REVIEW ARTICLE

Microarray Data Analysis for Differential Expression: a Tutorial

ERICK SUÁREZ, Act., Ph D*; ANA BURGUETE, MD, Ph D†; GEOFFREY J. MCLACHLAN, Ph D, DSc‡

DNA microarray is a technology that simultaneously evaluates quantitative measurements for the expression of thousands of genes. DNA microarrays have been used to assess gene expression between groups of cells of different organs or different populations. In order to understand the role and function of the genes, one needs the complete information about their mRNA transcripts and proteins. Unfortunately, exploring the protein functions is very difficult, due to their unique 3-dimentional complicated structure. To overcome this difficulty, one may concentrate on the mRNA molecules produced by the gene expression. In this paper, we describe some of the methods for preprocessing data for gene expression and for pairwise comparison from genomic experiments. Previous studies to assess the efficiency of different methods for pairwise comparisons have found little agreement in the lists of significant genes. Finally, we describe the procedures to control false discovery rates, sample size approach for these experiments, and available software for microarray data analysis. This paper is written for those professionals who are new in microarray data analysis for differential expression and want to have an overview of the specific steps or the different approaches for this sort of analysis.

Key words: Preprocessing data for microarrays, Pairwise comparison for microarrays, False discovery rate, Free microarray analysis software

Bioinformatics is a growing branch of biology and is highly interdisciplinary, using techniques and concepts from different fields, such as informatics, biostatistics, statistics, epidemiology, mathematics, chemistry, biochemistry, and physics. Bioinformatics grew as a field of study to maintain, organize, analyze, and make accessible large amounts of gene and genomic sequence information. Bioinformatics has a large impact on biological research. Giant research projects such as the Human Genome Project (HGP) would be meaningless without the bioinformatics component. The HGP was completed in 2003, after 13-years, and was coordinated by the U.S. Department of Energy and the National Institutes of Health. One of the main purposes of HGP are: i) to identify the approximately 25,000 genes in DNA and ii) to determine the sequences of the 3 billion chemical base pairs that make up human DNA. More details of the HGP can be seen in the following website: genomics.energy.gov.

Genomics encompasses the study of all features of genomes and individual genes at the DNA level, including mutations, polymorphisms, and phylogenetic relationships that are based on sequence differences. Another aspect of genomics is concerned with the pattern of transcription (gene expression) as a function of clinical conditions in response to natural or toxic agents or at different times during biological processes, such as the cell cycle. One of the aims of gene expression studies is to discover the genes that are up – and down- regulated under specific conditions. Because of the large amount of data that is generated from these experiments, special computational tools are required for obtaining, storing, and analysing data (1).

DNA microarray is a high throughput technology to simultaneously evaluate quantitative measurements for the expression of thousands of genes. Previously, expression analysis was performed in a low-throughput fashion, one gene at a time, typically by northern blot analysis. DNA microarrays have been used to discover new genes by assessing gene expression between groups of cells of different organs or different populations and have been used to identify disease biomarkers that may be important in genetic epidemiology (2). The applications of microarrays for the study of neurological diseases, like multiple sclerosis, Alzheimer's disease, or neuromuscular diseases are promising, both for generating new pathophysiological hypotheses and for enabling new molecular classifications (3). Microarray data analysis on cancer research has opened new avenues for diagnosis and therapeutic interventions (4). Our capabilities for diagnosis and understanding of infectious diseases have also been enhanced by using microarrays (5-6).

Several microarray platforms have been used to assess gene expression, such as: Agilent, CodelinkTM Bioarray,

* Department of Biostatistics and Epidemiology, School of Public Health, University of Puerto Rico-Medical Sciences Campus, San Juan, Puerto Rico; †Instituto Nacional de Salud Pública en México; ‡Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane Australia

Address correspondence to: Erick Suárez, PhD, Department of Biostatistics and Epidemiology, School of Public Health, University of Puerto Rico-Medical Sciences Campus, P.O. Box 365067, San Juan, Puerto Rico, 00936-5067. Email: erick.suarez@upr.edu

cDNA, Expression Array System, Febit, GeneChip, NimbleGen, and Xeotron (7). Microarray platform performance is measured by several indicators, such as specificity and sensitivity. Specificity is the ability to distinguish sequences up to a certain homology. Imperfect specificity is caused by cross-hybridization of other transcripts. Sensitivity is defined as the lowest target concentration at which an acceptable accuracy is obtained. Cross-platform integration of data has had limited success to date. Differences arise from the intrinsic properties of the arrays themselves, and the various processing and analytical steps involved (7).

To prepare microarrays using cDNA technology, glass or nylon micro plates are used onto which thousands of single stranded pieces of DNA of lengths of tens of nucleotides are placed. Each spot on the plate corresponds to a particular gene. In standard terminology, the cDNAs spotted onto the arrays are called probes, and those in the RNA sample are called target genes. In a single reaction, two different RNA samples can be labelled with different colours and simultaneously incubated with a microarray. Robots (arrayers) are required to place (or array) a large number of probes onto slides. After DNA probes are arrayed onto slides, they are air–dried. The probes are immobilized by UV irradiation to form covalent bonds between the thymidine residues in the DNA and the positively charged amine groups on the silane slides. After cross linking, excess DNA molecules are removed by washing the arrays at room temperature and the arrayed samples are denatured in water before hybridization, that is when two complementary sequences find each other, such as the immobilized target DNA and the mobile cDNA, and lock together (Figure 1). After hybridization, a laser scanner measures dye fluorescence of each colour at a fine grid of pixel. Higher fluorescence indicates higher amounts of hybridized cDNA, which, in turn, indicate higher gene expression in the sample.

The oligonucleotide array is made up of sets of oligonucleotide probes, usually 25 nucleotides in length, representing thousands of genes, that are synthesized directly (in situ) on a quartz of wafer by photolithography. For each gene, there are 11 to 20 pairs of oligonucleotide probes near the 3'. Therefore, each probe pair belongs to a probe set of one mRNA molecule produced by one gene, as follows:

Gene → DNA → mRNA → Probe set (11-20 pairs) → probe pair (25 bases)

A pair of probes consists of a sense and an antisense sequence. Multiple probes are used for each gene to distinguish between specific and non-specific hybridizations. The most widely used oligonucleotide array is the Affymetrix GeneChip -or Affy- (Figure 2).

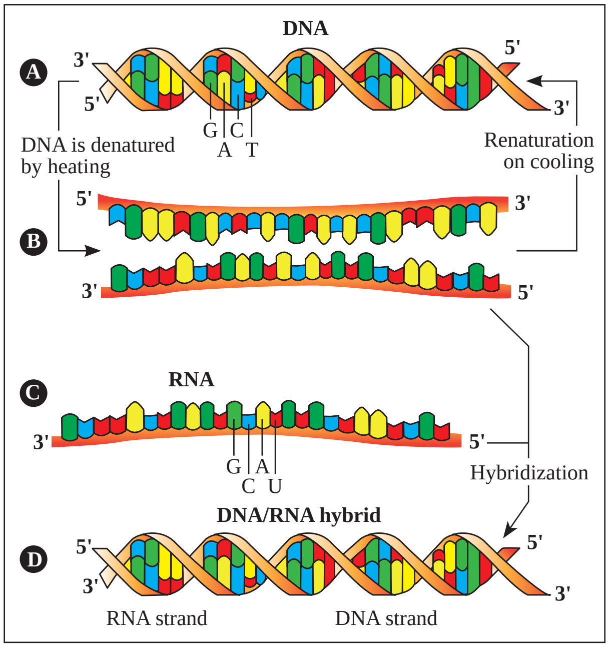


Figure 1. Nucleic Acid Hybrydization

The first type of probe in each pair is known as perfect match (PM) and is taken from gene sequence (25 bases). The second type probe is known as mismatch (MM) and is created by changing the middle (13th) base of the PM sequence to reduce the rate of specific binding of mRNA for that gene, as follows:

GGGAATGGGTCAGAA C GACTCCTATGTGGGTGGCT
Reference sequence

TTACCCAGTCTT C CTGAGGATACACCCAC Perfect Match
Oligo (PM)

TTACCCAGTCTT G CTGAGGATACACCCAC Mismatched
Oligo (MM)

The goal of MMs is controlling for experimental variation and non-specific binding of mRNA from other parts of the genome (8). These two probes (PM, MM) are referred to as a probe pair. RNA samples are prepared, labelled, and hybridized with array. Arrays are scanned and images are produced and analyzed to obtain an intensity value for each probe. These intensities represent how much hybridization occurred for each olinucleotide. The average of the PM-MM differences for all probe pairs is used as the expression index for the target gene (the DNA or RNA sequence of research interest) (9).

