PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
15th May 2011. Vol. 27 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GENE EXPRESSION ANALYSIS FOR TYPE-2 DIABETES MELLITUS – A STUDY ON DIABETES WITH AND WITHOUT PARENTAL HISTORY

[1]CHANDRA SEKHAR VASAMSETTY, [2]Dr. SRINIVASA RAO PERI, [3]Dr. ALLAM APPA RAO, [4]Dr. K. SRINIVAS, [5]CHINTA SOMESWARARAO

[1]Department of CSE, S R K R Engineering College, Bhimavaram – 534204, W.G.Dt., A.P. India,

[2]Professor in CS&SE, AU College of Engineering, Visakhapatnam - 530003, Andhra Pradesh , India
[3]Vice Chancellor, JNT University Kakinada, AP, India.
[4]Professor, Department of CSE, VR Siddartha College of Engg.,Vijayawada,AP, India.

[5]Department of CSE, S R K R Engineering College, Bhimavaram – 534204, W.G.Dt., A.P. India,

ABSTRACT

Diabetes mellitus, simply referred to as diabetes, is a group of metabolic diseases. There exits more than one type of diabetes, with each type having its own risks. Among the different types, types 1 and 2 are the most common ones. The cause of diabetes depends on the type. In each case, combinations of genetic and environmental influences are responsible for causing diabetes. Type 2 diabetes is primarily due to lifestyle factors and genetics. Microarray analysis is a method for analyzing expression levels of multiple genes at once. This method is especially suitable for identifying and classifying genes whose expression level differs in two samples. The present work focuses on identifying and classifying genes that cause type-II diabetes with two different samples, one with parental history and other without parental history. Mahalanobis Distance, Minimum Co-variance Determinant are the statistical methods used for identifying multivariate and univariate outliers for the identified inflammatory genes, the functional classification is performed by using Gene Ontology and pathway analysis. It is observed that 38 differentially expressed genes were identified out of 39400 genes tested between diabetes with and without parental history.

Keywords: *Type-2 Diabetes mellitus, Mahalanobis Distance, Gene Ontology, pathway analysis, Microarray analysis.*

1. INTRODUCTION

Diabetes is a chronic disease that is associated with considerable morbidity and mortality. Recent studies revealed that the incidence of diabetes mellitus is assuming epidemic proportions both in the developing and developed world. This has been attributed largely to westernized life style pattern. In view of this increasing incidence of diabetes, it is imperative that more sophisticated, fast, reliable and robust methods to need to be devised to develop the best use of information science and technology in relation to diabetes, decision support and clinical management. Molecular Biology research involves in this area through the development of the technologies used for carrying them out. DNA Microarray is one such technology which enables the researchers to investigate and address issues which were once thought to be non traceable. One can analyze the expression of many genes in a single reaction quickly and in an efficient manner [1]. DNA Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body [2].

A typical microarray experiment involves the hybridization of an mRNA molecule to the DNA template from which it is originated. Many DNA samples are used to construct an array. The amount of mRNA bound to each on the array indicates the expression level of the various genes.

1.1. Microarray Technology

The scientific principle underlying microarray technology is complementary hybridization between nucleic acids [3,4]. All gene expression DNA microarrays can be understood as high-throughput 'dot-blot' systems, where pieces of known DNA are anchored to a solid support, while targets are fluorescently labeled, free-floating

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

amplified RNA or complementary DNA (cDNA) species originating from the samples [5]. When the labeled sample is hybridized to the DNA microarray, each probe binds its complementary target. Analyzing the microarray with high-resolution fluorescent scanners allows assessment of the fluorescent signal strength that originates from the probe-bound target. This signal is presumed to be directly proportional to the abundance of the RNA species present in the investigated samples.

Most commonly, one sample is hybridized to one array, and the results are standardized and mathematically compared across microarrays, uncovering fluorescent intensity differences between microarrays. These intensity differences correspond to transcript abundance differences between samples. Current microarrays contain probes corresponding to many thousands of annotated genes in the human genome, allowing 'transcriptome profiling' from each of the samples. A variety of microarray platforms are available [6]. The choice of platform depends on factors like the number of genes represented on the array, cost and availability.

However, microarrays are more than a simple collection of independently performed dot blots for thousands of genes. Many expression changes are correlated in ways that suggest a causal dependence [7], and changes in relative abundances within individual samples provide valuable information about the complex pathways and cellular processes that are altered in a particular disease or condition. Furthermore, defined disease-specific expression patterns can be correlated with relevant pre-mortem information [8].

*1.1.1. **Types of Microarrays**:*

Depending upon the kind of immobilized sample used construct arrays and the information fetched, the Microarray experiments can be categorized in three ways:

*Microarray expression analysis***:** In this experimental setup, the cDNA derived from the mRNA of known genes is immobilized. The sample has genes from both the normal as well as the diseased tissues. Spots with more intensity are obtained for diseased tissue gene if the gene is over expressed in the diseased condition. This expression pattern is then compared to the expression pattern of a gene responsible for a disease.

*Microarray for mutation analysis***:** For this analysis, the researchers use gDNA. The genes might differ from each other by as less as a single nucleotide base. A single base difference between two sequences is known as Single Nucleotide Polymorphism (SNP) and detecting them is known as SNP detection

*Comparative Genomic Hybridization***:** It is used for the identification in the increase or decrease of the important chromosomal fragments harbouring genes involved in a disease.

