OPEN ACCESS

Ology Press

Review

Biostatistics and Epidemiology International Journal

Commonly used statistical methods for detecting differential gene expression in microarray experiments

SemaAtis Kuyuk,[1,2] Ilker Ercan[2]

[1]Department of Biostatistics and Medical Informatics, Halic University, Turkey

[2]Department of Biostatistics, Institute of Health Sciences, Uludag University, Turkey

Correspondence: SemaAtis Kuyuk, Department of Biostatistics and Medical Informatics, Halic University, Sutluce, Beyoglu, Istanbul 34440, Turkey, Email semaatis@halic.edu.tr

Received: October 31, 2017 | **Published:** December 13, 2017

Abstract

Bioinformatics tools provide the needs for carrying out microarray analysis. Differential gene expression analysis reveals the substantial aspects of biological pathways and takes a leading part for further hypothesis development. The principle in analyzing the gene expression data is the need for determination the genes of which expression models differ by phenotype or an experimental condition. Microarrays are employed to detect the distinctive gene expression profiles in a wide variety of experimental conditions. In the field of bioinformatics, different computer science technologies and statistical methods are considered. This multidisciplinary approach allows the understanding of the relationship between statistical methods, bioinformatics applications and computer science technologies. Images resulting from microarray experiments are not directly interpreted to reveal differences in sample condition. To make microarray experiments interpretable, a number of algorithms and statistical approaches need to be applied. The raw dataset has to go through preprocessing step prior to a series of subsequent analyses as significance analysis clustering and visualization of the relevant biological components and samples. Many microarray studies are designed to detect genes associated with different phenotypes. This review attempts to give an overview of statistical methods for identifying differentially expressed genes in microarray experiments. In this review, our aim is to summarize briefly basic methods as a guide for differential gene expression studies. All the summarized methods are given from the basic to more complicated in the review. .

Keywords: microarray analysis, gene expression, differential gene expression analysis

Introduction

Bioinformatics research studies are concerned with the analysis of large quantities of biological data with the help of computational techniques. In recent years, advances in molecular biology and information technology have allowed a major part of the genomes of various species to be investigated. Current bioinformatics studies are concerned with the structural and functional aspects of genes and proteins since the amount of data produced in the field of molecular biology is enormous. Most of these studies are related to the Human Genome Project. In summary, bioinformatics is an interdisciplinary area at the intersection of information technology, statistics and biology.[1,2]

The most basic tasks in the field of bioinformatics are the creation and development of biological databases. The majority of these databases consist of nucleic acid sequences and protein sequences derived from them. One of the most important applications in the field of bioinformatics is the analysis of sequence information. Bioinformatics investigates the genetic structures of all living organisms through the development of new information technologies to clarify complex biological questions.[1]

A primary goal of empirical genetic studies is the identification, quantification, and comparison of genetic differentiation among individuals, populations, species and studies.[3] Microarray technology has made it possible to measure the expression of thousands of genes simultaneously.[4] Statistical analysis of microarray data is started through software programs using CEL files defined as raw data. Prior to the start of the analysis, quality assessment of raw data is performed as the first step. In order to evaluate the homogeneity of the arrays and to compare the density distribution between the arrays, box graphs are plotted for each array using the densities of the logarithm2 base of the raw data. Images of the CEL files are obtained to observe the dimensional distributions of the densities on each array and to detect dimensional artifacts. MA-plots are used to compare the expression values for all possible pair of arrays with a probeset-wise median array. The MA plots are generated by plotting M values which are obtained by logarithmic ratios versus A values which are average logarithmic intensity values. The pre-normalization quality control step can be complemented by histograms drawn to assess the density distributions of each array.[5]

After quality control of raw data, background correction and normalization should be applied to the data using background correction methods such as RMA (Robust Multiple-Array Average) method. With the RMA method, the probe-level signal is removed from the background signal. Quantile normalization is performed by the RMA method and it is ensured that all the arrays have the same quantile. Using the RMA method, the expression set to be used in

Citation: Kuyuk S, Ercan I. Commonly used statistical methods for detecting differential gene expression in microarray experiments. *Biostatistics Epidemiol Int J*. (2017);0(0): 1-8. DOI: 10.30881/beij.00001

the analysis is generated by normalized and the background corrected intensities. After the background correction and the normalization methods are performed, box charts related to each array are drawn to re-evaluate the quality control. Following normalization and background correction, a list of genes that differ between two different conditions can be obtained by applying various statistical tests to the expression dataset to be used for analysis.[5]

Preprocessing of microarray data

Gene expression measurement is generally obtained as a measure of fluorescence intensity.[6] Background fluorescence can arise from many sources such as non-specific binding, residual precipitates after the washing step, optical noise from the scanner.[5,7]