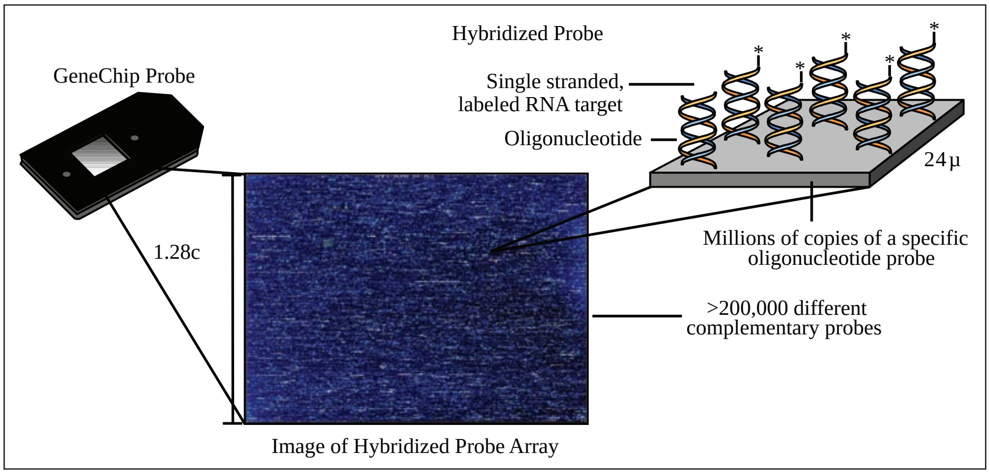

Hybridized Probe
GeneChip Probe
Single stranded, labeled RNA target
Oligonucleotide
24µ
1.28c
Millions of copies of a specific oligonucleotide probe
>200,000 different complementary probes
Image of Hybridized Probe Array

Figure 2. Oligonucleotide Chips (GENECHIP PROBE ARRAYS)

In order to understand the role and function of the genes, one needs the complete information about their mRNA transcripts and proteins. Unfortunately, exploring the protein functions is very difficult due to their unique 3-dimentional complicated structure and a shortage of efficient technologies. To overcome this difficulty, one may concentrate on the mRNA molecules produced by the genes of interest (gene expression) and use this information to investigate specific questions of the functional roles of the genes. Microarray experiments are often complex, generate large amounts of data, and warrant careful planning. Different books have already been published for Microarrays data analysis (8, 10-15). In this paper, we pretend to summarize some of the statistical approaches used for microarray data analysis, particularly for preprocessing data and for pairwise comparison from genomic experiments. We describe some of the procedures for image analysis, data normalization, and data summarization. In addition, we describe some of the alternatives to perform a pairwise comparison, to control the false discovery rate, to determine the sample size, and the software availability for microarray data analysis.

I) Preprocessing data

Preprocessing data in microarrays refers to the methods for controlling the effect of the different sources of variation during the experimental procedures before one obtains the genomic-level measurements (12). Microarrays are imaged using an optical scanner that must be subjected to background correction to adjust for nonspecific binding and fluorescence from other chemicals on the slide (16).

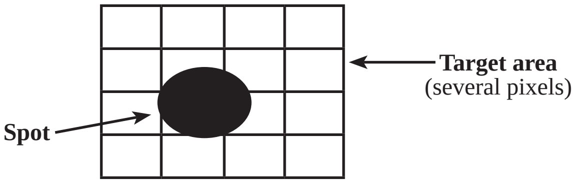
1.1) Image Analysis

One of the major objectives of microarray image analysis is to find the discrete spot locations and to quantify the spot intensities of gene expressions.

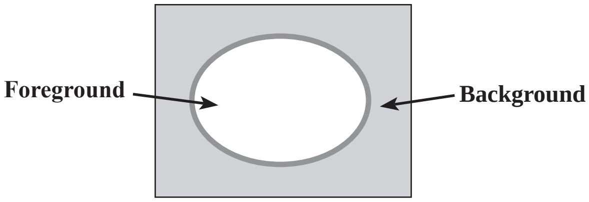
Using a laser scanner, Tagged Image File Format (TIFF) images of the gene spots are obtained. Gene spots are often composed of characteristic imperfections such as irregular contour, donut shapes, artefacts, and low or heterogeneous expression. It is often assumed that the signal observed is a combination of the true signal or foreground signal (from the specific hybridization of interest) and the background signal (due to non-specific hybridization and/or contamination). Estimation of background intensity is generally considered necessary for the purpose of performing background correction. The standard approach is simply to subtract the background estimate directly from the spot intensity, with the aim of improving accuracy (reducing bias). However, the background signal may increase due to dust, fibres, fingerprints, auto fluorescence of the coated glass, hybridization problems resulting from dehydration near the edge of the coverslips, or residual effects from inadequate washing. Exploratory data analysis has been the tool of choice for detection of problematic arrays. However, the largest values are orders of magnitude larger than the bulk of the data and these results in a non-informative image (12). A simple solution is to examine an image plot of the log intensities, as it is demonstrated with the data from a large acute lymphoblastic leukemia study (17) described in Figure 3.

In cDNA, usually two individual heterogeneous mRNA samples are labelled with either a red-fluorescent dye Cy5 (referred as R) or a green –fluorescent dye Cy3 (referred as G), respectively. Then, they are mixed and hybridized to the arrayed cDNA sequences. The ratio of the fluorescence measurements for red and green dye, obtained from TIFF files describes the relative abundance of the corresponding mRNA. The phase of image processing which attains two values for intensities, R and G, and one value, R/G, of relative abundance for a single spot can be divided into three basic steps:

i) **Addressing or gridding**. Identifying the areas (assigning coordinates) that belong to spots in an image, usually K rectangular zone (default K=16). The combined area of a spot and its background is called the target area (or target Patch), as follows:



ii) **Segmentation**. Partitioning the target area of every spot in two distinct regions, as foreground (the spot itself) and background:



Usually, for each grid (equally spaced zone), the lowest 2% of probe intensities are used to compute a background value for that grid.

iii) **Intensity extraction**. Extracting scalar values for the absolute and relative's intensities. In cDNA, the intensities of R and G and the ratio R/G (or log2(R/G).

Most methods assume circular shapes and require manual alignment of the grid location. In cDNA microarray images, the assumption of circular spot shape is not justified due to artifacts caused by the printing process and the hybridization technique. One of the major difficulties is that each cDNA clone usually contains several hundreds of pixels, and the locations and shapes of these spots may vary depending on the quality of the experiment and the scanner. Some scanners have higher sensitivity than others; the background values for the same slide will differ depending on which scanner is used to acquire the image. Therefore, the errors in image analysis can be produced from different sources and can be classified as follows (18-19):

1. *Variable size*: different diameters

2. *Variable contours*: sickle shape, donut shape, oval or pear shape, scratched or interrupted shape

3. *Normalization*: adjustment for effects which arise from variation in the technology or between the printed probes high background and/or low foreground.

4. *Spatial artefacts*: smeared or incorrectly segmented areas, caused by dirt on the slide or slide treatment.

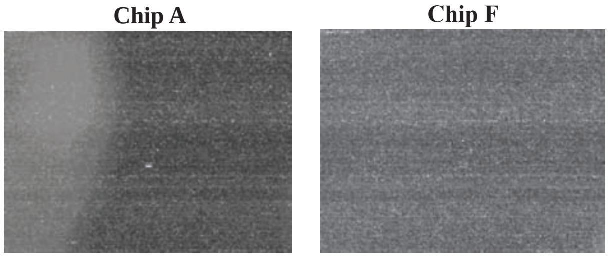
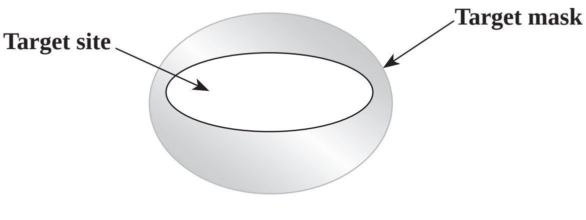


Figure 3. Image of probe intensities in log-scale for two arrays (or chips using the package ALLMLL (www.bioconductor.org). The log- scale in chip A demonstrates a strong spatial artefact (smeared or incorrectly segmented area) not seen in the chip F.

Several methods for image analysis have been adapted for microarray to deal with its specific problems. In general, they can be classified into spatial and distributional methods. Spatial methods try to capture the shape of a spot; one of these methods is to fix a circle with a constant diameter to all the spots in the image, which is clearly not satisfactory for all the spots. One of the distributional methods is based on a threshold value using the Mann-Whitney test; pixels are classified as foreground if their value is greater than the threshold and as background otherwise. Another distributional method is based on the histogram, it defines the background and the foreground as the mean (or median) intensities between some predefined percentiles values; by default, these are the 5th and 20th percentiles for the background and 80th and 95th percentile for the foreground. By computing the foreground intensities from a higher percentile range, this method usually yields a higher estimate of the foreground. The main advantage of these methods is their simplicity. Furthermore, as the resulting spots are not necessarily connected, these methods may perform well with donut-shaped spots. However, a major disadvantage is that quantification is unstable when a large target mask is set to compensate for spot size variation, as follows:



A recent method for segmentation in cDNA that uses a parametric approach for the densities distribution of the pixels intensities of the background and foreground has been published (20). For the background, it has been proposed the bivariate distribution as the underlying pixel intensities distribution, whose marginal densities for the Cy5 dye (Red, R) and Cy3 dye (Green, G) are independent three parameters gamma: α_i shape parameter, β_i scale parameters, and γ_i location parameter. For the underlying distribution for the foreground, the bivariate t distribution, with the following location parameters

$$\mu_R = \alpha_R \big/ \beta_R + \gamma_R + \o_R,\ \mu_G = \alpha_G \big/ \beta_G + \gamma_G + \o_G,\ \o_i \geq 0,$$

is adopted. Since the mean of the foreground intensity must be larger than the mean of the background, μ_i was parameterized as the mean of the background plus the nonnegative parameter. Also, it was assumed that the observed intensities (y_i's) are independent and identically distributed realizations from the following mixture of densities:

$$f(y; \Psi) = \pi_B * f_B (y; \alpha_i, \beta_i) + \pi_F * f_F * (y; \mu_i, \Sigma, \upsilon)$$

where $\Psi = (\alpha_i, \beta_i, \mu_i, \Sigma, \upsilon, \pi_i)$, π_i is the probability that a pixel belongs to background or foreground with $\pi_B + \pi_F = 1$, and f_B and f_F are the densities for background and foreground, respectively. To obtain the maximum likelihood estimates of the parameters Ψ, the EM algorithm was implemented. In the E-step, the posterior probability that y_j belongs to the *i*th component (background or foreground) of the mixture, given the value $\Psi^{(k)}$ of Ψ after the *k*th iteration, was expressed as follows:

$$\tau_{\ddot{y}}^{(k)} = \pi_i^k f_i (y_j ; \Psi^{(k)}) \Big/ f(y_j ; \Psi^{(k)})$$

When a solution was found in the EM algorithm for the posterior probability, assuming $\hat{\tau}_{\ddot{y}}$, a nonparametric kernel estimate was obtained to encourage neighbouring pixels, either in the background or in the foreground of the rectangle containing the spot. The authors of this proposed method (20) concluded that the new method for gridding, segmentation, and estimation to cDNA microarray images provided better segmentation results in spot shapes as well as intensity estimation than Spot and spot Segmentation R language software.