In this paper we used Microarray expression analysis for identifying and classifying genes causing Type 2 Diabetes Mellitus (T2DM). The prevalence of T2DM is rising worldwide. While environmental factors, such obesity and lack of physical activity, play an important role to the rapid increase in the prevalence of T2DM, genetic factors are also important for the increased risk of T2DM. Studies have estimated that risk for diagnosed T2DM increases approximately two- to fourfold when one or both parents are affected.

A lot of studies have showed an excess paternal transmission of T2DM in different populations. Genetic factors, such as mitochondrial DNA mutations, and environmental mechanisms, such as intrauterine environment, have been proposed for the explanation of the excess paternal transmission of T2DM. In the present study, we performed gene expression profile in subjects with type 2 diabetes mellitus with parental history Versus Healthy using micro array data.

Searching for all of the available information about each gene of interest is very time consuming. This is hampered further by the wide variations in terminology. Gene Ontology (GO) is a collection of controlled vocabularies describing the biology of a gene product in any organism

*1.2. **Normalizing DNA Microarray Data***

Normalization is a broad term for methods that are used for removing systematic variation from DNA microarray data. In other words, normalization makes the measurements from different arrays inter-comparable. The methods are largely dissimilar for different DNA microarray technologies. For example, robust multiparty average (RMA) is a commonly used method for preprocessing and normalizing Affymetrix data, but it can't be applied to any other data types. However, one part of the RMA method is quantile normalization that is applicable to all data types.

Typically log2-transformed data is used for further analysis. Most of the normalization functions produce data in this format by default. If this is not the case, it is indicated below after the normalization. After normalization and possible log-transformation, the data is saved in a tabular format for further analysis.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In carrying out comparisons of expression data using measurements from a single array or multiple arrays, the question of normalizing data arises. In this study we will consider Lowess Normalization method for normalization.

The global locally weighted scattered plot smoothing (LOWESS) normalization is a good choice because it provides a good balance on the following three factors: ideally the center of the distribution of log-ratios should be zero, the log-ratios should be independent of spot intensity, and the fitted line should be parallel to intensity axis. It has been reported that the log2(ratio) values can have a systematic dependence on intensity, which most commonly appears as a deviation from zero for low-intensity spots. Locally weighted linear regression (LOWESS) analysis has been proposed as a normalization method that can remove such intensity-dependent effects in the log2(ratio) values (see M/A plots below). The easiest way to visualize intensity-dependent effects is to plot the measured log2(red/green) ratio or (M) for each element on the array as a function of the log2(red*green) product intensities or (A). LOWESS method detects systematic deviation in the "ratio-intensity" plot and corrects them by carrying out a local weighted linear regression as a function of the log2(intensity) and subtracting the calculated best-fit average log2(ratio) from the experimentally observed ratio for each data-point.

1.3. Data Analysis Challenges

There are three main phases to microarray data analysis: pre-processing, inferential statistics and descriptive statistics. These phases of analysis are used to answer some of the key questions typically posed by biologists using microarrays.

Gene expression variations between samples are attributable to a combination of biological differences and experimental artefacts. The latter include variations associated with the sample (differences in the way RNA samples are isolated and processed, or different labeling efficiencies of fluorescently tagged nucleotides), the array (uneven spotting of DNA elements) or the hardware (variable performance of fluorescence scanners). Non-biological variations in gene expression can be reduced through proper experimental design (for example, by processing experimental and control samples in parallel, using microarrays from the same lot or by using dye swap experiments). Data normalization can also be applied to remove systematic biases in the data [9]: global normalization allows the comparison of data from two microarrays, and local normalization accounts for artifactual variations that are not constant across a range of signal intensities or across the surface of a microarray [10].

A fundamental challenge for researchers using microarrays is that there is currently no consensus for the appropriate data normalization procedures. We believe that the distribution of data should be normalized around zero, and local normalization procedures should be applied to datasets to account for gene expression values that change as a function of signal intensity. However, such corrections are not consistently applied. If three normalization procedures are applied to the same raw data set, it is likely that entirely distinct descriptions of regulated genes will be generated.

The goal of the second phase, inferential statistics, is to evaluate hypotheses about gene expression changes in terms of significance and confidence. We may state the null hypothesis that a given gene is not differentially regulated in five brain samples from individuals with schizophrenia relative to a similar number from matched controls, and then test whether we can reject it with a probability $P < 0.05$. However, there is little consensus on how the significance of gene expression changes should be applied.

2. RELATED WORK

Microarray experiments are now being used to profile expression levels of genes under changing experimental conditions. To analyze these profiles in an attempt to answer diverse biological questions, various techniques and ideas have been proposed. Of particular interest to many scientists is the identification of genes whose expression profiles are similar, since genes with similar cellular functions have been theorized to respond similarly to changing conditions [11]. As a result, an efficient similarity measure for microarray analysis is fundamental for understanding the cellular processes [12] and annotating unknown genes.