Measurement values may have undergone various adjustments in the device system, such as calibration. Thus, in the presentation of gene expression data, it must be explained how the values are generated by the device system.[5,6] These expression measures always contain a component called "background noise." Local background noise levels are measured from the areas of the glass slide that do not contain probes. The background correction tries to remove non-specific background noise and local variations of the overall signal level on each chip.[5] The most common method to remove the background effect is to remove the measured fluorescence intensity around the spots.[8]

Microarray gene expression data sets consist of ω_{gn} gene expression values, with $\mathrm{g} = 1,\ldots,\mathrm{G}$ genes and $\mathrm{n} = 1,\ldots,\ \mathrm{N}$ samples. ω_{gn} values are arranged in a $\mathrm{G} \times \mathrm{N}$ data matrix, where each gene corresponds to one row and each sample to one column. The readout gene expression value ω_{gn} can be statistically defined as the sum of the true gene expression value x_{gn} and the background noise B_{gn} components;[6]

$$\omega_{\mathrm{gn}} = \mathrm{x}_{\mathrm{gn}} + \mathrm{B}_{\mathrm{gn}} \qquad (1)$$

The structure and correction of the background noise depend on the microarray technology used. Spot array data provides an estimate of background noise B_{gn}, with uncorrected expression intensities ω_{gn} values. If the background estimate is expressed as $\hat{\mathrm{B}}_{gn}$, background corrected expression value, $\omega_{\mathrm{gn}}^{(\mathrm{c})}$ is given as follows[6];

$$\omega_{\mathrm{gn}}^{(\mathrm{c})} = \omega_{\mathrm{gn}} - \hat{B}_{\mathrm{gn}} = \left(\mathrm{x}_{\mathrm{gn}} + \mathrm{B}_{\mathrm{gn}}\right) - \hat{B}_{\mathrm{gn}} \qquad (2)$$

The most common methods used for background correction in microarray analysis are; The "Robust Multi-Array Average (RMA) Background Correction" method and the "MAS 5.0 Background Extraction" methods.[9]

RMA background correction: RMA background correction is a method that uses only Perfect Match (PM) intensities. PM values are corrected using a global model for the distribution of probe intensities.[7]

The model is based on the experimental distribution of probe intensities. Observed PM probes are modeled as a Gaussian noise component with μ average and σ^2 variance.[7]

To avoid negative expression values, the normal distribution is truncated at zero. If the observed density is assumed to be *Y*, the correction will be as follows;

$$E\left(S \middle| Y=y\right) = a + b\,\frac{\varnothing\left(\frac{a}{b}\right) - \varnothing\left(\frac{y-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{y-a}{b}\right) - 1} \qquad (3)$$

$\alpha = s - \mu - \sigma^2\alpha$ and $b = \sigma$ where S is an averaged exponential signal component with α mean. ϕ and Φ are the standard normal density and distribution functions, respectively.[7]

MAS 5.0 background correction: In the MAS 5.0 background correction method, the chip is divided into a rectangular grid with *k* rectangular regions. At each region, at least 2% of the probe intensities are used to calculate a background value for this grid. Then, each probe intensity is corrected based on a weighted average of the background values. The weights depend on the distance between the probes and the center of gravity of the grid.[7]

Weights are calculated as follows;

$$\omega_k\left(x,y\right) = \frac{1}{d_k^2\left(x,y\right) + s_0} \qquad (4)$$

Where $d_k\left(x,y\right)$ is a Euclidean distance from $\left(x,y\right)$ position to the center of gravity of region *k* and s_0 is correction coefficient.

In MAS 5.0 Background Correction method, both Perfect Match (PM) and Mismatch (MM) probes are corrected.[7]

RMA background correction has been one of the most commonly used pre-processing method in the recent literature.[10-12] Performed assessments of the measure's precision, consistency of fold change, and specificity and sensitivity of the measure's ability to detect differential expression and demonstrated the substantial benefits of using the RMA measure to users of the Gene Chip technology. They used data from spike-in and dilution experiments to conduct various assessments on the MAS 5.0, dChip and RMA expression measures. Irizarry have demonstrated that RMA has similar accuracy but better precision than the other two summaries and RMA provides more consistent estimates of fold change.[12]

The study of[13] implements seven data extraction methods including MAS 5.0 and RMA to calculate expression indices from Affymetrix microarray gene expression data and tested use of different background correction methods calculated the correlation coefficient for each pair-wise comparison of background correction methods.[15]

Quality assessment: It is necessary to evaluate the quality of the data before the normalization of the arrays. Quality control assessment should be carried out to determine whether the quality of experimental data is acceptable and whether any hybridization should be repeated.[5,7]

Various descriptive data plots are drawn to identify potential problems with hybridization or other experimental structures in the quality control evaluation process. Quality control plots are basically divided into diagnostic and spot statistics.[6,7]

