A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data

Sanghamitra Bandyopadhyay, Saurav Mallik, and Anirban Mukhopadhyay

Abstract—DNA microarray is a powerful technology that can simultaneously determine the levels of thousands of transcripts (generated, for example, from genes/miRNAs) across different experimental conditions or tissue samples. The motto of differential expression analysis is to identify the transcripts whose expressions change significantly across different types of samples or experimental conditions. A number of statistical testing methods are available for this purpose. In this paper, we provide a comprehensive survey on different parametric and non-parametric testing methodologies for identifying differential expression from microarray data sets. The performances of the different testing methods have been compared based on some real-life miRNA and mRNA expression data sets. For validating the resulting differentially expressed miRNAs, the outcomes of each test are checked with the information available for miRNA in the standard miRNA database PhenomiR 2.0. Subsequently, we have prepared different simulated data sets of different sample sizes (from 10 to 100 per group/population) and thereafter the power of each test have been calculated individually. The comparative simulated study might lead to formulate robust and comprehensive judgements about the performance of each test in the basis of assumption of data distribution. Finally, a list of advantages and limitations of the different statistical tests has been provided, along with indications of some areas where further studies are required.

Index Terms—Differentially expressed transcripts, parametric and nonparametric tests, multiple testing corrections, power of test

1 Introduction

DNA microarray is a useful technology for measuring the activity levels of thousands of biomolecules (such as genes/miRNAs) [1], [2] simultaneously over different experimental conditions or tissue samples. It is used to gain novel biological insights about a biological system through clustering [3], [4], biclustering [5], classification [6], [7], differential gene selection [8], [9], [10] single nucleotide polymorphism (SNP) detection [11], cancer subtypes selection [4], etc. The aim of differential expression (i.e., DE) analysis is to identify the transcripts whose expressions change significantly in terms of mean and/or standard deviation across different types of samples or experimental conditions. It may be noted that DE of transcripts is closely related to the choice of the statistical testing method adopted. Many genes/miRNAs are abnormally expressed (deregulated) under diseased conditions. Many genetic and epigenetic factors are responsible for such deregulations. Microarray experiments are carried out to recognize such aberrations in expressions [12] between two or more groups. Such differentially expressed genes/miRNAs are also called discriminator genes/miRNAs.

- *S. Bandyopadhyay and S. Mallik are with Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, West Bengal, India. E-mail: sanghami@isical.ac.in, sauravmtech2@gmail.com.*
- *A. Mukhopadhyay is with the Department of Computer Science and Engineering, University of Kalyani, Kalyani 741235, India. E-mail: anirban@klyuniv.ac.in.*

Manuscript received 2 Nov. 2012; revised 30 July 2013; accepted 6 Nov. 2013; date of publication 19 Nov. 2013; date of current version 7 May 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2013.147

It has been noticed that for small number of samples, the choice of any statistical test for identifying differentially expressed transcripts in the microarray data is too much contradictory. Inaccurate choice of any statistical test may lead to incorrect p-values. Thus, the estimation of differentially expressed genes/miRNAs may be incorrect and the resulting gene/miRNA list might be different for different test. The distribution of data for each group/population should be also taken into account, which might make impact on the performance of the test. There are many latest articles [13], [14], [15], [16], [17] which has provided the survey of different statistical tests for microarray data. Murie et al. [17] presented the comparison of different statistical tests for small number of samples in microarray data. But, this paper does not provide any validation of resulting genes/miRNAs. The number of tests covered in the survey paper is too small. Moreover, in the above mentioned survey papers, either the large number of testing methods has been reviewed neglecting the individual special phenomena of each test or a couple of tests have been covered up focusing on the individual special phenomena of each test. But, both of those are not covered in a single study. Therefore, in our paper, we have tried to elaborate both these aspects. Here, we have reported a comprehensive survey of different parametric and non-parametric testing [13], [14], [15], [16], [17], [18], [19], [20] methodologies used for finding differentially expressed genes/miRNAs [21] from microarray data sets. We have provided a comparative study of the performances of several testing methods based on three miRNA expression data sets of Lu et al. [22] and two mRNA expression data sets to identify differentially expressed miRNAs