There has been a growing interest in linking genes whose expression profiles are similar to construct co-expression networks. These networks and their highly modular sub networks are invaluable sources of information for system-level gene processes [13,14]. Similarity of two genes can be deduced from expression levels of these genes across all samples [15, 13, 16]. However, the noise inherent in microarray datasets limits the sensitivity of such analysis. Since any microarray measurement is likely to fluctuate due to many possible sources of error, a similarity based solely on expression measurements of two genes is more error-prone

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

than a similarity based on expression measurements of many genes. In addition, inferring the similarity of two genes based on their relations with a set of other genes will be in accordance with the biological hypothesis about gene products acting as complexes to accomplish certain cellular level tasks [17]. Thus, here we investigate use of extrinsic similarity measures to analyze microarray studies.

M. Kathleen Kerr et.al demonstrated [18] that ANOVA methods can be used to normalize microarray data and provide estimates of changes in gene expression that are corrected for potential confounding effects. This approach establishes a frame work for the general analysis and interpretation of micro array data.

The probability that a false identification is committed can increase sharply when the number of tested genes gets large. Correlation between the test statistics attributed to gene co-regulation and dependency in the measurement errors of the gene expression levels further complicates this problem. Anat Reiner et.al addressed [19] this problem by adapting the False Discovery Rate (FDR) controlling approach. Comparative analysis shows that all the four FDR controlling procedures control the FDR at the desired level.

D. L. Wilson et.al presented [20] two methods for the normalization of the micro array data to remove biases towards one or the other fluorescent dyes used to label each mRNA sample allowing for proper evaluation of differential gene expression. One method deals with smooth spatial trends in intensity across micro arrays. Second method deals with normalization of a new type of cDNA micro array experiment where large proportion of the genes on the microarrays is expected to be highly differentially expressed.

Hong-Ya Zhao et.al applied [21] a multivariate mixture model to model the expression level of replicated arrays, considering the differentially expressed genes as the outliers of the expression data. In order to detect the outliers of the multivariate mixture model, a statistical method based on the analysis of Kurtosis Coefficient (KC) is applied to the micro array data. They used the RT-PCR method and two statistical methods, Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE) to verify the expression levels of outlier genes identified by KC algorithm.

Dan Nettleton et.al developed [22] a non parametric multivariate method for identifying gene categories whose multivariate expression distribution differs across two or more conditions. By comparing the performance to several existing procedures via the analysis of a real data set and showed that this method has good power for differentiating between differentially expressed and non- differentially expressed gene categories.

Huaizhen Qin et.al proposed [23] a computationally simple method for finding differentially expressed genes in small micro array experiments. This method incorporates a novel stratification based tight clustering algorithm, principal component analysis and information pooling. They applied this method to three real micro array data sets. Comprehensive simulation shows that this method is substantially powerful than the popular SAM and eBayes approaches.

Bogdan Done et al proposed [24] a technique that improves previous method for predicting novel GO annotations by extracting implicit semantic relationships between genes and functions. In this work, they use a vector space model and a number of weighting schemes. The technique described is able to take into consideration the hierarchical structure of the Gene Ontology (GO) and can weight differently GO terms situated at different depths.

Purvesh Khatri et al proposed [25] an impact analysis approach that considers crucial biological factors to analyze regulatory pathways at systems biology level. This approach calculates perturbations induced by each gene in a pathway, and propagates them through the entire pathway to compute an impact factor for the given pathway. They proposed an alternative approach that uses a linear system to compute the impact factor. Their proposed approach eliminates the possible stability problems when the perturbations are propagated through a pathway that contains positive feedback loops. Additionally, the proposed approach is able to consider the type of genes when calculating the impact factors.

3. ANALYSIS PERFORMED

Data from three samples were hybridized on Human 40 K OchiChip Array. Gene expression values were obtained after quantification of TIFF images. Data has 40,320 X 3 data-points (or probes). Empty spots and control probes were removed before proceeding with data analysis.
Analysis process involved:

1. Differential expression analysis.
2. Functional classification of differentially expressed genes.

3.1. Differential Expression Analysis

In any micro array study the primary objective is to assess mRNA transcript levels of samples

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

under different experimental conditions. Which of the thousands of genes show significant difference in expression levels in the samples is the question of importance. Appropriate statistical techniques are required to furnish the accurate information on differentially expressed genes if there are no or limited replicates due to practical constraints in majority of the experiments.

For experiments with single sample in different conditions, we assume that the log intensity values of gene expression for the two samples are linearly related, following bivariate normal distribution, contaminated with outliers. In a contaminated bivariate distribution, the main body of the data is characterized by bivariate normal distribution and constitutes regular observations. The non-regular observations, described as outliers, represent systematic deviations. These outliers are often suspected as possible candidates for differential expression genes.

Here we use an exploratory approach consisting of two-stages to detect outliers from bivariate population and determining differentially expressed candidates from these outliers. The approach provides the fold-change value considering the scatter of observations and thereby provides up and down regulated genes across the samples.

3.2. Functional Classification

To determine biological significance of differentially expressed genes, functional classification was performed.

3.2.1. Gene Ontology

GO provides a dynamic controlled vocabulary and hierarchy that unifies descriptions of biological, cellular and molecular functions across genomes.