For oligonucleotide arrays, the suggested purpose of the MM probes was that they could be used to adjust the PM probes for probe-specific non-specific binding by subtracting the intensities of the MM probe from the intensities of the corresponding PM probes. The reason for including an MM probe is to provide a value that compromises most of the background cross-hybridization and stray signal affecting the PM probe. It also contains a portion of the true signal. If the MM value is less than the PM value, it is physically possible to estimate for background. One of the concerns in the MM adjustment is that some MM probes may have intensities higher than their corresponding PM probes. Thus, when raw MM intensities are subtracted from PM intensities, negative expression values can occur, which makes no sense, because an expression value should not be below zero (12). To solve this problem, an idealized value is estimated based on the knowledge of the whole probe set, an ideal mismatch for probe pair "j" in probe set "i" is obtained as follows (21):

$$\mathrm{IM}_{ij} = \begin{cases} \mathrm{MM}_{ij}, & \mathrm{IM}_{ij} < \mathrm{PM}_{ij} \\ \dfrac{PM_{ij}}{2^{SB_i}}, & \mathrm{MM}_{ij} > \mathrm{PM}_{ij} \text{ and } \mathrm{SB}_i > \tau \\ \dfrac{PM_{ij},}{2^{\left(\frac{\tau}{1+\frac{\tau - SB_i}{\kappa}}\right)}}, & \mathrm{MM}_{ij} \geq \mathrm{PM}_{ij} \text{ and } \mathrm{SB}_i > \kappa \end{cases}$$

where SB_i is the specific background for each probe set (robust average of log2(PM/MM) among all pair probes in each probe set), τ and κ are tuning constants, referred to as the contrast τ (with default values of 0.03) and the scaling κ (with default value of 10). The first case where the mismatch value provides a probe-specific estimate of stray signal is the best solution. In the second case, the estimate is not probe-specific, but at least provides information specific to the probe set. The third set case involves the least informative estimate, based only weakly on probe-set specific data. The adjusted PM intensity is obtained by subtracting the corresponding IM from the observed PM intensity (12).

1.2) Normalization

Normalization aims to adjust microarray data for effects which arise from variation in the technology rather than from biological differences between the RNA samples or between the printed probes. The need for normalization arises naturally when dealing with experiments involving multiple arrays. Imbalance between the red and green dyes may arise from differences between the labelling efficiencies or scanning properties of the fluorescence, complicated perhaps by the use of different scanner settings. The dye-bias will also generally vary with spatial position on the slide. Positions on a slide may differ because of differences between the print-tips on the array printer, variation over the course of the print-run, non-uniformity in the hybridization, or from artefacts on the surface of the array which affect one colour more than the other. Finally, differences between arrays may arise from differences in print quality, from differences in environmental conditions when the plates were processed, or simply from changes in the scanner settings (18, 22).

cDNA microarrays generate one- or two-channel data. In the latter, the arrays are hybridized to a mixture of two samples, each labelled with two dyes (Cy3, Cy5). In one channel use, each array is hybridized to a single sample, labelled with a single dye. The two-channel data allow for internal correction of a number of commonly occurring artefacts (i.e., defective print tips and fainting of the signal in large regions of an array). Variation across arrays will reflect the genetic, experimental, and environmental differences under study, but will also include variations introduced during sample preparation, manufacturing of the arrays, and processing of the arrays (labelling, hybridization, and scanning). A first step to explore the possibility of data normalization is to draw a scatter plot between the M=log2(R/G) and A=log2 ($\sqrt{RxG}$), where, R and G for the background-corrected red and green intensities for each spot. It is convenient to use base-2 logarithms for M and A so that M is units of 2-fold change and A is in units of 2-fold increase in brightness. On this scale, M=0 represents equal expression, M=1 represents a 2-fold change between the RNA samples, M=2 represents a 4-fold change, and so on (22). If M-A plots exhibit any obvious curvature deviating from the horizontal line at zero, normalization is recommended (Figure 4).

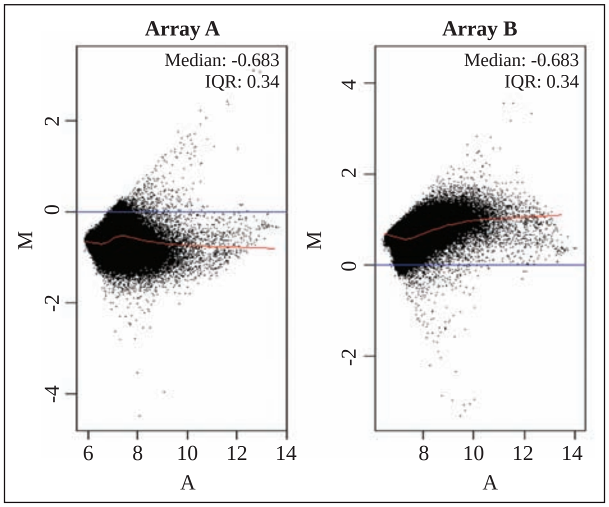


Figure 4. MA Plot in two arrays plotted with common pseudo-array reference and the loess trend. Original data from Ross, et al. (17), as described in figure 3.

In oligonucleotide arrays, different exploratory plots can be used to detect obscure sources of variation and the need for normalization. For example, one may consider direct array-to-array comparison of PM values via box plots of $\log_2$(PM), $\log_2$(MM), $\log_2$(PM/MM), or PP-MM. Alternatively, one can explore intensity–related biases for each pairwise array comparison via M-A plots of M=$\log_2$(PM_k/PM_l) versus abundance A=$\log_2(\sqrt{PM_k{}^*PM_i}\,)$ for two different arrays .

The process of cDNA normalization can be separated into two main components: location and scale. In general, methods for location and scale normalization adjust the centre and spread of the distribution of log-ratios. The normalized intensity log-ratios M_{norm} are generally given by:

$$M_{norm} = \frac{M - \ell}{s} \text{ where } M = \log_2 \frac{Cy5}{Cy3},$$

ℓ and s denote the location and scale normalized values, respectively. The two-channel normalization method is recognized by the definition of the forms ℓ and s. For example, in global median normalization, the parameter ℓ is assumed to be the same for all spots on an array, whereas in global A-dependent normalization it is assumed to be a smooth function of A = $\log_2(\sqrt{Cy5^*Cy3}\,)$, and the function is estimated using the scatter-plot smoother loess (23).

In oligonucleotide arrays, normalization methods have been classified as complete data method and baseline array method (18). The complete data methods combined information from all arrays to form the normalization relation. The baseline array method uses information from one array as a reference; for example, the array having the median of the median intensities.

The Cyclic loess is a complete data method, which is based on the M versus A plot. The procedure is as follows:

1. Calculate $M_k = \log_2(x_{ki}/x_{kj})$ and $A_k = \log_2\sqrt{x_{ki}{}^*x_{kj}}$ for every probe k in any two arrays (i,j), where x's are the intensities.
2. Fit the loess curve of M versus A, $\hat{M}_k$
3. Calculate the normalization adjustment: $M'_k = M_k - \hat{M}_k$
4. Adjust the probe intensities:

$$x'_{ki} = 2^{A_k + \frac{M'_k}{2}}, \; x'_{kj} = 2^{A_k - \frac{M'_k}{2}}$$

When more than two arrays are considered, the method is extended to look at all distinct pairwise combinations. So, after looking at all pairs of arrays for any array k, there are p-1 adjustments, where p is the number of arrays. Then, the adjustments are equally weighted and applied to the set of arrays. Bolstad, et al. (18) has reported that this method could be somewhat time consuming.

Another complete data method is the quantile normalization (18). The goal of this method is to make the distribution of probe intensities for each array in a set of the same arrays, that is, to impose the same empirical distribution of intensities to each array. The motivation of this method is that a quantile-quantile plot (q-q plot) shows that the distribution of two data vectors is the same if the plot is a straight diagonal line; thus, to make a set of data to have the same distribution, we need to project the points of our quantile plot onto the diagonal; this method is extended to n-dimension. The quantile method is a specific case of the transformation $x'_i = F^{-1}[(G(x_i)]$, where G is estimated by the empirical distribution of each array and F is the empirical distribution of the averaged sample quantiles. The method will be adequate when distribution of the normalized data will around zero using the MA plot (Figure 5). The extension of the quantile method can be implemented where F-1 and G are more smoothly estimated. However, previous studies by Bostand, et al. (18) have shown that the performance of the quantile normalization is slightly better than the cyclic loess.