Diagnostic Plots: The diagnostic plots include various plots such as MA-plots for evaluating intensity bias and histograms for examining signal-to-noise ratios for each channel. Diagnostic plots are usually used to observe non-linear trends between two channels.[7]

a. MA plots: *M* and *A* are commonly used variables in the analysis of two-color arrays. *A* is defined as follows;

$$A = log_2 \sqrt{Cy5\ .Cy3} = \frac{1}{2}\left[log_2(Cy5)+log_2(Cy3)\right] \qquad (5)$$

Cy5 and Cy3 denote green and red dye intensities for a given spot, respectively. *A* variable is a measure of the total intensity of the logarithmic transformation of a spot. Thus, if the combined red and green intensities are high for a particular spot, the *A* value will also be high.[7,9]

M variable is defined as follows;

$$M = log_2 \frac{Cy5}{Cy3} = log_2 \left(Cy5\right) - log_2 \left(Cy3\right) \qquad (6)$$

The *M* variable is the logarithmic transformation of the intensity ratio. The *M* value shows which of the red and green dyes are more binding to a particular spot array.[7]

MA plots are used to investigate density bias. *A* disproportionate amount of spot above or below the x-axis on the graph indicates a problem in the array. MA plots are an indication of whether normalization within the array is required.[6,7]

Alvord et al.demonstrated the use of some of the exploratory plots including boxplots, volcano plots and MA plots for the expression level data on the soybean genome[14]. Lu et al.'s study, can be cited as an example of MA-plot application in method comparison studies, in which MA-plots were created on the raw data and normalized data to compare normalization methods.[15]

b. Histograms: In microarray designs, it is very important to obtain the histograms of the p-values of tests conducted to identify different gene expression. Histograms are graphs that are easy to interpret and contain considerable information. A histogram is an indication of whether there is a signal in the gene and whether the genes are differently expressed. Histograms also allow for estimation of how many genes are differentially expressed in reality.[16]

Spot statistics plots: Spot Statistics help to predict the structures of spot and hybridizations. The main plots that can be obtained with spot statistics are spatial plots, box plots, scatter diagrams and volcanic plots.[16]

a. **Spatial plots:** Spatial plots are used to reveal irregular spot and hybridization structures. Spatial plots are used to observe the spatial distributions of the intensities on each array and to detect the artifacts. Spatial plots play a fundamental role in determining the background correction, depending on whether there are any dimensional artifacts on the arrays.[7]

b. **Box plots:** Box plots are one of the most commonly used plots for displaying spot and hybridization structures. At the same time, box plots can be drawn to understand the scale differences between different arrays. It is necessary to evaluate the box plots to see if between-array normalization is required. The homogeneity of the arrays can be observed quite clearly from the box plots.[7,16]

c. **Scatter diagrams:** Scatter diagrams used to compare the expression values of two samples are the most commonly used plots for visualizing microarray data.[17] In the first step of the microarray data analysis, a scatter diagram is drawn between the two intensity channels to view the general structures and variability. Scatter diagrams are also commonly used to find out slides lying away from the center, which have abnormal expression structures.[18]

d. **Volcano plots:** Volcano Plots are used to summarize fold change and t-test criteria. A volcano plot is constructed by plotting the negative log of the p-value on the y-axis and log of the fold change between the two conditions on the x-axis. For each gene, there is a point on the graph that represents the t-statistic and the fold change[16] (Figure 1).

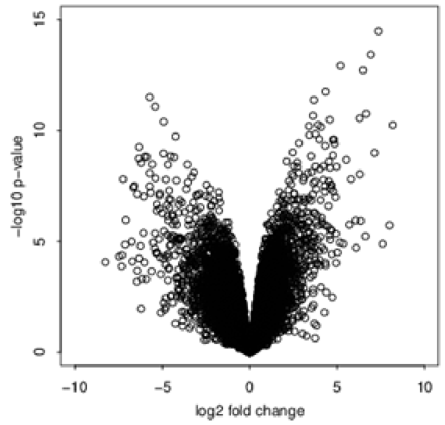

Figure 1 Volcano Plot.

Normalization: The purpose of the normalization phase is to adjust the data according to the technical variation. Variations can cause measurement differences between general fluorescence intensity levels of various arrays. The normalization process is necessary to make the measured values obtained from different arrays comparable.[9] Normalization methods depend on which microarray technology is used. Generally, logarithmically transformed data are used for further analysis.[19]

The most commonly used methods of normalization are as follows.[7]

1. Scaling Normalization Method
2. Nonlinear Normalization Methods
3. Quantile Normalization
4. Cyclic Loess Normalization
5. Contrast Normalization

Scaling normalization method: The Scaling Normalization Method scales all the arrays with the same average intensity by choosing a reference array. The constructed procedure is to determine a reference array and then create a linear regression between each array and the chosen array without a constant term. Subsequently, the regression line is used as a normalization relation.[7]