and genes, respectively at 0.05 p-value cutoff. Thereafter, miR-Ontologies of the top five significant miRNAs for each test are described. For validating the resulting differentially expressed miRNAs, we used the standard miRNA database, PhenomiR 2.0. Some special phenomena of testing methods are also described briefly. Subsequently, we have prepared different simulated data sets of different sample sizes (from 10 to 100 per group/population) and thereafter the power of each test has been calculated individually. The comparative simulated study may lead to formulate robust and comprehensive judgements about the performance of each test on the basis of assumption of data distribution. The above methodologies provide a highly convenient and robust way to mark differentially expressed genes/miRNAs which are very useful for biological and clinical analysis. A table has been provided indicating the summary of the comparative study of different statistical tests on the basis of the experimentally calculated power for normally distributed data as well as non-normally distributed data. Finally, a list of advantages and limitations of the different statistical tests are provided, and some directions in which further investigation is needed are indicated.

The rest of the paper is organized as follows. Section 2 contains a description of different statistical testing methods for identifying differential expression and other related issues such as multiple testing corrections. Sections 2.5 and 2.6 describe the different types of errors in statistical tests and the performance metrics used for evaluating the performance of different tests, respectively. The data sets on which these statistical tests are applied, are mentioned in Section 4.3. The results are described in Section 4. A discussion on the results of the different tests is provided in Section 5. Finally, the conclusions are presented in Section 6.

2 Statistical Methodology for Identifying Differential Expression

In this section, we provide a review on different statistical methods for determining differentially expressed genes/miRNAs in microarray data.

2.1 Analysis of Pre-Test

Before applying any statistical test, it is mandatory to utilize some pre-filtering processes, normality test and normalization respectively to obtain result with reduced error.

2.1.1 Pre-Test Filtering Processes [23]

In high-dimensional data, multiple testing is necessary for adjustment of error during different statistical tests; but the multiple testing may produce the result with low statistical power. There are different filtering processes have been used to assess the differential expression before applying any statistical test. One fundamental procedure is to check overall variance of the data according to each gene (i.e., row) and filter out the genes having very low variance. This is an example of nonspecific or unsupervised filters. There are a Matlab function "genevarfilter' by which some user-defined percentile (say, 5 or 10 or 20 percentile) of the genes having low variance can be removed from gene list.

Due to the low variance of the gene, sometime lower p-value is produced which seems to be significant, but actually it is insignificant. If overall variance filtering procedure is applied before using t-test, then discoveries due to the small number of samples can be avoided. In Limma t-statistic [19], the overall variance filter is already merged with it to avoid discoveries with small effect sizes. Actually, filtering increases the number of discoveries. It is helpful if the false positive (FP) rate is controlled correctly. In fact, certain nonspecific filters may invalidate type I error control, as the filter statistic is not based on the sample class-labels. Inappropriate filtering affects adversely the type I error rate control. Suppose, the data matrix is of size $m \times n$, where m denotes the total number of genes and n denotes the total number of samples. Let, the data for a gene i are $Y_i = (Y_{i1}, \ldots, Y_{in})^t$. Therefore, if $Y_{i1}, \ldots, Y_{in}$ are fully independent and normally distributed for each $i \in H_0$, then both the overall variance and overall mean filter statistics are marginally independent of the 2-sample t-test for the gene (where, H_0 denotes set of indices for true nulls); i.e., the unconditional distributions before filtering and the conditional marginal distributions of test statistics after filtering are same. Thereby, the unadjusted p-values after both filtering and test statistic will be the correct. When the number of samples is large, the utilization of an empirical null distribution provides a potential solution that the effects of conditioning can be correctly incorporated.

In fact, the filter and test statistics are not necessary to be independent on each other when the null hypothesis is false (i.e., the corresponding gene is differentially expressed). Moreover, positive correlation between the two statistics under the alternative hypothesis is needed for increment of detection power by filtering.

Another type filtering is fold change filtering. This type of filtering is combined with the test statistic in volcanoplot. It is available online at: http://www.mathworks.in/help/bioinfo/ref/mavolcanoplot.html. Here, the genes which have fold changes greater than a certain threshold value (i.e., lower cutoff) or less than another threshold value (i.e., upper cutoff), are considered to be filter-passed genes. There is another software which is available for making the necessary pre-filtering computations in the "genefilter" package for Bioconductor which is available online at: http://www.bioconductor.org/packages/2.12/bioc/html/genefilter.html.