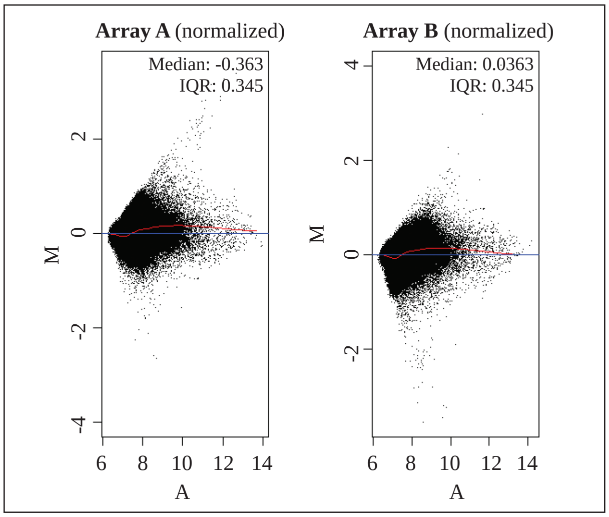


Figure 5. MA Plot in two normalized arrays plotted with common pseudo-array reference and the loess trend. Data source from Figure 4.

Scaling method is a baseline array method. This is the standard normalization method in Affymetrix (version 4.0 and 5.0) that is carried out on probe set expression measures. Bolstand, et al. (18) assess this method at probe level. This method chooses a baseline array, in particular, the array having the median of the median intensities. Then, all arrays are normalized to this 'baseline' as follows:

1. Compute the trimmed mean intensity (excluding highest and lowest 2% probe intensities) of the baseline array, called $\tilde{x}_{base}$
2. Compute the trimmed mean intensity (excluding highest and lowest 2% probe intensities) of the array "i", called $\tilde{x}_i$
3. Let, $\beta_i = \dfrac{\tilde{x}_{base}}{\tilde{x}_i}$
4. Then, the intensities for the normalized array would be: $x' = \beta_i x_i$

This is equivalent to selecting a baseline array and, then, for every other array fitting a linear regression, without an intercept term, removing the highest and lowest intensities. Affymetrix has proposed using the scaling normalization after the computation of expression values, but it may also be used on probe-level data.

1.3) Summarization

Before the statistical analysis is performed, a probe reduction has to be defined in oligonucleotide arrays, that is, to combine the multiple probe intensities for each probe set to produce an expression value for each gene. Efron, et al. (2001) (21) have evaluated the probe reduction using the following expression:

$$M_i = \mathit{mean}\{\log(\mathrm{PM}_{\ddot{y}}) - c*\log(MM_{\ddot{y}})\};\ j=1,..,20 \text{ probe}$$

Results from a study to assess the transcriptional responses to ionizing radiation have shown that the probe reduction with c = 0.5 has a mild advantage over c = 1 or c = 0 (24).

Another method for probe reduction was developed by Irizarry, et al. (2003) based on a log scale linear additive model, which is referred to as the log scale robust multi-array analysis (RMA) (21). The motivation of the model is due to the large variation at probe level data, since probes with larger mean intensities have larger variances. The RMA model is defined as follows:

$$T(PM_{\ddot{y}}) = e_i + a_j + \varepsilon_{\ddot{y}}$$

where T represents the transformation that background corrects, normalizes, and $\log_2$ PM intensities, e_i represents the $\log_2$ scale expression value found on array i = 1,…,I, a_j represents the log scale affinity effects for probe j = 1,…, J, and $\varepsilon_{\ddot{y}}$ represents the error. A robust linear fitting procedure, such as median polish, has been used to estimate the log scale expression values e_i. This model does not consider the subtraction of the MM intensities because the empirical results have demonstrated that mathematical subtraction does not translate to biological subtraction. The authors have concluded that substantial benefits of using the RMA measure instead of the GeneChip technology where the computer software automatically calculates average difference values (22).

II) Statistical Analysis

Statistical methods for microarray data analysis can be classified in the following two major groups: i) methods that identify differentially expressed genes, and ii) methods that classify the functional dependency of genes. The objective of the first method is to identify those genes that are consistently expressed at different levels under different conditions using the classical statistical test (t-test, ANOVA, Mann-Whitney test,…) controlling the probability of false declaration (25-27). The second method pretends to identify the shared patterns of expression across genes to classify new diseases of subtype of diseases for subsequent validation and prediction, and, ultimately, to develop individualized prognosis and therapy, using cluster analysis methods (28-29). In this paper, we describe and compare some of the methods used for pairwise comparison from genomic experiments, controlling the false discovery rate and the assumption in the density distribution (normal vs. empirical distribution) of the statistics test for the unaffected genes.

The simplest experimental design is the comparison of two groups (diseased persons vs. healthy; treatment A vs. treatment B; exposed vs. not-exposed) in microarray data analysis (30). Usually, the databases for this scenario are described in row and columns, where rows indicate the genes and the columns, the arrays associated to each group (Table 1).

Table 1. Data structure in microarray data analysis for pairwise comparison

	Diseased	**Healthy**				
Genes	**Array 1**	**…..**	**Array n_1**	**Array 1**	**…….**	**Array n_2**
1	X_{11}		X_1n_1	Y_{11}		Y_1n_2
2	X_{21}		X_2n_1	Y_{21}		Y_2n_2
:						
:						
m	X_{m1}		X_mn_1	Y_{m1}		Y_mn_2

2.1) Ordinary t-test

The statistical methods used to identify differentially expressed genes in the two groups are based on fold change, i.e., $\overline{X}$(diseased) – $\overline{Y}$ (healthy) for any gene. In order to assess the significance of these fold-changes, some researchers have used the ordinary *t*-statistics for each gene, as follows:

$$t_j = \frac{\overline{X}_j - \overline{Y}_j}{\sqrt{Var(\overline{X}_j - \overline{Y}_j)}} = \frac{\overline{X}_j - \overline{Y}_j}{\overline{s}_j} \sim t_{df}$$

The distribution of the tj will be expected to be symmetrically distributed around zero and their respective *p*-values will have an inverse J-shape distribution, as it is demonstrated with the data of the study of Chriaretti, et al. (31), where BCR/ABL cells are compared with cytogenetically normal cells (Figure 6).

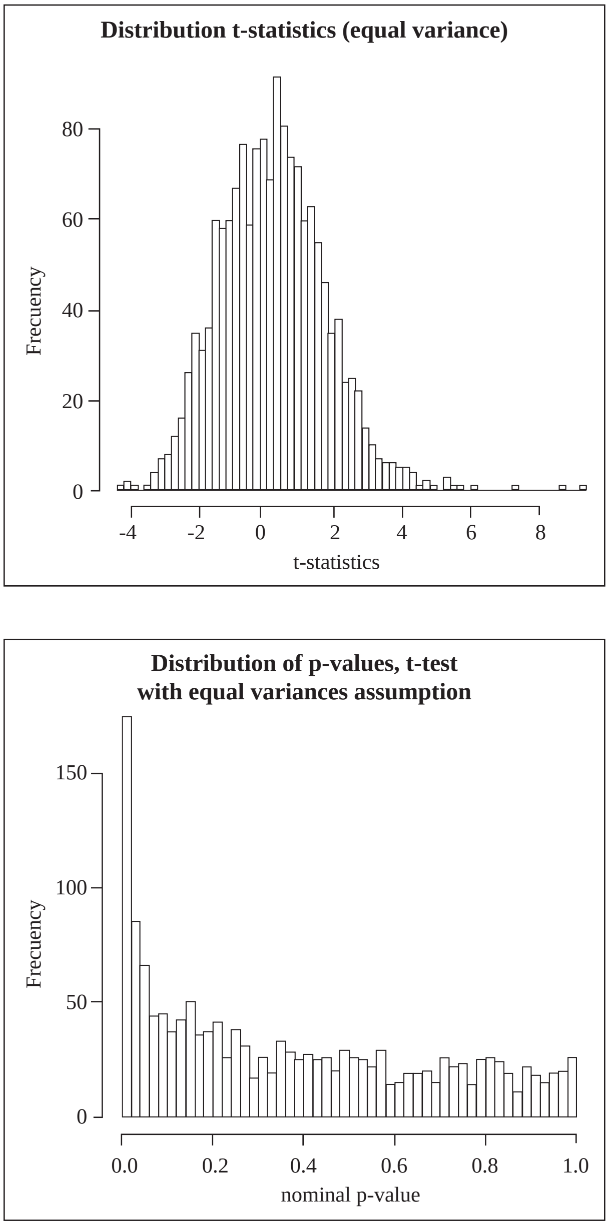


Figure 6. Data on Acute Lymphoblastic Leukaemia from Chiaretti et al. (31), where BCR/ABL (n=37) cells are compared with cytogenetically normal cells (n=42). Data were normalized with RMA (intensities are on log2-scale). For this example, intensities above 100 were selected in at least 75% of the samples, and the interquartile range of log2-intensities >0.5 (m=1541 genes). The p-values distribution shows: 296 genes with p <0.05, 23 genes with p<.01, 45 with p <0.001, 21 with p<0.0001, and 10 with p<0.00001.

2.2) Modified t-statistics

Similar statistical tests have been used with an adjustment in the sample standard deviation as follows (modified *t*-statistics):

$$\hat{t}_j = \frac{\overline{X}_j - \overline{Y}_j}{\bar{s}_j + \bar{s}_0} \sim t_{\upsilon_j}$$

where $\ddot{s}_0^{2}$ is the 90th percentile of the sample standard deviations (or any other percentile, ie., 50% among all genes. Specifically, $\bar{s}_0$ is chosen to make the coefficient of variation of t_j approximately constant as a function of $\bar{s}_j$; this has the added effect of dampening large values of t_j that arise from genes whose expression is near zero (24, 32).