Nonlinear normalization methods: Non-Linear Normalization Methods perform non-linear corrections between arrays. These methods generally perform better than linear corrections such as the scaling method.[19]

Citation: Kuyuk S, Ercan I. Commonly used statistical methods for detecting differential gene expression in microarray experiments. *Biostatistics Epidemiol Int J.* (2017);0(0): 1-8. DOI: 10.30881/beij.00001

There are many nonlinear normalization methods in the literature. Workman et al. proposed a nonlinear normalization method using array signal distribution analysis and cubic splines.[18] Chen et al. proposed a subset normalization to adjust for location biases combined with global normalization for intensity biases.[19] Edwards[19] proposed a nonlinear LOWESS normalization method used in one channel cDNA microarrays mainly for correcting spatial heterogeneity.

Quantile normalization: The purpose of quantile normalization is to impose the same experimental density distribution on each array. If two data vectors have the same distribution, a Q-Q graph has a straight diagonal line with 1 slope and 0 intercept.[20]

Drawing quantiles of two data vectors against each other and designing each point on a 45-degree diagonal line leads to a transformation that allows both data vectors to have the same distribution.[20]

Cyclic loess normalization: The cyclic loess method is a generalization of the global loess method in which the Cy5 and Cy3 channel intensities are normalized using MA graphics. When dealing with single channel array data, array pairs are normalized according to each other. The Cyclic Loess method normalizes the intensities for an array set by working in dual form.[21]

Contrast normalization: In contrast normalization, the data is transformed into a contrast set and a nonlinear MA-plot normalization is performed. Afterward, inverse transformation is applied.[7]

Normalization procedures are essential as the first step of expression analyses for adjusting artifacts on samples and making samples comparable. Choice of normalization procedure plays a fundamental role in the final results of gene expression analysis. There are several methods for normalization in the literature. Quantile normalization procedure is one of the most commonly used between these methods.

Qian et al. used quantile algorithm for normalization in their study which is aimed to identify differentially expressed genes and compare the expression profiles.[22] The raw expression data was preprocessed using the RMA which includes default configuration for background correction, normalization and calculation of expression values in the most of microarray studies as Zhang et al.'s study and Kupfer et al.'s study.[11,14,23] Wu et al. demonstrated that cyclic loess normalization procedures performed better than quantile normalization procedures at reducing the number of false-positive up-regulated miRNAs.[24]

Statistical methods used in differential gene expression analysis: The principle in analysing the gene expression data is the need for determination the genes of which expression models differ by phenotype or an experimental condition.[1] A simple approach to selecting genes is to use the "fold change" criteria. This is only possible if there are no or only a few repetitions. However, an analysis based only on the fold change does not allow for the assessment of the significance of expression differences in the presence of biological and experimental variations that may vary from gene to gene. This is the main reason for using statistical tests to evaluate differential expressions. Parametric tests generally have a higher power if the underlying model assumptions are met.[9]

When doing the statistical analysis of microarray data, an important question is determining which scale to analyze the data. Generally, a logarithmic scale is used to approximate the distribution of the repeated measures for each gene to roughly symmetric and normal.[7] The variance-stabilizing transformation derived from an error model for microarray measurements can be used to make the variance of the measured intensities independent of their expected value. This may be advantageous for gene-based statistical tests based on variance homogeneity.[16] This will reduce variance differences between experimental conditions arising from differences in intensity level. However, it should not be forgotten that differences in the variance between conditions may also have gene-specific biological causes.[19,20]

t-test comparisons for one or two groups, variance analysis for multiple groups, and trend tests are frequently used models to assess differential gene expression. Due to lack of knowledge about the co-regulation of genes, linear models are usually calculated separately for each gene. When lists of relevant genes are identified, researchers can begin coordinated regulatory studies to further model their common actions.[7,20]

A statistical testing approach for each gene is common. However, this approach has some difficulties. Most importantly, a large number of hypothesis tests are being performed. This potentially leads to a number of false positives. Multiple test procedures allow the assessment of the overall significance of the results of a group of hypothesis tests. These procedures focus on specificity by controlling type I (false positive) error rates such as experimental error rate or false discovery rate. These controls are statistical methods used in multiple hypothesis tests to correct for multiple comparisons. However, testing multiple hypotheses remains a challenge. Because an increase in specificity is related to a loss of sensitivity, as provided by the p-value correction methods. Therefore, the possibility of detecting the true positivity decreases.[7,9]

Most microarray experiments involve only a few repetitions for each condition, making it difficult to predict gene-specific variances. Different methods have been developed to take advantage of variance information from all gene data.[18]

Fold change criteria: A simple microarray experiment is carried out to determine the expression differences between two conditions. Each condition can be represented by one or more RNA samples. The simplest method used to test expression differences of genes is the "Fold Change" criterion.[16]

