comp - component, c1, c2, c3 & c4 - control groups, T31, T32, T33, T34 - 3 weeks pacing induced heart failure (T3) test groups, T41, T42, T43, T45 - 4 weeks pacing induced heart failure (T4) test groups, Hclust - Hierarchical clustering distance by average linkage method.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 9 (15): 759-765 (2013)

Figure 5: a) Box plots of all groups for raw data; **b)** Box plot of all groups for normalised data; **c)** Scree plot of arrays as variables; **d)** Scree plot of gene as variables; **e)** Biplot of arrays as variables; **f)** Biplot of genes as variables; **g)** Average linkage hierarchical clustering; **h)** K Means clustering.

Results & Discussion:

Normalisation of Heart failure data

Before applying statistical analysis, the normality of the gene expression data should be checked. The **Figure 4** shows step by step statistical procedure for doing the heart failure microarray data. The heart failure microarray raw data, were extremely positively skewed **(Figure 5a).** The raw data were processed using log transformation with base 2 and quantile normalized for 12 arrays in the three groups namely control, T3 and T4 to reduce variations across arrays **(Figure 5b).**

Analysis of Variance (ANOVA)

The exploratory microarray data analyses were carried out to shot list the differentially expressed genes in two or more known groups or classes. The one way ANOVA was used and the test was carried out in parallel for all the genes. It was found that out of 23912 genes, 1802 genes were differentially expressed in the three groups at 5% level of significance and 496 genes were differentially expressed at 1% level of significance. The top 40 most differentially expressed genes in the three groups were selected and corresponding p values were given the **Table 1 (see supplementary material).**

Principal Component Analysis (PCA)

For the selected 40 genes, PCA based on correlation matrix with samples as variables was performed. The first two PC from the table 2 explained more than 96% of the total variations. From the scree plot, it is observed that first two components are sufficient **(Figure 5c).** In the Biplot all control samples (C) labelled as Gr1 were grouped together, all T2 samples labelled as Gr2 were grouped together and all T3 groups labelled as Gr3 were grouped together. The angle between arrows of groups two and three was small, group 2 was between group 1 and group 3 as expected from definitions of the groups **(Figure 5e).** Gene with rank 14 and 1 is likely to be up regulated in groups (3 and 2) and down regulated in group1(control) whereas gene 29 is likely to be up regulated in group 1 and down regulated in groups (2 and 3). Gene with rank 14 codes for HSP40 (DNAJB6), which acts as a chaperone and plays a pivotal role during stress conditions and heart failure **[8].** Gene with rank 1 is an uncharacterised protein. Gene with rank 29 codes for ubiquitination factor E4A (UBE4A), which is involved in the degradation process of excess, unwanted and mis-fold proteins which is an important event to save the cells during heart failure **[9].** Similarly, for the 40 selected genes, PCA based on covariance matrix with genes as variables was performed. The number of variables was large and first two PC from the **Table 2 (see supplementary material)** explained 76% of the total variations.

From the scree plot, it is observed that first two components are sufficient **(Figure 5d).** From the biplot **(Figure 5f),** only two groups were identified among the genes and the genes which are up regulated and down regulated in groups 2 & 3 and the outlying genes in the groups are given in the **(Table 2)**. Genes such as PGRMC1, CYBB, AGPAT9 are playing major role in heart diseases. PGRMC1 and its homologues regulate cholesterol synthesis by activating the P450 protein Cyp51 which is an important target in cardio vascular disease. CYBB regulates NADPH oxidase activity and thereby protects from severe ischemia/reperfusion injury during HF. AGPAT9 is involved in phospholipid biosynthesis and its up regulation increases insulin resistance which is highly linked with cardio vascular disease. On molecular level it was not possible to distinguish between 3 weeks pacing induced heart failure and 4 weeks pacing induced heart failure however, these two groups were different from the control group. The genes which are up regulated in groups 2 and 3 can be used as biomarkers in case of heart failure models.