2.3) Moderate t-test

Linear models for microarray (**Limma**) data is another approach for analysing differential expression (33). For example, the expectation of the expression for gene "j" can be defined for pairwise comparison, as follows:

$$E[y_j] = X\alpha_j = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (\mu_{Yj}\ \alpha_j) = \mu_{Yj} + \alpha_j = \mu_{Yj} + (\mu_{Xj} - \mu_{Yj})$$

$$\text{var}(y_j) = W_j\ \sigma_j^2 \text{var}(\hat{a}_j) = V_j\ s_j^2$$

where y_j contains the expression data for the gene j, *X* is the design matrix of a full column rank to define the reference group, $\mathbf{W}_j$ is a known non-negative definite weight matrix, $\mathbf{V}_j$ is a positive matrix not depending on s_j^2 and α_j is a vector of coefficients to determine the effect groups. To take advantage of the parallel structure whereby the same model is fitted to each gene, the variances σ_j^2 across the genes, given a prior distribution ($\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\chi_{d_0}^2$), is estimated as follows:

$$\tilde{s}_{gj}^2 = E(\sigma_g^2 | s_j^2) = \frac{d_0 s_0^2 + d_j s_j^2}{d_0 + d_j}$$

where $\tilde{s}_{gj}^2$ is called the posterior mean of σ_j^2 given s_j^2 and s_0^2 is the prior estimator σ_j^2 with d_o degrees of freedom; so, the t statistics to compare two groups is defined as follows:

$$\tilde{t}_{gj} = \frac{\overline{X}_j - \overline{Y}_j}{\bar{s}_{gj}\sqrt{\upsilon_{gj}}} \sim t_{d_0 + d_j} | H_0 : \mu_{gx} = \mu_{gy}$$

where υ_{gj} is the *j*th diagonal element of $\mathbf{X}^T\mathbf{V}_g\mathbf{X}$. The statistics is called the moderated t-statistics and represents a hybrid classical/Bayes approach in which the posterior variance has been substituted in the ordinary t-statistics in place of the usual sample variance. If d_o=0, the moderate t-statistics reduces to the ordinary *t*-statistics (33).

2.4) Bayesian Methods

Bayesian methods for differential gene expression with mixture model approach have also been applied, as follows (34):

$$\hat{z}_g = \Pr(z_g = 1 | x_g, y_g, p, \psi)$$

$$= \frac{p p_A (x_g, y_g | \psi)}{p^* p_A(x_g, y_g|\psi) + (1-p)^* p_0(x_g, y_g|\psi)}$$

Where z_g indicates the posterior probability of change ($\mu_{gx} \neq \mu_{gy}$), with a prior probability of change *1-p* (*p* indicates the probability of no change), and p_a and p_o denote the joint marginal density of the measured intensities of gene g under the assumption of differential expression ($\mu_{gx} \neq \mu_{gy}$) and no differential expression ($\mu_{gx} = \mu_{gy}$), given ψ (vector of unknown hyperparameters).

2.5) Rank-sum statistics

Another statistic that could be formed for differential expression is the rank-sum statistics. For example, in the two groups comparison, let r_{ji} be the rank of the *i*th expression within gene "j"; then, the rank-sum statistics for this gene is: $r_j = \sum_{\text{group1}} r_{ji}$, where the summation is taken over the genes in group 1. An extreme r_j value in either direction would indicate a difference in gene expression. The statistics t_j tests for a difference in means, whereas r_j tests for a more general difference in distribution. Usually, one is more concerned with a difference in mean gene expression, so t_j is a more powerful statistics to use for this test (32).

2.6) Permutation methods

To control the correlation among genes, it has been suggested to use permutation method by estimating the t-statistics under the null hypotheses by permutations of sample labels (Table 2).

Table 2. Example of one permutation for pairwise comparison

Original Dataset: T=treatment, C=Control

Gene	T	T	T	C	C	C
1	123	78	56	**34**	**45**	**89**
2	34	48	90	**24**	**46**	**23**
3	23	78	56	**58**	**78**	**15**

One possible permutation (change the label, no the data)

Gene	C	C	T	C	T	T
1	**123**	**78**	56	**34**	45	89
2	**34**	**48**	90	**24**	46	23
3	**23**	**78**	56	**58**	78	15

Under this method, the *p*-value for each gene is given as the fraction of permutations yielding a test statistic that is at least as extreme as the observed one. If the null distribution of t_j is calculated on the basis of just the data on the *j*th gene, then it suffers from a granularity problem; for example, there are only ten ways to divide six microarrays into equal sized groups. The null distribution has a resolution on the order of the number of permutation. If we perform B permutation, then the *p*-value will be estimated with a resolution of 1/B. When we combine the permutations across the genes and assume that each gene has the same null distribution, then the resolution will be 1/(m*B) and the p-value will be (13, 35):

$$p^B_j = \sum_{b=1}^{B} \frac{\#\{ j:|t^{(b)}_{0\,j}| \geq |t_j|, j=1,...B\}}{m*B}$$

where $t^{(b)}_{0\,j}$ is the null version t_j after the *b*th permutation of the class labels. The drawback of pooling the statistics $t^{(b)}_{0\,j}$ across the genes is that we are assuming that the null distribution is true for all genes, but only a proportion of $\pi_o = \frac{m_o}{m}$ are null. Based on this permutation method, Westfall and Young (1993) make an adjustment in the *p*-values to control the family-wise error rate, as follows (36):

$$\tilde{p}^B{}_j = \Pr(\min_{k=1,2,...,m} p^B{}_k \leq p_j \mid H_0) = \#\{b : \min p^B_k \leq p_j\}/B$$

For example, suppose the minimal unadjusted *p*-value, p_j, was .00005, then, among the randomized data sets (permuted sample labels) count how often the minimal *p*-value is smaller than 0.00005; if this appears in 2% of all case, $\tilde{p}_{\min}$=.02 (Table 3). For description of the gene products in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner you can visit the web site of the GO project (www.geneontology.org).

Table 3. The most significant genes using different methods for computing the p-values. Data from figure 6. The permutation methods were computed with B=1,000,000.

	***p*-values**			
Genes	**Ordinary**	**Limma**	**Permutation**	**Westfall/Young**
ABL1	3.76E-14	2.06E-14	0.0000010	0.0000010
ABL1	4.79E-13	2.00E-13	0.0000010	0.0000010
ABL1	2.45E-10	8.62E-11	0.0000010	0.0000020
KLF9	2.79E-08	7.98E-09	0.0000010	0.0000170
AHNAK	0.0000003	0.0000001	0.0000010	0.0002510
ZNF467	0.0000005	0.0000010	0.0000030	0.0007310
FYN	0.0000011	0.0000006	0.0000030	0.0006890
CASP8	0.0000012	0.0000006	0.0000020	0.0013550
TUBA1	0.0000013	0.0000006	0.0000020	0.0008240
FHL1	0.0000060	0.0000029	0.0000050	0.0047650
FYN	0.0000135	0.0000099	0.0000100	0.0096320
SV2A	0.0000154	0.0000108	0.0000160	0.0208330
CRIP1	0.0000326	0.0000161	0.0000320	0.0267770
TPD52L2	0.0000481	0.0000365	0.0000600	0.0406950
NA	0.0000511	0.0000556	0.0000580	0.0575760
ENG	0.0000550	0.0000393	0.0000740	0.0567370
CD97	0.0000564	0.0000364	0.0000550	0.0448570
SOCS2	0.0000611	0.0000355	0.0000240	0.0389150
GYPC	0.0000742	0.0000491	0.0001190	0.0775450
FSCN1	0.0000829	0.0000487	0.0000810	0.0567310

2.7) Significance Analysis of Microarrays

The significance analysis of microarray (SAM) is another method for group comparison using the permutation method, but rather than using the standard rule of the form $|t_j| > c$ to call genes significant (i.e. having symmetric cut points ±t), SAM derives cutoff points c_1 and c_2 and uses the rejection rule $t_j < c_1$ or $t_j > c_2$. The procedure of SAM is described in the following steps, using the modified t-statistics (37): i) Compute the modified statistics: $t_1, t_2, ….,t_m$; ii) Compute the ordered statistics: $t_{(1)}, t_{(2)}, ….,t_{(m)}$; iii) Take B sets of permutations of the group labels:

Permutation	**Ordered statistics**
1	$t^{*1}{}_{(1)}, t^{*1}{}_{(2)}, ….,t^{*1}{}_{(m)}$
2	$t^{*2}{}_{(1)}, t^{*2}{}_{(2)}, ….,t^{*2}{}_{(m)}$
:	:
B	$t^{*B}{}_{(1)}, t^{*B}{}_{(2)}, ….,t^{*B}{}_{(m)}$

iv) Estimate the expected order statistics by:

$$\bar{t}_{(j)} = \frac{\sum_{b=1}^{B} t^{*b}_{(j)}}{B} \text{ for j=1, 2, ..., m}$$

v) Plot the observed $t_{(j)}$ score versus the expected $\bar{t}_{(j)}$ score. For a fixed threshold Δ, starting at the origin ($\bar{t}_{(j)} = 0$, $t_{(j)} = 0$), and moving up to the right, find the first $j = c_2$ such that $t_{(j)} - \bar{t}_{(j)} \leq \Delta$. All genes past c_2 are called "significant positive". Similarly, start at the origin, move down to the left and find the first $j = c_1$ such that $t_{(j)} - \bar{t}_{(j)} \geq \Delta$. All genes past c_1 are called "significant negative" (Figure 7). SAM is a more powerful test in situations where more genes are overexpressed than underexpressed (32).

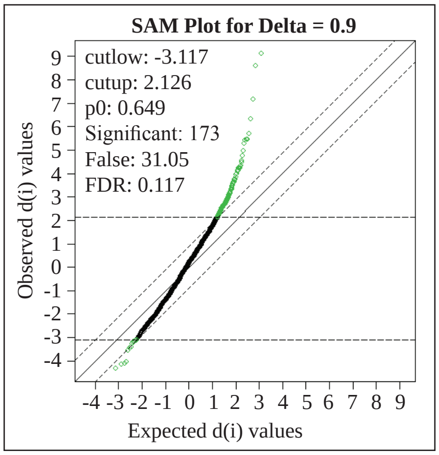


Figure 7. SAM plot. Data on Acute Lymphoblastic Leukaemia as in figure 7. With Δ=.9 and B=1000, then c_1= -3.11, c_2= 2.12, and 173 "significant genes" were found using the SAM method.