3.2.2. Pathway Analysis

To determine pathways associated with differentially expressed genes, pathway analysis was performed

3.2.1.1. Gene Ontology Analysis

Molecular Function: Genes involved in NADH dehydrogenase (ubiquinone) activity, glutamate dehydrogenase [NAD(P)+] activity, CDP-diacylglycerol-glycerol-3-phosphate-3-phosphtidyltransferase activity are upregulated in D&PH with respect to H. Gene involved in protein kinase B binding, enzyme inhibitor activity, acyl-CoA oxidase activity, phosphatidylinositol transporter activity, acyltransferase activity are downregulated in D&PH with respect to H.

Biological Process: Genes involved in synaptic vesicle membrane organization and biogenesis, polysaccharide metabolic process, regulation of growth rate, nucleosome assembly are upregulated in D&PH with respect to H. Genes involved in immune response, regulation of glycolysis are downregulated in D&PH with respect to H.

Cellular Component: Genes localized in cohesin core heterodimer, oligosaccharyl transferase complex, nucleosome, respiratory chain complex II are upregulated in D&PH with respect to H. Genes localized in isoamylase complex, protein kinase CK2 complex, proteasome activator complex, 6-phosphofructokinase complex are downregulated in D&PH with respect to H.

3.2.2.1. Pathway Analysis

Genes involved in Inositol phosphate metabolism, Starch and sucrose metabolism, Nitrogen metabolism, Oxidative phosphorylation, Androgen and estrogen metabolism, Glycan biosynthesis and metabolism pathways, Metabolism of cofactors and vitamins pathways, MAPK signalling pathway, ECM-receptor interaction, Neuroactive ligand-receptor interaction, Regulation of actin cytoskeleton, Cell communication pathways, Nervous system pathways, Neurodegenerative disorders pathways are upregulated in D&PH Vs H. Genes involved in Glycolysis / Gluconeogenesis, Propanoate metabolism, Carbon fixation, Biosynthesis of steroids, Fatty acid metabolism, Histidine metabolism, Phenylalanine metabolism, Tyrosine metabolism, Urea cycle and metabolism of amino groups, Cell cycle, Insulin signalling pathway, PPAR signaling pathway, Antigen processing and presentation are downregulated in D&PH Vs H.

4. SHARING AND COMPARING DATASETS

On the technical end, data sharing of transcriptome datasets is becoming relatively easy. Microarray data repositories (such as Gene Expression Omnibus [26] at the National Centre for Biotechnology Information, and Array Express [27,28] at the European Bioinformatics Institute) can accommodate even the largest datasets, and the deposited data are readily accessible by the whole scientific community. In an effort to standardize microarray data reporting, Brazma *et al.*[29] proposed a set of guidelines, Minimum Information About a Microarray Experiment (MIAME), to define parameters that uniformly describe each dataset, such as experimental design, sample preparation, hybridization procedures and use of controls[30]. Some journals, including *Nature Neuroscience*, now require public disclosure of the data in this format at the time of publication. Sharing of microarray data is also required from all researchers using the three NINDS/NIMH established microarray core facilities

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(http://arrayconsortium.cnmcresearch.org). Data generated by the consortium becomes publicly available 6 months after completion of the project.

Comparing microarray datasets is much more challenging [31]. First, we can compare outcomes of experiments. Was the expression pattern present in both experimental series? In this comparison, we rely on processed and analyzed data from different sources, accepting the data analysis that was done by the researchers who generated the datasets. This 'meta-analysis' is very useful, as replication of findings across different cohorts remains one of the critical aspects of post-mortem brain research. However, negative outcomes of such comparisons are difficult to interpret: methodological differences can substantially influence the results.

Second, we can compare RNA level changes based on the analysis of raw data generated by different laboratories. The availability of raw data permits researchers to systematically explore the changes in RNA levels using any preferred pre-processing or other data-analysis technique. Such *post-hoc* comparisons require that there be no major technical confounds between the datasets to compare, although this is almost never the case. Even if the same microarray platform and processing procedures are used, the operators, batches of reagents and microarray processing equipment differ. Furthermore, the samples are not processed in parallel. All this may introduce variability in the data and could confound the outcome of the *post-hoc* comparisons. Therefore, raw data comparison remains an important challenge.

5. MATERIALS AND METHODS

5.1. Stage- I: Multivariate Outlier Detection:

Outlier detection is one of the important tasks in any data analysis, which describe abnormalities in the data. Many methods have been proposed in the literature for detecting univariate outliers based on robust estimation of location and scale parameters. The standard method for multivariate outlier detection involves robust estimation of parameters in the *Mahalanobis Distance* (MD) measure and then comparing MD with the critical value of c2 distribution. The values larger than the critical value are treated as outliers of the distribution.

5.2. Mehalanobis Distance:

The covariance matrix is used for the quantification of the size and shape of the multivariate data, which is taken into account in the Mehalanobis distance. For a multivariate sample Xij, where *i* = 1,2,3,...n (number of genes) and *j* = 1,2,3...p (number of samples), the Mahalanobis distance is defined as,

$MD_i = ((X_{ij} - m)^T C^{-1} (X_{ij} - m))^{0.5}$

Where m is estimated multivariate location parameter and C is the estimated covariance matrix. The location and the covariance parameters are determined using Minimum Covariance Determinant estimation method. The MCD estimator is determined by that subset of observations of size *h*, which minimizes the determinant of the covariance matrix computed only from the *h* observations. The location estimator is the average of these *h* observations, whereas the scatter estimate is proportional to the variance covariance matrix.