There is also a filter which is developed on the basis of the maximal within-class mean. Suppose, there are two groups/classes. For simplification of the strategy, it can be assumed that the data follow Gaussian distribution, and all genes have known common variance σ_2, by which $\overline{\sigma}_2 = 2\sigma_2/n$ becomes the variance for the within-class means. The filter statistic is that choose the genes whose $\max\{\overline{Y}_{i,1g}, \overline{Y}_{i,2g}\}$ exceeds the user-defined cutoff u^*.

2.1.2 Normalization

After pre-filtering process, it is necessary to prepare the data of each gene from different scales to a common scale. Thus, a normalization method (viz., zero-mean, min-max, median, sigmoid, statistical column normalization [24],

quantile normalization [25], variance stabilizing normalization or VSN [26] techniques) should be applied genewise on the data. (For details, see supplementary file "RSurvey-supplementary3.pdf", which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2013.147.)

2.1.3 Normality Test

After normalization, normality tests are necessary to utilize on data of each gene to determine whether it follows a normal distribution or not for each group/population. There are different methods are available for the normality tests (viz., Jarque-Bera test [27], the Anderson-Darling test [28], Lilliefors test [28], Shapiro-Wilk test [28], etc.).

After normality test, we can apply different parametric tests on the genes which follow normal distribution and different nonparametric tests on the genes which do not follow normal distribution. For very small number of samples (say, 1-3 samples), we may apply fold change method only as no statistical test can make sense on this data.

2.2 Fold Change

Fold change is a fundamental method for identifying differentially expressed genes/miRNAs. According to literatures, there are two definitions about fold change. The standard definition of the fold-change (FC) of gene/miRNA (g) is the ratio of the means between two groups [29] (i.e., $FC = \frac{\overline{x}_{1g}}{\overline{x}_{2g}}$, where $\overline{x}_{1g}$ is the mean of samples of group 1 and $\overline{x}_{2g}$ is the mean of samples of group 2). According to Choe et al. [30], FC is the difference of the means between two groups (i.e., $FC = (\overline{x}_{1g} - \overline{x}_{2g})$), specially when all the values are log2-transformed values. Thus, standard deviation of the groups is not considered here. It is not always a good estimation of differential expression of microarray data, especially when examining the expression of less abundant transcripts.

2.3 Parametric Tests

Parametric Tests are called as classical tests. In parametric tests, it is assumed that the data follow a normal distribution and they also show the same variance property of the groups. If these assumptions do not apply, non-parametric tests must be utilized. Parametric tests normally relate to data expressed [31] in absolute numbers or values rather than ranks. Different types of popular parametric tests, by which we are able to identify the significantly differentially expressed genes/miRNAs, are described below in brief.

2.3.1 T-Test

The t-test [16] is the most commonly used statistical test. The "two-sample t-test" compares the difference between two means in relation to the variation in the data. It allows us to measure a p-value using the t-test statistic. The p-value is calculated from t-table or cumulative distribution function (cdf). This p-value indicates the probability of observing a t-value as large or larger than the actually observed t-value where the null hypothesis is given true. By convention, if there is less than 5 percent probability of getting the observed differences by chance, we can say that we found a statistically significant difference between the two groups. Suppose, for each gene/miRNA g, group 1: n_1 treated samples, with mean $\overline{x}_{1g}$ and standard deviation s_{1g}; and group 2: n_1 controlled samples, with mean $\overline{x}_{2g}$ and standard deviation s_{2g}

$$t - statistic : t = \frac{(\overline{x}_{1g} - \overline{x}_{2g})}{se_g}. \quad (1)$$

Here, se_g denotes the standard error of the groups' mean, thus, $se_g = sPooled * \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$, where *sPooled* is the pooled estimate of the population standard deviation; i.e., $sPooled = \sqrt{\frac{(n_1-1)*s_{1g}^2+(n_2-1)*s_{2g}^2}{df}}$. Here, *df* is degree of freedom of the test. It is stated as $df = (n_1 + n_2 - 2)$. This strategy is used assuming that variance of two groups are equal. Assuming equal variability of two groups is known as "Behrens Fisher" problem [32]. It occurs when somebody wants to make inferences about the means of two normally distributed groups without assuming the equal variability. For recovery from this, it is necessary to check whether the variance of two groups are equal to each other or not. If equal, then use earlier mentioned t-statistic, otherwise use the following t-statistic: $t = ((\overline{x}_{1g} - \overline{x}_{2g})/\sqrt{\frac{s_{1g}^2}{n_1}+\frac{s_{2g}^2}{n_2}})$. Here we use unpooled estimates of the population standard deviations. It's called as Welch's t-test [32].