III) False Discovery Rate

For the purpose of statistically comparing the expressions in a single gene, one is usually concerned with the Type I error (probability of rejecting the null hypothesis when it is true), and Type II error (probability of accepting the null hypothesis when it is false). When differential expressions are assessed for multiple genes, multiple tests are performed. The most commonly controlled method when testing multiple hypotheses is the family wise error rate (FWER), which is the probability of yielding one or more false positive out of all hypotheses tested. For example, the Bonferroni method declared each test significant if $p < \alpha/m$, where m is the total number of tests, it then follows that FWER $\leq \alpha$. Although this method is quite generally applicable, it is usually not a good choice for microarray studies because it has a very low power, i.e. the probability of correctly identifying differentially expressed genes is very low; so many potentially interesting genes may be missed (16). One of the methods for controlling the false positives among the genes differentially expressed and those declared significant was developed by Benjamin and Hochberg (38). To illustrate this method, the following contingency table is used:

Genes differentially expressed	Declared non-significant	Declared Significant	Total
Non-True	U	V	mo
True	T	S	m1
Total	U+T	V+S	M

V+S is an observable variable, while U, V, S, and T are unobservable random variables. The errors committed by falsely rejecting null hypotheses can be viewed through the unobserved random variable Q=V/(V+S), proportion of the rejected null hypotheses which are erroneously rejected. When V+S =0, Q is defined equal to zero. The expectation of Q was defined as the False Discovery Rate (FDR) by Benjamini, et al. (38):

$$\text{FDR}=E(Q)=E\{V/(V+S)\}$$

In a critical review on microarray studies for cancer outcome, only 9 of 23 studies published in 2004 controlled the number of false-positive differentially expressed genes (39).

Different approaches have been used to estimate the FDR (24, 27, 40). Efron, et al. (24) proposed a mixture density of the statistics (Z) to compare the expression of two populations (diseased vs. healthy) to estimate the local FDR, which is an empirical Bayes version of the Benjamini & Hochberg (38) methodology focusing on densities, as follows:

$$\text{fdr} = \pi_0 \frac{f_0(Z)}{f(Z)}$$

where π_0 is the probability that a gene is unaffected, $f(Z) = \pi_0 * f_0(Z) + (1 - \pi_1) * f_1(Z)$ is the mixture density, $f_0(Z)$ the density of Z for unaffected genes (i.e. the normal distribution) and $f_1(Z)$ the density of Z for affected genes. The ratio $\frac{f_0(Z)}{f(Z)}$ is taken from the set of $\{Z_i\}$ and the empirical distribution of Z_i using permutation analysis, and $\hat{\pi}_0 \equiv \min_z\{\frac{f(Z)}{f_0(Z)}\}$.

Storey, et al. (41) proposed the following estimation of FDR:

$$F\hat{D}R(\alpha) = \frac{\hat{\pi}_0(\lambda) * m * \alpha}{\#\{p_i \leq \alpha\}}$$

where $\hat{\pi}(\lambda) = \frac{\#\{p_i > \lambda; i = 1,\ldots,m\}}{m(1-\lambda)}$, λ is a tuning parameter, and α is a threshold to declare significant results when $p < \alpha$. If λ=0, then, $\hat{\pi}(\lambda)$=1 which is going to be too conservative in genome-wide data sets; however, if we set λ close to 1, the variance of $\hat{\pi}(\lambda)$ will increase making the estimate of the FDR's more unreliable. Due to the possibility of no significant results (V+S=0), the positive false discovery (pFDR) was defined as follows:

$$pFDR = E\left[\frac{V}{V+S} \mid V + S > 0\right]$$

The positive term of pFDR describes the fact that it was conditioned on at least one positive finding having occurred. Storey (41) has proposed to use a Bayesian approach to estimate pFDR, when m identical tests are performed with statistics T_i (i.i.d random variables) to assess H_o vs. H_1, as follows:

$$pFDR(\Gamma) = \Pr(H = 0 \mid T \in \Gamma)$$

$$= \frac{\pi_0 * \Pr(T \in \Gamma \mid H = 0)}{\pi_0 * \Pr(T \in \Gamma \mid H_0) + \pi_1 * \Pr(T \in \Gamma \mid H_1)}$$

$$= \frac{\pi_0 * \Pr(\text{Type I error on } \Gamma)}{\pi_0 * \Pr(\text{Type I error on } \Gamma) + \pi_1 * \Pr(\text{Power of } \Gamma)}$$

where Γ is the significant region, $T_i|H_i \sim (1-H_i)*F_0+H_i*F_1$ (mixture density) for some null distribution F_0 and alternative distribution F_1, $H_i \sim \text{Bernoulli}(\pi_i)$, $\pi_0 = 1 - \pi_1$ is the implicit prior probability that a group of genes are not differentially expressed. To assess the significance of each test, an analogous quantity of the p-value was proposed by Storey (41) in terms of the pFDR, as follows:

$$q\text{ - }value(t) = \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} [\text{pFDR}(\Gamma_\alpha)]$$

$$= \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} [\Pr(H_0 \mid T \in \Gamma_\alpha)]$$

The *q*-value for a particular gene is the expected proportion of false positives incurred when calling that gene significant. The *q*-value, a Bayesian version of the *p*-value (posterior Bayesian *p*-value), is a measure of the strength of an observed statistics with respect to the pFDR (41).

Recently, McLachlan, et al. (27) published another method to estimate the FDR by using a normal mixture approach. The steps to reach this estimation are:

i) Assuming that M_r genes are selected as significant with the following rule: $\hat{\tau}_0(z_j) \le c_0$ where $\hat{\tau}_0(z_j) = \hat{\pi}_0 \hat{f}_0(z_j) / \hat{f}\,(z_j)$ is the estimate of the posterior probability that the jth gene is not differentially expressed, $z_j = \Phi^{-1}(1 - p_j)$, p_j is the *p*-value of the statistics to be used to assess the evidence against the null hypotheses (ordinary t-statistics, modified t, moderated t,…) for each gene, Φ is the N(0,1) distribution function, π_0 is the prior probability of a gene belonging to the set of genes that are not differentially expressed, f_o (z_j) is the null density of z_j, $f(z_j)$ is the mixture density of z_j defined as (42):

$$f(z_j) = \pi_0 f_0(z_j) + (1 - \pi_0) f_1(z_j)$$

ii) Estimate the FDR as follows:

$$F\hat{D}R = \frac{\sum_{j=1}^{M} \hat{\tau}_0(w_j) I_{[0,c_0]}(\hat{\tau}_0(w_j))}{M_r}$$

where $I_{[0,c_0]}(\hat{\tau}_0(w_j))$ is an indicator function, which is one if $\hat{\tau}_0(w_j) \le c_0$ otherwise is zero, and $M_r = \sum_{j=1}^{M} I_{[0,c_0]}(\hat{\tau}_0(w_j))$ Usually, it is assumed that $f_i(z_j)$ follows the normal density with parameters $\mu_0 = 0$, $\sigma^2_0 = 1$ for $f_o(z_j)$ and μ_1, σ^2_1 for $f_i(z_j)$; however, this assumption is not always true across genes (24). If the $f_i(z_j)$'s are the density of the normal distribution, the null hypotheses is called the theoretical null hypotheses; if $f_i(z_j)$'s are the empirical distributions of z_j, then the null hypotheses is called the empirical null hypotheses. In McLachlan, et al. (27), the estimation of the parameters $(\pi_0, \mu_i, \sigma^2_i)$ were affected by maximum likelihood via the EM algorithm, using the EMMIX program with the following initial value of π_0:

$$\pi_0^{(0)}(\xi) = \frac{\#\{z_j : z_j < \xi\}}{M * \Phi(\xi)}$$

for an appropriate value of ξ; as a consequences for the theoretical densities, the initial values for the mean and the variance of the alternative hypotheses were:

$$\hat{\mu}_1^{(0)} = \bar{z}/(1 - \pi_0^{(0)}) \text{ and}$$
$$\hat{\sigma}_1^2 = \{s_z^2 - \hat{\pi}_0^{(0)} - \hat{\pi}_0^{(0)}(1 - \hat{\pi}_0^{(0)})\hat{\mu}_1^2\} / (1 - \hat{\pi}_0^{(0)})$$

There is a trade-off of the choice of ξ. In most cases, as ξ grows smaller, the bias of grows larger, but the variance becomes smaller. When the empirical densities are considered, for the initial value of π_0, the z_j are sorted in descending order, then the first M_o smallest z_j's are assigned to the non-differential group and the remaining M-M_0 to the alternative group; the means and the variances are taken from the corresponding classes to be formed. Based on the data from Figure 6, the theoretical and the empirical densities provided different estimation of the π_0, as a consequence, the number of significant genes were different, but only slight changes were observed in the $F\hat{D}R$ (Table 4).

Table 4. FDR estimation, number of significant genes and the best estimation of π_0 under different conditions: Theoretical vs. Empirical densities and different threshold for declaring significant results ($\hat{\tau}_0(z_j) \le c_0$). Data from Figure 6 using EMMIX program.