The method calculates the logarithm rate between expression levels of the two conditions. It then evaluates whether all genes are greater than a threshold value determined for differential expression. The Fold change method is not considered reliable because it does not take statistical variability into consideration. The fold-change method is subject to bias if the data have not been properly normalized.[18]

For gene i, fold change is defined by as follows;

$$\mathrm{FC} = \frac{\overline{x}_i}{\overline{y}_i} \qquad (7)$$

Where x_i and y_i denote the expression levels in the control and treatment groups for the i^{th} gene, respectively.[16]

Student t-test: The Student's t-test is one of the most basic statistical methods used to determine differentially expressed genes if there is constant variance. Student t-distribution should be used if the genes are independent of each other and the observations are normally distributed.[20]

Let x_1 and x_2 be two independent gene expression data under two conditions; For example, normal and disease groups, two samples t-statistic for a given gene is;

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (8)$$

Where $\bar{x}_1$ and $\bar{x}_2$ denote the average expression level of a given gene in the control and disease group, respectively. *S* is the pooled standard deviation. Pooled variance is s^2 is estimated by:

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \quad (9)$$

Where s_1^2 and s_2^2 are the variances for control and disease groups, respectively.

The calculated test statistic is compared to the critical value of $n_1 + n_2 - 2$ degrees of freedom where n_1 and n_2 are the sample sizes for the control and disease group, respectively. Since the t-test makes use of the variances of the samples, it also draws attention to the shortcomings of the fold change approach.[20]

The standard t-test is frequently used to identify differentially expressed genes between two conditions in microarray studies. Shi et al. used student's t-test to obtain differentially expressed genes between embryonic stem cells and urinary induced pluripotent stem cells.[25]

Satterthwaite-welch t-test: Homogeneity of variances is rarely seen in microarray experiments. Heterogeneous sample or cell samples may cause heterogeneous variance in microarray experiments. Changing the correlation structure of expression change by the transcription factor can also cause heterogeneous variance. Therefore, Welch (Satterthwaite's) t-test method would be more appropriate for independent samples with different variances.[21]

Let x_1 and x_2 be two independent gene expression data under two conditions, The Welch t-statistic is calculated as follows for a given gene;

$$t_w = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (10)$$

Where $\bar{x}_1$ is the mean value of given gene in the control group with n_1 sample size, $\bar{x}_2$ is the mean value in the disease group with n_2 sample size, and s_1^2 and s_2^2 are the variances for the control and disease groups, respectively.[21]

The Welch method has a special correction for the degree of freedom under the variances of different samples;

$$\mathcal{V} = \frac{(\omega_1 + \omega_2)^2}{\omega_1^2/(n_1-1) + \omega_2^2/(n_2-1)} \quad (11)$$

ω_i is the estimated squared standard error for sample i;

$$\omega_i = \frac{s_i^2}{n_i} \quad (12)$$

Barajas et al. used Welch test to compare imaging and histopathologic variables in their study.[26] Mosig et al. compared monocyte gene expression profiles from FH patients with healthy controls using a Welch t-test.[27]

Paired sample t-test: The dependent design is often used in two-channel experiments where gene expression comparisons involving a natural match of experimental units are made. Student t-test and Welch t-test are used for independent samples. The dependent sample t-test should be used when the samples are dependent.[21]

Let $\bar{x}_j$ and s_j be the mean difference and standard deviation of differences, respectively. Paired sample test statistic is given by;

$$t = \frac{\bar{x}_d}{s_d/\sqrt{n}} \quad (13)$$

With *n-1* degrees of freedom.

Moderated t-test: Moderated t-test is one of the most common methods used in microarray studies. The modification is to add a small value to the standard deviation, which reduces the variability of the t-value, while the t-statistic is being calculated. The main purpose of the moderated t-test is to reduce the statistical significance of genes with a small standard deviation to avoid false positives.[19,28]

Moderated t-statistic is described by as follows;

$$T_i^{'} = \frac{\bar{x}_i - \bar{y}_i}{s_i + s_0} \quad (14)$$

Where s_0 is a selected constant to reduce the variability of $T_i^{'}$.

The moderated t-statistic improves the ranking performance if the purpose is to create a short sorted list.[21,29] It has been proven in many studies that the moderated t-statistic performs better than the classical t-test and the fold change criteria.[30,31]

Moderated t-test is most frequently used method to identify differentially expressed genes between two conditions in the literature. Shi et al. obtained differentially expressed genes in the five datasets of coronary heart disease using moderated-t test.[32] Dolah et al. used moderated t-test for genes in dolphin skin differentially expressed according to sex.[33]

Wilcoxon rank sum test: The Wilcoxon rank sum test is a nonparametric method used to compare gene expressions for two groups when x_i expression values are not normally distributed. In order to compare the samples, the ranks of the observations are used instead of the original observations.[20]