5.3. Stage-II: Univariate Outlier detection:

Let S denote the original set of observations.

Let S_{out} and S_{in} be the subsets of S containing outlier and inlier observations respectively. Thus, $S_{out} \cup S_{in} = S$ and $S_{out} \cap S_{in} = \{\emptyset\}$, i.e. the two subsets are mutually exclusive.

We denote

$S_{out} = \{(\log_2(X_{i1}), \log_2(X_{i2})) / MD_i > c \text{ for } i=1,2,3...n\}$ and

$S_{in} = \{(\log_2(X_{i1}), \log_2(X_{i2})) / MD_i < c \text{ for } i=1,2,3...n\}$

where 'c' is the cut-off for a given quantile and n is the total number of genes.

We define a statistic, $Z = \log_2 (X_2 / X_1) = \log_2(X_2) - \log_2(X_1)$

Which is the log of the ratio of intensity values for different genes for the two samples.

Here X1 is treated as reference, while X2 is treated as test sample. The statistic provides a measure of differential expression (DE) of genes across the samples. The genes showing at least k-fold change (usually k=2, i.e. Z=1) across the samples are considered to be DE genes. The appropriate choice of k is important since it influences the number of DE genes. Here we propose a rationale for selecting *k* for a given percentage of bivariate outliers.

We generate values for the statistic for the entire set as,

$$Z = \{ z_i, i = 1,2,3,\ldots n \}$$
$$= \{ \log_2(X_{i2} / X_{i1}); \; i = 1,2,3,\ldots n\}$$

The statistic is used to obtain Mahalanobis distance measure as,

$$MD_i^* = \left[\frac{Z_i - m}{S_e}\right]^2 \text{ for } i = 1,2,3,\ldots n$$

The transformed distance measure is supposed to follow chi-square distribution with one degree of freedom. The empirical distribution function of MD* could be obtained and compared with that of the cumulative distribution of chi-square with one

PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
15th May 2011. Vol. 27 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

degree of freedom. A cut-off could be selected for MD* such that the observations greater than the cut-off could be declared as outliers. We search for an optimal cut-off, so that the univariate subset of outliers does not include any of the bivariate inliers. In other words, if Rout is a subset of univariate outliers and Sin the subset of bivariate inliers of S, then the optimal cut-off could be obtained as,

$$C^*_{opt} = \inf [\, C_i^* / R_{out} \cap S_{in} = \{\varnothing\}\,]$$

The optimal cut-off could be obtained programmatically thereby yielding a set of univariate outliers that overlap with a subset of multivariate outliers.

The cut-off value could be used in Mehalanobis distance measure to obtain the z-value as,

$$Z = (S_e)\sqrt{C^*_{opt}} + m$$

This z-value determines the log fold change resulting into bivariate outliers that could be the potential candidates for differential expression.

6. INDENTIFIYING THE GENES

In the present context, there are two individuals, one from each of the categories namely diabetes with parental history (D&PH) and healthy (H). The expression levels of 39400 genes for each individual were obtained and compared pair wise. Prior to analysis, the data for each combination was normalized using Lowess normalization.This analysis was carried out for each of these combinations independently based on above said procedure. Prior to analysis, the data for each combination was normalized using Loess normalization. Below we present the analysis for each combination along with the interpretations.

Diabetic with no parental history 1 vs Diabetic with parental history [D&NPH1 vs D&PH]

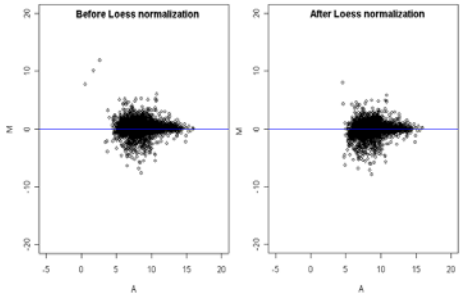

Figure 1: MA-plots showing scatter of expression values before and after loess normalization for diabetic with parental no history (1) vs. diabetic with parental history comparison.

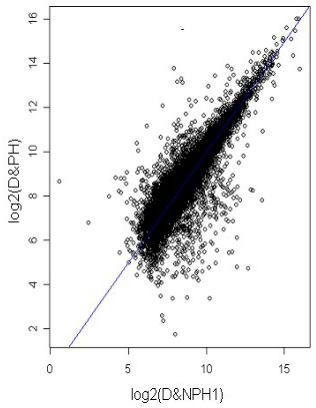

Figure 2: Scatter plot of log intensities for diabetic with parental no history (1) vs.diabetic with parental history comparison after loess normalization.

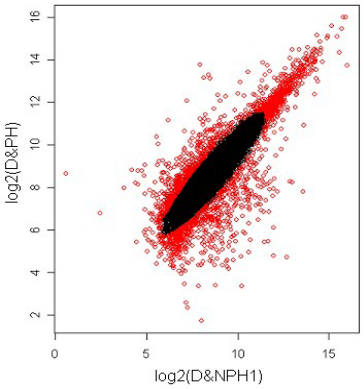

Figure 3: Bivariate outliers based on Mahalanobis distance measure for p=0.10 for diabetic with parental no history (1) vs. diabetic with parental history comparison.