		Theoretical Density			Empirical Density		
t-statistics	c_0	# of Sig. Genes	$F\hat{D}R$	$\hat{\pi}_0$	# of Sig. Genes	$F\hat{D}R$	$\hat{\pi}_0$
Ordinary	**0.1**	89	.04	0.532	6	.033	.971
	0.2	169	.097		8	.070	
	0.3	256	.16		9	.091	
Permutation	**0.1**	147	.049	.350	6	.080	.937
	0.2	296	.103		11	.102	
	0.3	452	.161		18	.174	
Limma	**0.1**	91	.04	.533	8	.043	.971
	0.2	170	.096		9	.067	
	0.3	250	.151		9	.067	

Storey, et al. have proposed the following estimation of the FDR using the SAM methods (32):

$$F\hat{D}R_{\Delta'}(\Delta) = \hat{\pi}_0(\Delta')\frac{R^0(\Delta)}{R(\Delta)} = \hat{\pi}_0(\Delta')\frac{\sum_{b=1}^{B}\#\{t_j^{*b} : t_j^{*b} \le t_1(\Delta) \text{ or } t_j^{*b} \ge t_2(\Delta)\}/B}{\#\{t_j : t_j \le t_1(\Delta) \text{ or } t_j \ge t_2(\Delta)\}}$$

where $\hat{\pi}_0(\Delta') = \frac{M - R(\Delta')}{M - R^0(\Delta')}$ which is the overall proportion of true null hypotheses (unchanged genes). The SAM methodology takes Δ' such that $R^0(\Delta') = M/2$ (i.e., half the null statistics fall in the rejection region defined by Δ'). In figure 7, the SAM method estimated that there were 173 significant genes out of 1541, and the estimate of FDR was 0.114.

IV) Sample Size

The sample size for microarray data is an area of continuous research. When a group comparison (i.e., diseased vs. healthy) is the objective of the study, several methods have been already proposed (40, 43-45). Most of these methods determine the sample size controlling the FDR with the assumptions of independent observations among genes. When controlling the FDR=V/(V+S), the following two complementary screening tests are affected: (i) the false negative rate (FNR=T/m_1) or the proportion of truly DE genes missed by the experiment, and (ii) the sensitivity (S/m_1), the proportion of truly DE genes identified by the experiment, also known as the average power. For example, given the following contingency table in a microarray setting for pairwise comparison:

Genes differentially expressed	Declared non-significant	Declared Significant	Total
No-True	U+ε	V-ε	mo
True	T-ε	S+ε	m1
Total	U+T	V+S	M

Assuming ε is an integer number, then FDR'=(V-ε)/(V+S), FNR'=(T-ε)/m1, and the sensitivity'=(S+ε)/m1=1-FNR'. As a consequence,

a) if ε >0, then the FDR' < FDR, FNR' < FNR and sensitivity'> sensitivity.

b) if ε <0, then the FDR'> FDR, FNR'> FNR and sensitivity' < sensitivity.

Therefore, to determine the most adequate sample size, when a microarray experiment is planned to compare two groups, a simultaneous assessment of the FDR and the sensitivity has to be performed. In order to carry on this assessment, Pawitan, et al. (46) have proposed the following expressions:

$$FDR = \frac{\pi_0\{1 - F_0(c)\}}{1 - F(c)} \qquad sensitivity = 2(1 - F_1(c)$$

where $F(c) = \pi_0 F_0(c) + (1 - \pi_0)F_1(c)$, F_o(c) is the central *t*-distribution with 2n-2 degrees of freedom, F_1(c) is the non-central *t*-distribution with 2n-2 degrees of freedom and non-centrality parameters $\pm\sqrt{n/2}\,D/\sigma$, D/σ is the assumed non-zero log-fold change, π_0 is the probability that a gene is unaffected, c is a given critical value to declare significant differences, and 2(1-*F*(c)) is the proportion of declared DE genes.

This sort of assessment can be performed using the R-package OC plus for computing FDR, sensitivity curves, and sample size (46). For example, the effect of the FDR and sensitivity for two sample sizes (n=10, 50) for different critical values of t-statistics are presented in Figures 8 and 9 when *t*-Statistics is the ordinary two-sample *t*-statistics with pooled variance and, the Log-fold change D=1, which is the mean difference in log2-scale and in standard deviation units ('log-fold change =1' indicates a ratio of 2σ for the mean of Group 1 versus the mean of Group 2), and π_0=0.9.

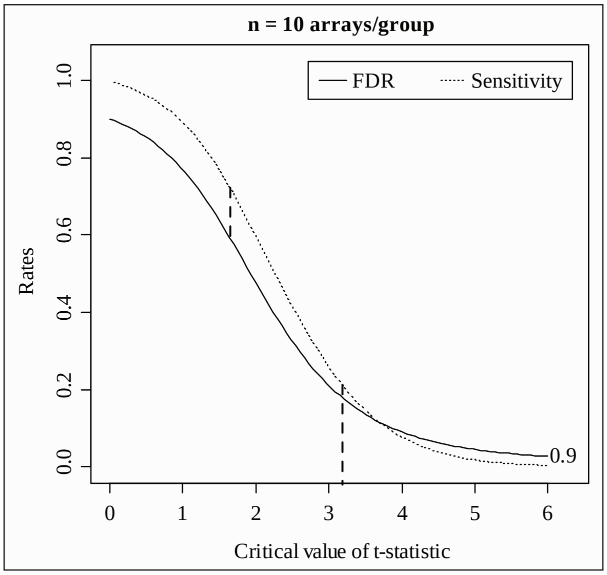


Figure 8. FDR and Sensitivity with n=10 arrays/group

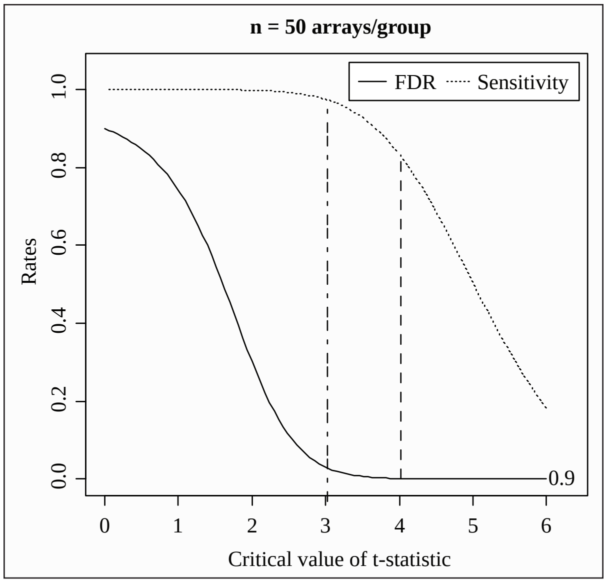


Figure 9. FDR and Sensitivity with n=50 arrays/group

When the sample size is 10 per group and the sensitivity is ~80%, the FDR is very high (>70%). If the critical value is > 3, the FDR is reduced (<25%), but the sensitivity also is reduced close to 30%. On the contrary, when the sample size is 50 per group and the sensitivity is ~80%, the FDR is very low (<5%). If the critical value is > 3, the same pattern is observed, high sensitivity and very low FDR. So, an adequate sample size per group will be close to 50.

V) Software

Several R packages for microarray data analysis are available for free at Bioconductor (www.bioconductor.org). One of these packages is R/maanova, which is extensible, interactive environment for microarray analysis. R/maanova can be used for data quality checks and visualization, data transformation, ANOVA model fitting (fixed and mixed effects model), Statistical tests including permutation, confidence interval with bootstrapping, and cluster analysis. Gentleman, et al. (12) is an excellent reference for starting programming in R for microarray data analysis.

DNA-Chip Analyzer (dChip Harvard University) and BRB-Array tools are also free microarray analysis software. DNA-Chip Analyzer is a Windows software package for probe-level (e.g. Affymetrix platform) and high-level analysis of gene expression microarrays and SNP microarrays. Gene expression or SNP data from various microarray platforms can also be analyzed by importing as external dataset. At the probe level, dChip can display and normalize data, and the model-based approach allows pooling information across multiple arrays and automatic probe selection to handle cross-hybridization and image contamination. High-level analysis in dChip can be performed, among them are: comparing samples, hierarchical clustering, view expression and SNP data along chromosome, and linkage analysis (www.dchip.org). BRB-ArrayTools have utilities for processing expression data from multiple experiments, visualization of data, multidimensional scaling, clustering of genes and samples, and classification and prediction of samples. BRB-ArrayTools features drill-down linkage to NCBI databases using clone, GenBank, or UniGene identifiers, and drill-down linkage to the NetAffx database using Probeset ids. It can be used to analyze both single-channel and dual-channel experiments. The package is implemented as an Excel add-in so that it has an interface that is familiar to biologists (http://linus.nci.nih.gov/~brb/download.html).

VI) Conclusions

The methodology for pairwise comparison is an area in development, there are still several issues under discussion for preprocessing data (different platforms to collect microarray data, different segmentation procedures, different approaches for normalization, use of the mismatched probes?), statistical inference (different t-statistics, theoretical vs. empirical densities, different methods to control the proportion of false positive declarations, or problems in controlling the correlation among the genes and among the tissues), sample size and power analysis with correlated genes (closed formula or sensitivity analysis), and validation (Is there a gold standard to measure gene expression?, What the criteria under which a finding can be said to be validated) (10).

The mixture-model methods seems to be the standard procedure in the assessment of differential expressions when the proportion of false positive declarations is controlled using either the theoretical or empirical densities. The Bayesian approach has been used for differential expression under the structure of a linear model, combining the classical and Bayes approach in which the posterior variance has been substituted in the ordinary *t*-statistics in place of the usual sample variance. Also, for the FDR estimation, a Bayesian approach has been used; however, different methods are still used to determine the proportion of null hypotheses (π_0). A Bayesian version of the *p*-values has been developed, which is called the *q*-value (posterior Bayesian p-value) to estimate the expected proportion of false positives incurred when calling a particular gene significant. Due to the expected correlation across the genes, the permutation methods have been the recommended procedure to estimate the *p*-values for pairwise comparison. One of the permutation methods is called SAM, which has been recommended when more genes are overexpressed than underexpressed.