Cluster Analysis

For the selected 40 genes, three clusters were identified and correctly classified **(Figure 5g & Table 2 (see supplementary material))** for the three groups (control, T3 and T4) using cluster analysis by average linkage method with genes as variables. In cluster one, genes with rank 35, 15, and 18 were outlying genes, in cluster two, 1, 38 and 14 were outlying genes and in cluster three, 25, 37, and 33 were outlying genes. Gene with rank 35 codes for kininogen which is involved is blood coagulation system. Kinins, peptide products of kininogens may be involved in hypertensive and diabetic diseases **[10].** As already mentioned in PCA results, Genes with rank 1 and 14 play a major role in heart failure models. The other outlying genes need to be studied further. The **(Figure 5h)** represents the scatter plot of k means clustering where the two elbows indicate the two possibilities viz. 2 - cluster solution and 3 - cluster solutions.

Conclusion:

Oligonucleotide or cDNA micorarray has also been used in cardiovascular disease related gene changes **[11].** Microarray experiments using two colour comparisons has potential pitfall for data analysis. We don't measure gene expression directly, but rather fluorescence intensity by a scanner. Many factors influence the intensity and produce multicentric effect, creating a need for bias correction or normalisation between two colour systems. We have demonstrated in this work, methods for selecting differentially expressed genes. In conclusion, an approach to select differentially expressed genes in cDNA microarray was proposed and studied. This offers basis for advanced datamining approaches.

There are limited statistical methods to deal with multidimensional data. The same microarray data set can lead to very different conclusions by using different data analysis techniques and different clustering algorithms **[12].** The group of samples, gene clusters, outlying genes and the relationship between the samples and genes were analysed using PCA and cluster analysis. Ojaimi et al compared found that a number of processes including normalization of gene regulation during decompensation, appearance of new up regulated genes and maintenance of gene expression all contribute to the transition to overt heart failure with an unexpectedly small number of genes differentially regulated **[2].** In this paper, three group comparisons were done simultaneously. The genes such as PGRMC1, CYBB, AGPAT9, DNAJB6, UBE4A and KNG1 are

differentially expressed in the canine heart failure model. These genes may also be critically involved in human heart failures. Hence, further studies should be done to understand the role of these genes in the etiology of heart failure cases.

Acknowledgement:
The authors thank Ms. Shainaba A.S., Research Scholar, Department of Bacteriology, National Institute for Research in Tuberculosis for helping to understand the biological background of the work.

References:

[1] Bui AL *et al. Nat Rev Cardiol*. 2011 **8:** 30 [PMID: 21060326]
[2] Ojaimi C *et al. Physiol Genomics*. 2007 **29**: 76 [PMID: 17164392]
[3] Yao F *et al. BMC Bioinformatics*. 2012 **13**: 24 [PMID: 22305354]
[4] http://www.sciencedirect.com/science/article/pii/S0065308X10730128#f0025
[5] Selvaraj S & Natarajan J, *Bioinformation*. 2011 **6:** 95 [PMID: 21584183]
[6] Stone M *et al. Comput Methods Biomech Biomed Engin*. 2010 **13:** 493 [PMID: 20635265]
[7] Lee H *et al. Schizophr Res*. 2011 **128**: 143 [PMID: 21353765]
[8] Hayashi M *et al. Nat Med*. 2006 **12:** 128 [PMID: 16327803]
[9] Kaneko-Oshikawa C *et al. Mol Cell Biol*. 2005 **25:** 10953 [PMID: 16314518]
[10] Fong D *et al. Hum Genet*. 1991 **87:** 189 [PMID: 2066106]
[11] Hwang JJ *et al. Physiol Genomics*. 2002 **10:** 31 [PMID: 12118103]
[12] Yeung KY & Ruzzo WL, *Bioinformatics*. 2001 **17:** 763 [PMID: 11590094]

Edited by P Kangueane
Citation: **Perumal & Mahalingam,** Bioinformation 9(15): 759-765 (2013)

BIOINFORMATION

open access

Supplementary material:

Table 1: Top 40 most differentially expressed genes

Gene No	Gene ID	Rank	Gene Name	F statistics	p value
23901	1606183_at	1	Transcribed locus	108.02	3.59e-05
4294	1586576_s_at	2	PGRMC1	102.04	5.27e-05
1468	1583750_at	3	GOLGA2	93.90	7.40e-05
23339	1605621_at	4	Transcribed locus	73.99	1.03e-04
6454	1588736_at	5	Transcribed locus	82.86	1.03e-04
19265	1601547_at	6	Transcribed locus	87.25	1.33e-04
20115	1602397_at	7	CYBB	64.53	1.79e-04
5370	1587652_s_at	8	Transcribed locus	49.93	1.91e-04
9606	1591888_at	9	AGPAT9	63.20	2.06e-04
41	AFFX-r2-Bs-phe-3_at	10	Control	50.61	2.46e-04
6264	1588546_s_at	11	SAR1A	62.45	2.55e-04
14549	1596831_at	12		53.63	2.77e-04
414	1582696_s_at	13	TSHB	48.92	2.82e-04
16289	1598571_s_at	14	DNAJB6	55.89	3.25e-04
14674	1596956_at	15	Transcribed locus	40.77	3.36e-04
12521	1594803_at	16	Transcribed locus	42.73	3.43e-04
11935	1594217_s_at	17	Transcribed locus	39.54	3.69e-04
15350	1597632_at	18		44.14	3.96e-04
7009	1589291_at	19	Transcribed locus	49.79	4.07e-04
3870	1586152_at	20	IVNS1ABP	47.78	4.51e-04
8084	1590366_s_at	21	LOC478421 (CI-9KD)	47.41	4.55e-04
18499	1600781_at	22	ATP5C1	38.37	4.63e-04
9637	1591919_at	23	FEZF2	36.69	4.79e-04
12898	1595180_at	24	Transcribed locus	38.51	4.91e-04
13675	1595957_at	25		39.70	4.92e-04
9738	1592020_s_at	26	SYPL1	47.48	4.93e-04
4585	1586867_at	27	Transcribed locus	37.87	5.08e-04
593	1582875_at	28	RPGR	46.08	5.23e-04
12525	1594807_at	29	UBE4A	36.90	5.56e-04
8811	1591093_at	30	Transcribed locus	42.95	5.62e-04
10296	1592578_at	31	Transcribed locus	39.71	5.74e-04
2565	1584847_at	32	Transcribed locus	41.19	5.83e-04
15615	1597897_at	33	Transcribed locus	61.18	6.50e-04
20018	1602300_at	34		35.60	6.61e-04
10130	1592412_at	35	KNG1	42.99	6.67e-04
22138	1604420_s_at	36	PHF10	39.55	6.69e-04
4438	1586720_at	37	PERP	40.13	6.89e-04
4863	1587145_s_at	38	RPS4X	40.45	6.92e-04
11129	1593411_at	39	PCMD2	39.37	6.92e-04
21810	1604092_at	40	SNX5	37.47	7.01e-04

Legend:
Gene No - Gene number from the data base ; Gene Id - Id from Platform data table;
Gene Name - Single letter gene name , F Statistics - derived from one way ANOVA, p value - indicates the probability of getting a mean difference between the groups as high as what is observed by chance. The lower the p-value, the more significant the difference between the groups.

Table 2: Results of Principal Component (PCA) and cluster analysis

Variables		Principal Components				
		1	2	3	4	5
	Standard deviation	3.3203	0.7217	0.4712	0.2517	0.2318
Arrays	Proportions of variance	0.9187	0.0434	0.0185	0.0053	0.0045
	Cumulative proportion	0.9187	0.9621	0.9806	0.9859	0.9904
Genes	Standard deviation	5.0670	3.4280	1.8331	1.5116	1.3754
	Proportions of variance	0.5210	0.2380	0.0682	0.0464	0.0384
	Cumulative proportion	0.5210	0.7600	0.8277	0.8741	0.9125
Arrays	Upregulated in control and					9

	down regulated in groups 2 & 3	
	Upregulated in groups 2 & 3 and down regulated in control	14,1
Genes	Upregulated in control and down regulated in groups 2 & 3	13, 29, 25, 40, 22, 21, 20, 37, 28, 24, 33, 12, 6, 39, 35, 15 and outlying gene in the category - 18
	Upregulated in groups 2 & 3 and down regulated in control	23, 5, 38, 32, 27, 8, 1, 2, 14, 34, 10, 17, 16, 7, 11, 26, 30, 31, 9, 3, 4 and outlying genes in the category - 19, 36

Genes	Cluster	Member
Cluster	1	35, 15, 18, 28, 39, 24, 6, 12, 5, 34, 11, 30, 26, 9, 31, 19, 36, 10, 23
Member	2	1, 38, 14, 13, 21, 29
(Average Linkage)	3	20, 22, 40, 16, 4, 27, 8, 2, 32, 25, 37, 33, 3, 7, 17