Wilcoxon rank sum statistic is given by;

$$z_g = \sum_k Rank\left(x_{g1k}\right) \quad (15)$$

Where $Rank\left(x_{g1k}\right)$ is the ranking between all the x_{g1k} and x_{g2k} values of x_{g1k}. Sorting starts from 1 for the smallest value and the highest value is $n_{g1} + n_{g2}$. If all the x_{g1k} values are smaller than

Citation: Kuyuk S, Ercan I. Commonly used statistical methods for detecting differential gene expression in microarray experiments. *Biostatistics Epidemiol Int J.* (2017);0(0): 1-8. DOI: 10.30881/beij.00001

x_{g2k} values, then z_g will be $\sum_{k=1}^{n_{g1}} k$. If all the x_{g1k} values are higher than x_{g2k} values, then z_g will be $\sum_{k=n_{g2}+1}^{n_{g1}+n_{g2}} k$. All other possibilities are between these two values [20].

ANOVA F-test: ANOVA is used to investigate the significance of the effects of factors that may affect gene expression in microarray data analysis. The use of the ANOVA model for two-color microarray designs was first proposed by Kerr et al.[7] and has been an important part of the literature. Jiang et al. tested barley genes for differential expression by one-way analysis of variance and controlled FDR according to the standard method of Benjamini and Hochberg.[13]

The ANOVA model for microarray data can be determined in two steps.[34] The first stage is the normalization model;

$$y_{ijgr} = \mu + A_i + D_j + AD_{ij} + r_{ijgr} \qquad (16)$$

Where y_{ijgr} is the logarithm of signal intensity for the *i*th array, *j*th dye, *g*th gene and *r*th measurement. µ is the overall average expression level. *A* is the effect of the array at the measured intensity, *D* is the effect of the dye, and *AD* is the effect of interaction between dye and array. The first step is to form r_{ijgr} term from the measured intensities. In the second step, gene-specific effects are modeled in terms of residuals of the normalization method.[21] The gene-specific model is expressed as follows;

$$r_{ijgr} = G + VG_{ij} + DG_j + AG_i + \in_{ijr} \qquad (17)$$

G is the average intensity for a particular gene, AG_i is the effect of an array on this gene, DG_j is the effect of dye on this gene, and $\in_{ijr}$ is the residual value. The variety-by-gene (VG) term is the main interest in the analysis, and reflects the variability in expression levels between samples for a particular gene. VG_{ij} is expressed as a "catch-all" term for the effects related to samples. In the simplest case, VG_{ij} is an indicator of the condition represented by the sample for *j* dye and *i* array. In more complex experiments, the design structure at the biological sample level is reflected in the *VG* terms.[34]

When there are duplicated spots in an array, the model should include additional terms for labeling and spot effects. The Gene-specific model can be modified by removing the terms *DG* and *AG* for single-color data.[34]

Hypothesis testing involves comparing two models. The null hypothesis suggests that there are no differential expressions (the VG values are equal to zero) and the alternative hypothesis suggests that there are differential expressions between conditions (the VG values are not equal to zero).[29] F statistics are calculated with residual sum of squares;

$$F1 = \frac{(AKT_0 - AKT_1)/(\nu_0 - \nu_1)}{AKT_1/\nu_1} \qquad (18)$$

$$F2 = \frac{(AKT_0 - AKT_1)/(\nu_0 - \nu_1)}{\sigma^2_{pool}} \qquad (19)$$

$$F1 = \frac{(AKT_0 - AKT_1)/(\nu_0 - \nu_1)}{(AKT_1/\nu_1 + \sigma^2_{pool})/2} \qquad (20)$$

AKT_0 and AKT_1 are the residual sum of squares with ν_0 and ν_1 degrees of freedom, for zero and alternative models, respectively. σ^2_{pool} is the pooled variance between all the genes.[34]

Moderated F-test: Moderated t statistics lead to F statistics that can be used for test hypotheses about any set of comparisons. The appropriate quadratic forms of the moderated t statistics follow F distributions.[35]

Where the average of the contrast estimators is β_g , suppose that all comparisons for a given gene will be tested. While U_g is the diagonal matrix, C is the contrast matrix, and V_g is the positive definite matrix, the correlation matrix of $\hat{\beta}_g$ will be $\mathrm{R}_g = \mathrm{U}_g^{-1}\mathrm{C}^T\mathrm{V}_g\mathrm{C}\mathrm{U}_g^{-1}$. Let r be the column order of C, $\mathrm{Q}_g^T\mathrm{R}_g\mathrm{Q}_g = \mathrm{I}_r$ and $\mathrm{q}_g = \mathrm{Q}_g^T\mathrm{t}_g.F_g$ is obtained by as follows;

$$F_g = q_g^T q_g / r = t_g^T Q_g Q_g^T t_g / r \sim F_{r,d_0+d_g} \qquad (21)$$

If the columns of Q_g are chosen to be eigenvectors covering the spacing of R_g , the diagonal matrix will be $\mathrm{q}_g = \mathrm{Q}_g^T\mathrm{t}_g$. F_g is obtained by as follows;

$$F_g = q_g^T \Lambda_g^{-1} q_g / r \sim F_{r,d_0+d_g} \qquad (22)$$

Moderated F-statistic tests whether any of the comparisons for a given gene is zero. The result of the test is whether or not the gene is differently expressed in any comparison. In complex experiments involving too many comparisons, it is primarily preferable to select genes based on moderated F-statistics.[35]