The distribution of log fold change values was obtained and the outliers were detected for the optimum cut-off value (c*). Figure 4 shows the thresholds for 2-fold change, thereby providing the up and down regulated genes. Out of 3940 outlier genes, 686 were detected as up-regulated, while 682 were detected as down-regulated genes with respect to the individual with diabetic and no parental history (1). Thus, for diabetic with no parental history (1) vs. Diabetic with parental history comparison, 1368 were found to be differentially expressed out of 39400, which amounts to 3.4% of the total genes under study.

PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
15th May 2011. Vol. 27 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

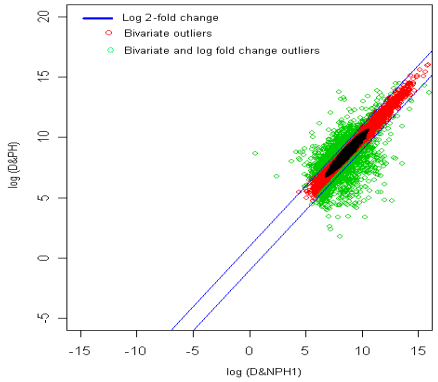

Figure 4: The thresholds for 2-fold change values.

The green spots are the differentially expressed outlier genes for diabetic with parental no history (1) vs. diabetic with parental history comparison. Here the modified threshold was same as conventional 2-fold change.

Diabetic with no parental history 2 vs Diabetic with parental history [D&NPH2 vs D&PH]

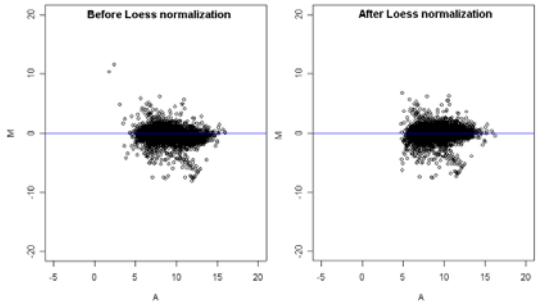

Figure 5: MA-plots showing scatter of expression values before and after loess normalization for diabetic with parental no history (2) vs. diabetic with parental history comparison.

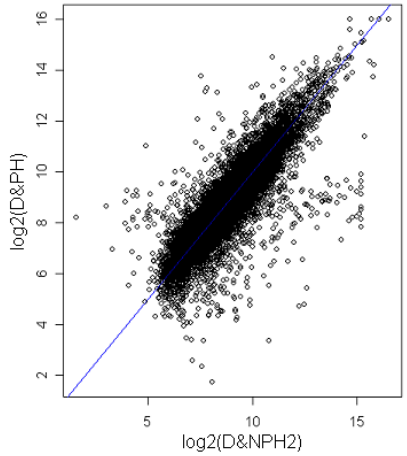

Figure 6: Scatter plot of log intensities for diabetic with parental no history (2) vs. diabetic with parental history comparison after loess normalization.

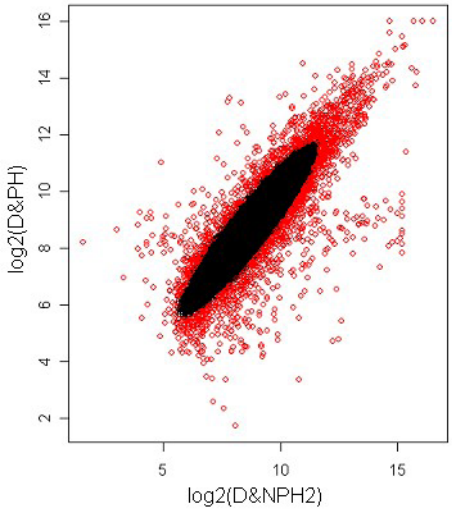

Figure 7: Bivariate outliers based on Mahalanobis distance measure for p=0.10 for diabetic with no parental history (2) vs. diabetic with parental history comparison. The distribution of log fold change values was obtained and the outliers were detected for the optimum cut-off value (c*). Figure 8 shows the thresholds for 2-fold change, thereby providing the up and down regulated genes. Out of 3940 outlier genes, 676 were detected as up-regulated, while 979 were detected as down-regulated genes with respect to the individual with diabetic and no parental history (2). Thus, for diabetic with no parental history (2) vs. Diabetic with parental history comparison, 1655 were found to be differentially expressed out of 39400, which amounts to 4.2% of the total genes under study.

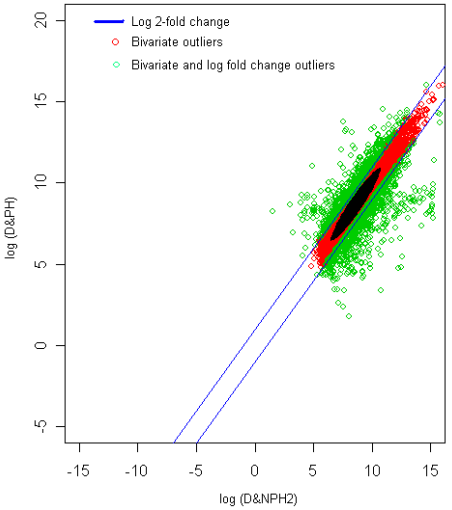

Figure 8: The thresholds for 2-fold change values. The green spots are the differentially expressed outlier genes for diabetic with parental no history (2) vs. diabetic with parental history comparison. Here the modified threshold was same as conventional 2-fold change.

PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
15th May 2011. Vol. 27 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

7. FUNCTIONAL CLASSIFICATION OF DIFFERENTIALLY EXPRESSED GENES

To determine biological significance of differentially expressed genes, functional **classification was performed using Gene Ontology**. Z-scores give statistical significance, indicating relative representation up-regulated/down-regulated genes in each function.

To determine pathways associated with differentially expressed genes, **pathway analysis** was performed. Pathway reports are provided in supplementary material. Numbers in parentheses indicate number of up-regulated / down regulated genes and total number of genes (in uploaded data), present in that particular pathway respectively.

7.1. GENE ONTOLOGY ANALYSIS

7.1.1. DIABETES WITH HISTORY VS DIABETES WITHOUT HISTORY *D&PH Vs D&NPH1.*

1. **Molecular Function**: Genes involved in MHC class II receptor activity, gamma-aminobutyric acid:hydrogen symporter activity, chemokine receptor activity, interleukin-4 receptor activity, interleukin-7 receptor activity, arachidonate 5-lipoxygenase activity, complement receptor activity are upregulated in D&PH Vs D&NPH1.

Genes involved in ammonia ligase activity, transaldolase activity, 4- alpha-glucanotransferase activity, choline:sodium symporter activity, interleukin-8 receptor activity are downregulated in D&PH Vs D&NPH1.

2. **Biological Process:** Genes involved in cell activation, macromolecule biosynthetic process, hydrogen peroxide biosynthetic process, immune response, regulation of glycolysis are upregulated in D&PH Vs D&NPH1.

Genes involved in blastocyst growth, aromatic compound biosynthetic process, nitric oxide biosynthetic process, regulation of glycolysis are downregulated in D&PH Vs D&NPH1.

3. **Cellular Component:** Genes localized in ribonucleosidediphosphate reductase complex, interleukin-18 receptor complex, interleukin-1 receptor complex, mitochondrion interleukin-5 receptor complex are upregulated in D&PH Vs D&NPH1.

Genes localized in proteasome activator complex, isoamylase complex, CAAX-protein geranylgeranyltransferase complex, protein kinase CK2 complex, oxoglutarate dehydrogenase complex, MHC class I peptide loading complex are downregulated in D&PH Vs D&NPH1.

7.1.2. *D&PH Vs D&NPH2*

1. **Molecular Function**: Genes involved in structural constituent of ribosome, MHC class II receptor activity, ferroxidase activity, NAD(P)H oxidase activity are upregulated in D&PH Vs D&NPH2.

Genes involved in 4-alpha-glucanotransferase activity, phosphomannomutase activity, receptor signaling protein tyrosine kinase activity are downregulated in D&PH Vs D&NPH2.

2. **Biological Process:** Genes involved in intracellular sequestering of iron ion, ribosome biogenesis and assembly, hydrogen peroxide biosynthetic process are upregulated in D&PH Vs D&NPH2.

Genes involved in hemostasis, developmental growth, lipid glycosylation, regulation of glycolysis are downregulated in D&PH Vs D&NPH2.

3. **Cellular Component:** Genes localized in ribosome, ferritin complex are upregulated in D&PH Vs D&NPH2.

Genes localized in CAAX-protein geranylgeranyltransferase complex, isoamylase complex, apolipoprotein B mRNA editing enzyme complex, lipopolysaccharide receptor complex, proteasome activator complex are downregulated in D&PH Vs D&NPH2.

7.2. PATHWAY ANALYSIS

7.2.1. DIABETES WITH HISTORY VS DIABETES WITHOUT HISTORY1 [D&PH Vs D&NPH1].

Genes involved in signal transduction, Regulation of actin cytoskeleton, Antigen processing and presentation, Complement and coagulation cascades, Axon guidance, Neurodegenerative disorders pathways are up regulated in D&PH Vs D&NPH1.

Genes involved in carbohydrate pathways are down regulated in D&PH Vs D&NPH1.

7.2.2. DIABETES WITH HISTORY VS DIABETES WITHOUT HISTORY2 [D&PH Vs D&NPH2].

Genes involved in Oxidative phosphorylation, Metabolism of cofactors and vitamins pathways, Immune system pathways, Nervous system pathways, metabolic disorders pathways are up regulated in D&PH Vs D&NPH2.

Genes involved in Lipid metabolism pathways, Amino acid metabolism pathways, Glycan biosynthesis and metabolism pathways, Ubiquitin mediated proteolysis, Signal transduction pathways, Signalling molecules and interaction pathways, Insulin signalling pathway, PPAR signalling pathway are down regulated n D&PH Vs D&NPH2.

PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
15th May 2011. Vol. 27 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

8. GENES INVOLVED IN INFLAMMATORY RESPONSE

Diabetes with family history vs Diabetes without family history 1 **(D&PH VS D&NPH1)**	*ALK, CCL13, CCR8, CDKN1A, EDN1, FGF1, IFIT1, IL12RB1, IL20, IL22, IL2RG, IL8RA, ITGB2, MMP20, SLK, TNFRSF12A, UBC, XCR1*
Diabetes with family history vs Diabetes without family history 2 **(D&PH VS D&NPH2)**	*ALK, BLR1,C5, CCL15, CCL16, CCR7,CCR8,CX CL11, CXCL12, FN1, FTH1, GBP1, HLA-A, IFIT1, IL12A, ITGB2, KIT, LTB MMP20, PPARD, RHOA, RPS27A, TAC1, TLR4, TNFAIP6, TNFRSF11A, TNFRSF12A*

9. CONCLUSION

Gene Expression Analysis is performed between samples of diabetes with Parental History and without parental history using micro array analysis. The microarray data is normalized using Lowess Normalization method. The analysis is repeated for two different sets of samples. Gene Ontology and pathway analysis are performed to find out the pathways associated with these differentially expressed genes. It is observed that 38 inflammatory genes were identified out of 39400 genes. A study on the different factors influencing the identified differentially expressed genes is under progress.

REFRENCES:

[1] John Ten Bosch, Chris Seidel, Sajeev Batra, Hugh Lam, Nico Tuason, Sepp Saljoughi, and Robert Saul, "Validation of Sequence-Optimized 70 Base Oligonucleotides for Use on DNA Microarrays", OPERON a QIAGEN COMPANY,2000.

[2] Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ., "Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays", Department of Molecular Biology and Genomics and Department of Infectious Diseases, Pfizer Global Research and Development, Ann Arbor, MI 48105, 2000, USA.

[3] Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science270, 1995, 467–470.

[4] Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. 14, 1996,pp.1675–1680..

[5] Cheung, V.G. et al. Making and reading microarrays. Nat. Genet. 21, 1999, pp.15–19.

[6] Hoffman, E. et al. Guidelines: Expression profiling - best practices for data generation and interpretation in clinical trials. Nat. Rev. Genet. 5, 2004,pp.229–237.

[7] Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 2003,pp.249–255.

[8] Blalock, E.M. et al. Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses.Proc. Natl. Acad. Sci. USA 101, 2004,pp.2173–2178.

[9] Quackenbush, J. Microarray data normalization and transformation. Nat. Genet. 32 (Suppl.), 2002, pp.496–501.

[10] Smyth, G. & Speed, T. Normalization of cDNA microarray data. Methods 31,2003, pp.265–273.

[11] Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95(25),1998,pp.14863–14868.

[12] Stuart, J., Segal, E., Koller, D., Kim, S.: A gene coexpression network for global discovery of conserved genetic modules. Science 302(5643), 2003,pp.249–255.

[13] Zhang, B., Horvath, S.: A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology 4,2005.

[14] Carter, S., Brechbhler, C., Griffin, M., Bond, A.T.: Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics, 20004,pp.2242–2250.

[15] Lee, H., Hsu, A., Sajdak, J., Qin, J., Pavlidis, P.: Coexpression analysis of human genes across many microarray data sets. Genome Research 14, 2004,pp.1085–1094.

[16] Datta, S., Datta, S.: Methods for evaluating clustering algorithms for gene expression data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

using a reference set of functional classes. BMC Bioinformatics , 2006.

[17] Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. PNAS, 2003.

[18] M. Kathleen Kerr, Mitchell Martin, and Gary A Churchill, " Analysis of Variance for Gene Expression Microarray Data", Journal of Computational Biology, Vol.7, No.6,2000,pp.819-837.

[19] Anat Reiner, et.al. "Identifying differentially expressed genes using false discovery rate controlling procedures", Bioinformatics-Oxford University press, vol.19, No.3,2003,pp.368-375.

[20] D.L. Wilson, et.al. "New Normalization methods for cDNA microarray data", Bioinformatics-Oxford University press, vol.19, No.11, 2003, pp.1325-1332.

[21] Hong-Ya Zaho, et.al,(2004) "Identification of Differentially Expressed Genes with Multivariate Outlier Analysis", Journal of Biopharmaceutical Statistics, vol. 14, Issue 3, pp.629-646.

[22] Dan Nettleton, Justin Recknor and James M. Reecy, " Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis", vol. 24, No. 2, 2008,pp.192-201.

[23] Huaizhen Qin, Tao Feng, et.al,"An efficient method to identify differentially expressed genes in microarray experiements, " Oxford University press, vol. 24 no. 14 , 2008,pp.723-729.

[24] Bogdan Done , Purvesh Khatri , Arina Done, Sorin Draghici,"Detriotpredicting Novel Human Gene Ontology Annotations Using Semantic Analysis" IEEE/ACM Transactions On Computational Biology And Bioinformatics ,2010.

[25] Purvesh Khatri,Sorin Draghici,Adi L. Tarca,Sonia S. Hassan,Roberto Romero,"A system biology approach for the steady-state analysis of gene signaling networks", CIARP'07 Proceedings of the Congress on pattern recognition 12th Iberoamerican conference on Progress in pattern recognition, image analysis and applications,2007.

[26] Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 2002,pp.207–210.

[27] Brooksbank, C. et al. The European Bioinformatics Institute's data resources. Nucleic Acids Res. 31,2003,pp.43–50.

[28] Brazma, A. et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 31,2003, pp.68–71.

[29] Brazma, A. et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. 29, 2001, pp. 365–371.

[30] Causton, H.C. & Game, L. MGED comes of age. Genome Biol. 4, 2003.

[31] Mirnics, K. Microarrays in brain research: the good, the bad and the ugly.Nat. Rev. Neurosci. 2, 2001,pp.444–447.