Overall, there are still several areas of development in microarray data analysis. We hope that this introduction to microarray data analysis will atract more investigators from different fields to develop new approaches, particularly in the areas of quality control and validation. Microarray data analysis is a powerful instrument that could be used to identify the global expression responses of genes in specific environmental conditions in order to better understand the social disparity in the health-disease process.

Acknowledgments

We thank Laura Bretaña for her help in editing this document. In addition, the first author wants to thank the Department of Mathematics, University of Queensland (Brisbane, Australia) for their hospitality during his sabbatical year. This study was supported by the School of Public Health of the University of Puerto Rico and by the following awards: (i) RCMI clinical research infrastructure initiative (RCRII) award, 1P20 RR11126, from the National Center for Research Resources (NCRR), NIH (ii) U54CA096297, MDACC/PRCC Partnership for Excellence in Cancer Research award, NIH/NCI.

Resumen

Análisis de datos en microarreglos relacionados con ADN es una tecnología nueva que nos permite evaluar simultáneamente la expresión genética de miles de genes. Esta tecnología ha sido utilizada para analizar la expresión genética entre grupos de células de diferentes órganos o de diferentes poblaciones. Con el propósito de entender las funciones de los genes, es necesario contar con la información sobre las funciones de las proteínas y de la transcripción del mRNA. Desafortunadamente, explorar las funciones de las proteínas es muy difícil debido a su estructura compleja tridimensional. Para resolver esta dificultad, nos podemos concentrar en las moléculas de mRNA a través de la expresión genética. En este artículo describimos algunos de los métodos para el pre-procesamiento de datos en expresión genéticas y el análisis comparativo de dos grupos en un experimento genómico. Estudios previos, realizados para evaluar la eficiencia de diferentes métodos para comparar dos grupos, han resultado en una limitada concordancia en la listas de genes significativos. Finalmente, describimos los procedimientos para el control de la tasa de descubrimientos falsos, la determinación de tamaño de muestra en estudios comparativos para el análisis de datos en microarreglos y los programas de computación disponibles para este tipo de análisis. Este artículo está escrito para los profesionales de la salud interesados en el análisis comparativo de datos en microarreglos y que deseen tener una introducción de los diferentes pasos que se deben realizar para llevar a cabo este tipo de análisis.

References

1. Pasternak J. An Introduction to Human Molecular Genetics: Mechanisms of Inherited Diseases. USA, John Wiley and Sons, 2005.
2. Suárez E, Sariol CA, Burgette A, McLachlan G. A tutorial in genetic epidemiology and some consideration in statistical modeling. P R Health Sci J 2007;26:401-421.
3. Ducray F, Honnorat J, Lachuer J. 2007. DNA microarray technology: principles and applications to the study of neurological disorders. Rev Neurol 2007;163:409-420.
4. Cowell JK, Hawthorn L. The application of microarray technology to the analysis of the cancer genome. Curr Mol Med 2007;7: 103-120.
5. Palacios G, Quan PL, Jabado OJ, et al. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. Emerg Infect Dis 2007;13:73-81.
6. Sariol CA, Munoz-Jordan JL, Abel K, et al. Transcriptional activation of interferon stimulated genes but not of cytokine genes after primary infection of rhesus macaques with dengue virus type 1. Clin vaccine immunol 2007;11:756-766.
7. Hardimann G. Microarray platforms-comparisons and contrasts. Pharmacogenomics 2004;5:487-502.
8. Parmigiani G, Garrett E, Irizarry R, Zeger S. The Analysis of Gene Expression Data: Methods and Software. USA, Springer-Verlag New York Inc., 2003.
9. Cheng Li, Tseng G, Wing Hung. Model-based analysis of olinucleotide array and issues in cDNA microarray analysis. In Terry Speed (Editor), Statistical Analysis of Gene Expression Microarray Data. USA, Chapman & Halls/CRC, 2003.
10. Allison D, Cui X, Page G., Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nature Rev Genet 2006;7:55-65.
11. Allison D, Page G, Beasley T, Edwards J. DNA Microarrays and Related Genomics Techniques: Design, Analysis and Interpretation of Experiments. USA, Chapman & Hall/CRC, 2006.
12. Gentleman R, Carey V, Huber W, et al (Editors). Bioinformatics and Computational Biology Solutions Using R and Bioconductor. USA, Springer Science+Business Media, Inc, 2005.
13. McLachlan G, Do K, Ambroise Ch. Analyzing Microarray Gene Expresión Data. John Wiley & Sons, 2005.
14. Mont D. Bioinformatics: Sequence and Genome Analysis, Second Edition. USA, Cold Spring Harbor Laboratory Press, 2004.
15. Speed T. Statistical Analysis of Gene Expression Microarray Data. USA, Chapman & Hall/CRC, 2003.
16. Verducci J, Melfi V, Lin Sh, et al Microarray analysis of gene expression: considerations in data mining and statistical treatment. Physiol Genomics 2006;25:355-363.
17. Ross M, Mahfouz R, Onciu M, et al. Gene expression profiling of pediatric acute myelogenous leukemia. Blood 2004;104: 3679-3687.
18. Bolstad B, Irizarry R, Astrand M, Speed T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185-93.
19. Yang Y, Buckley M, Dudoit S, Speed T. Comparison of Methods for Image analysis on CDNA Microarray Data. J Comput Graph Stat 2002;11:108-36.
20. Baek J, Sook So Y, McLachlan G. Segmentation and intensity estimation of microarray images using a gamma-t mixture model. Bioinformatics 2007;23:458-465.
21. Irizarry R, Bolstad B, Collin F, Cope L, Hobbs B, Speed T. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003;31:e15.
22. Smyth G, Speed T. Normalization of cDNA Microarray Data. Methods 2003;31:265-73.
23. Yang Y, Paquet A. Preprocessing Two-Color Spotted Arrays. In Gentleman, et al (editors), Bioinformatics and Computational Biology Solutions Using R and Bioconductor. USA, Springer Science+Business Media, Inc, 2005.
24. Efron B, Tibshirani R, Storey J, Tusher V. Empirical Bayes Analysis of a Microarray Experiment. J Am Stat Assoc 2001;96:1151-1160.
25. Jeffery I, Higgins D, and Culhane A. Comparison and evaluation of methods for generating differentially expressed gene list from microarray data. BMC Bioinformatics 2000;7:359.
26. Lian I, Chang C, Liang Y, Fann C. Identifying differentially expressed genes in dye-swapped microarray experiments of small sample size. Comput Stat Data Anal 2007; 51:2602-2620.
27. McLachlan G, Bean R, Ben-Tovim Jones L. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics 2006;22:1608-1615.
28. Chipman H, Trevor H, Tibshirani R. Clustering microarray data. In Terry Speed (Editor), Statistical Analysis of Gene Expression Microarray Data. USA, Chapman & Halls/ CRC, 2003.
29. McLachlan G, Chang S. Mixture modelling for cluster analysis. Stat Methods Med Res 2004;13:347-361.
30. Bueno J, Gilmour S, Rosa G. Design of Microarray Experiments for Genetics Genomics Studies. Genetics 2006;174:945-957.
31. Chiaretti S, Li X, Gentleman R, et al. Blood 2004;103:2771-2778.

32. Storey J, Tibshirani R. SAM Thresholding and False Dicovery Rates for Detecting Differential Gen Expression in DNA MIcroarrays. In Parmigiani G, Garrett E, Irizarry R, Zeger S. (editors). The Analysis of Gene Expression Data: Methods and Software. USA, Springer, 2003.
33. Smyth GK. Linear Models and Empirical Bayes Mehtods for Assessing Differential Expression in Microarray Experiments. Stat Appl Genet Mol Biol 2004;3:Article 3.
34. Lo K, Gottardo R. Flexible empirical Bayes models for differential gene expression. Bioinformatics, 2007;23:328-335.
35. Storey J, Tibshirani R. Statistical Significance for Genomewide studies. Proc Natl Acad Sci U S A 2003;100:9440-9445.
36. Ge Y, Dudoit S, Speed T. Resampling-based multiple testing for microaray data analysis. Technical Report # 633. Department of Statistics, University of California, Berkeley, 2003.
37. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98: 5116-5121.
38. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful approach to Multiple Testing. J R Stat Soc Ser B 1995;57:289-300.
39. Dupuy A, Simon RM. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. J Natl Cancer Inst 2007;99:147-157.
40. Pounds S, Cheng Ch. Improving false discovery rate estimation. Bioinformatics 2004;20:1737-1745.
41. Storey J. The Positive False Discovery Rate: A Bayesian interpretation and the q-value. Ann Stat 2003;31:2013-2035.
42. McLachlan G, Peel D. Finite Mixture Models. USA, John Wiley & Sons, Inc, 2000.
43. Ferreira J, Zwinderman A. Approximation Power and Sample Size Calculations with the Benjamini-Hochberg Method. Int J Biost 2006;5:Article 8,1-35.
44. Gadbury G, Page G, Edwards J, et al. Power and sample size estimation in high dimensional biology. Stat Methods Med Res 2004; 13:325-338.
45. Lee M, Whitmore G. Power and sample size for DNA microarray studies. Stat Med 2002;21:3543-3570.
46. Pawitan Y, Michiels S, Koscielny S, et al. False discovery rate, sensitivity and sample size microarray studies. Bioinformatics 2005;21:3017-3324.