False discovery rate: The great majority of discussions on error rate in the literature relate to experimental and per comparison error rates. In recent years, the false discovery rate (FDR) proposed by Benjaminini and Hochberg[35] has become very widespread. FDR is defined as the expected value of the ratio of false rejections to total rejections by the authors. Benjamini and Hochberg point out that controlling the FDR is more reasonable than controlling the experimental or per comparison error rates.[36]

Let $\mu_1, \mu_2, \ldots, \mu_J$ be the means to be compared, we are interested in testing $m = \frac{J(J-1)}{2}$ pair of hypotheses. U_n is the number of correctly rejected hypotheses from $m - R_n$ rejection sets. $h_0 - V_n$ is the number of pairs incorrectly rejected.[36]

Benjamini and Hochberg point out that the false rejected null hypothesis can be expressed by the Q random variable.[31]

Citation: Kuyuk S, Ercan I. Commonly used statistical methods for detecting differential gene expression in microarray experiments. *Biostatistics Epidemiol Int J.* (2017);0(0): 1-8. DOI: 10.30881/beij.00001

$$Q = {h_0 - V_n}\Big/{\left(h_0 - V_n\right) + V_n} \qquad (23)$$

When $m - R_n = 0$, Q is defined as zero. If there is no rejection, the error rate is zero. Benjamini and Hochberg defined the FDR as the average of Q.[36]

Consequently;

$$Error\ rate\ per\ comparison = E\left(\frac{h_0 - V_n}{m}\right) \qquad (24)$$

$$P\left(h_0 - V_n \geq 1\right) = Experimental\ Error\ Rate \qquad (25)$$

$$E\left(Q\right) = E\left(\frac{h_0 - V_n}{\left(h_0 - V_n\right) + U_n}\right) = E\left(\frac{h_0 - V_n}{m - R_n}\right) \qquad (26)$$

Conclusion

In this review, we have attempted to give a broad overview of the statistical methods used for differential gene expression analysis. The microarray has made possible the simultaneous investigation of thousands of genes. Microarray technology helps researchers learn about various kinds of diseases. The identification of differentially expressed genes has a great importance to understand biological issues. Scientists can find out the expression levels of thousands of genes by using microarrays. Many of the genetic discoveries show that fundamental need is to implement an appropriate statistical method for finding the differentially expressed gene lists. For this reason, both information technologies and statistical methods have been adapted and developed according to this need. The use of bioinformatics tools for the studies related to gene expression datasets has seen a massive increase in recent years. It is without a doubt, progress in the field of bioinformatics in the future will be at the forefront of other branches of science to define genetic structures. Bioinformatics has empowered the scientific researchers to understand the significant points of genetic issues and various kinds of diseases.

References

1. Chikhale NJ, Gomase VS (2007) Bioinformatics Theory and Practice. (1st edn), Himalaya Publishing House, Mumbai.
2. Arhipova I. The role of statistical methods in computer science and bioinformatics. *ICOTS-7*. 2011;2:276−278.
3. Dogan I, Dogan N. Statistical measures for genetic differentiation: Review. *Turkiye Klinikleri J Biostat*. 2016;8(2):180−186.
4. Kocak M. Identifying cyclic genes in time-course gene expression studies using proc traj. *Turkiye Klinikleri J Biostat*. 2015;7: 47−54.
5. Stekel D. Microarray Bioinformatics. (1stedn), United States of America by Cambridge University Press, New York, USA, 2003.
6. Lee T. Analysis of Microarray Gene Expression Data. (1stedn), Kluwer Academic Publishers, New York, USA, 2004.
7. Gentleman R, Carey VJ, Huber W, et al. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. (1stedn), Springer, China, 2005.
8. Berrar DP, Dubitzky W, Granzow M. Practical Approach to Microarray Data Analysis. (1stedn), Springer, New York, USA, 2002.
9. Sımon M. Methods of microarray data analysis II. (1stedn), Kluwer Academic Publishers, Dordrecht, Netherlands, 2002.
10. Kulyté A, Ehrlund A, Arner P, et al. Global transcriptome profiling identifies KLF15 and SLC25A10 as modifiers of adipocytes insulin sensitivity in obese women. *PLoS ONE*. 2017;12(6):e0178485.
11. Zhang H, Zhang X, Huang J, et al. Identification of key genes and pathways for peri- implantitis through the analysis of gene expression data. *Exp Ther Med*. 2017;13(5):1832−1840.
12. Irizarry RA, Bolstad BM, Collin F, et al. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*. 2003;31(4):e15.
13. Jiang N, Leach JL, Hu X, et al. Methods for evaluating gene expression from Affymetrix microarray datasets. *BMC Bioinformatics*. 2008;9:284.
14. Alvord WG, Roayaei AJ, Quinones AO, et al. A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R. *Brief Bioinform*. 2007;8(6):415−431.
15. Lu T, Costello MC, Croucher P, et al. Can Zipf's law be adapted to normalize microarrays? *BMC Bioinformatic*. 2005;6:37.
16. Gohlmann H, Talloen W. Gene Expression Studies using Affymetrix Microarray. (1stedn), Taylor & Francis Group LLC, New York, USA. 2010.
17. Loewe RP, Nelson PJ. Microarray bioinformatics. *Methods Mol Biol*. 2010;671:295−320.
18. Park T, Yi S, Lee S, et al. Diagnostic plots for detecting outlying slides in a cDNA microarray experiment. *Biotechniques*. 2005;38(3):463−471.
19. Tuımala J. DNA Microarray Data Analysis Using Bioconductor. (1stedn), CSC - IT center for science, Espoo, Finland. 2008.
20. Sreekumar J, Jose KK. Statistical tests for identification of differentially expressed genes in cDNA microarray experiments. *IJBT*. 2008;7:423−436.
21. Wit E, Mcclure RJ. Statistics for Microarray: Design, Analysis, and Inference. (1stedn) John Wiley & Sons Ltd, Chichester, USA. 2004.
22. Qian Y, Sun H, Xiao H, et al. Microarray analysis of differentially expressed genes and their functions in omental visceral adipose tissues of pregnant women with vs. without gestational diabetes mellitus. *Biomed Rep*. 2017;6(5):508−512.
23. Kupfer P, Guthke R, Pohlers D, et al. Batch correction of microarray data ksubstantially improves the identification of genes differentially expressed in Rheumatoid Arthritis and Osteoarthritis. *BMC Med Genomics*. 2012;5:23.
24. Wu D, Hu Y, Tong S, et al. The use of miRNA microarrays for the analysis of cancer samples with global miRNA decrease. *RNA*. 2013;19(7):876−888.
25. Shi L, Cui Y, Zhou X, et al. Comparative transcriptomic analysis identifies reprogramming and differentiation genes differentially expressed in UiPSCs and ESCs. *Biosci Trends*. 2017;11(3):355−359.
26. Barajas RF, Hodgson JG, Chang JS, et al. Glioblastoma Multiforme Regional Genetic and Cellular Expression Patterns: Influence on Anatomic and Physiologic MR Imaging. *Radiology*. 2010;254(2):564−576.
27. Mosig S, Rennert K, Buttner P, et al. Monocytes of patients with familial hypercholesterolemia show alterations in cholesterol metabolism. *BMC Med Genomics*. 2008;1:60.
28. Park T, Yı SG, Kang SH. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*. 2003;4:33−46.
29. Jae K. Statistical Bioinformatics: for Biomedical and Life Science Researchers. (1stedn), John Wiley & Sons, New Jersey, USA. 2010.
30. Rıs MLH, Zhao X, Kahev F, et al. Gene expression profile analysis of T1 and T2 breast cancer reveals different activation pathways. *ISRN Oncology*. 2013;1−12.

Citation: Kuyuk S, Ercan I. Commonly used statistical methods for detecting differential gene expression in microarray experiments. *Biostatistics Epidemiol Int J*. (2017);0(0): 1-8. DOI: 10.30881/beij.00001

31. Yu L, Gulati P, Fernandez S, et al. Fully moderated t-statistic for small sample size gene expression arrays. *Stat Appl Genet Mol Biol*. 2011;10(1):42−64.

32. Shi Y, Yang S, Luo M, et al. Systematic analysis of coronary artery disease datasets revealed the potential biomarker and treatment target. *Oncotarget*. 2017;8(33):54583−54591.

33. Van Dolah FM, Neely MG, McGeorge LE, et al. Seasonal Variation in the Skin Transcriptome of Common Bottlenose Dolphins from the Northern Gulf of Mexico. *PLoS ONE*. 2015;10(6):e0130934.

34. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*. 2003;4(4): 210−220.

35. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:1−25.

36. Ozkaya G, Ercan I. Examining multiple comparison procedures according to error rate, power type, and false discovery rate. *JMASM*. 2012;11(2):348−360.

Citation: Kuyuk S, Ercan I. Commonly used statistical methods for detecting differential gene expression in microarray experiments. *Biostatistics Epidemiol Int J*. (2017);0(0): 1-8. DOI: 10.30881/beij.00001