Now, there is a special case where normality assumption in t-test may become incorrect. Let, all genes have known common variance σ^2, so that $\overline{\sigma}^2 = 2\sigma^2/n$ becomes the variance for the within-class averages. The filter statistic is to compare U_g^I = max{$\overline{x}_{1g}$, $\overline{x}_{2g}$} to the user-defined threshold u^* and to choose those genes whose U_g^I exceeds the threshold. The test statistic is: $U_g^{II} = (\overline{x}_{1g} - \overline{x}_{2g})/\sqrt{2\overline{\sigma}}$, which is analogous to standard t, but applies the known variance in the denominator.

According to Fig. 2a, in the unconditional distribution (i.e., the dotted curve) of the test statistic, within-class mean μ is much above u^*; and in the conditional distribution of the statistic, μ is either near to u^* (see the dashed curve) or far below u^* (see the solid curve). Initially, for the dotted curve, the distribution of the statistic for the non-differentially expressed genes is a standard normal or Gaussian. But, whenever non-normality and size of tails increase gradually, then at first μ reaches at the level of u^* and then gradually goes down far below u^*. When μ exceeds highly the threshold u^*, then the conditional null distribution has much **heavier tails**. It is actually the most extreme stage. Here, the genes produce the non-normal conditional distributions. At that moment, such genes can rarely pass through the filter-stage due to producing low μ in both classes. The genes which have μ closer to u^*, pass with appreciable frequency and can be able to produce significantly non-normal conditional distributions. Therefore, the normality assumption for the last two cases is totally inappropriate for calculating t-statistic. The quantity σ controls the range of values of μ that has a reasonable chance of passing the filter and producing problematic conditional distributions for U_g^{II}. Thus, the range of values of μ and the relevance of the simulation example for controlling the actual error rate contracts as either the number of samples (i.e., n) increases or variance (i.e., σ)

decreases. In Fig. 2b, a simulated example has been presented where there is a skewed (unsymmetric) data distribution which recommends to use any non-parametric test, not any parametric test like t-test.

2.3.2 ANOVA 1 Test

ANOVA 1 (Analysis of Variance 1) [16] is such a statistical test in which the means of several groups/populations might or might not be all equal, and thereby infers t-test to more than two populations. Doing multiple two-sample t-tests will produce a result that has a high chance of making type I error [33], [34]. Therefore, ANOVA 1 is useful to compare the means of two, or more groups. It makes the comparison among the group means by estimating comparisons of variance estimates. The purpose of analysis of variance is to determine the differences in means of groups for making statistical significance. This is carried out by analyzing the variance, i.e., the variance is partitioned into the component which is caused by random error (i.e., Sum of Square within groups) and the components that are caused by the differences between means of groups. The latter variance components are then tested for statistical significance. If the significance is true, we reject the null hypothesis of no differences between means and accept the alternative hypothesis that the means (in the population) are different from each other. ANOVA 1 is based on the fact that two independent estimates of the population variance can be obtained from the sample data. A ratio has been formulated for the two estimates that is denoted by F. Suppose, for group 1: n1 treated samples, sum of group 1 values $= T_1$, mean $\overline{X_1}$ and for group 2: n2 controlled samples, sum of group 2 values $= T_2$, mean $\overline{X_2}$. Now, total number of groups$= P$ (here, $P = 2$), total number of values in two groups = $N = \sum_{j=1}^{P} n_j$ and $T = \sum_{j=1}^{P} T_j$. Let, $I = \frac{T^2}{N} II = \sum_{j=1}^{P}(\sum_{i=1}^{n_j} x_{ij}^2)$, $III = \sum_{j=1}^{P}(\frac{T_j^2}{n_j})$. Thus, sum of Squares between groups $(SS_{between}) = III - I$ and sum of Squares within groups $(SS_{within}) = II - III$ and moreover $SS_{total} = II - I$. Now, degree of freedom between groups $(Df_{between}) = (P - 1)$ and degree of freedom within groups $(Df_{within}) = \sum_{j=1}^{P}(n_j - 1)$. Here, $Df_{total} = \sum_{j=1}^{P} n_j - 1$, $Variance_{between} = MeanSquare_{between}(MS_{between}) = \frac{SS_{between}}{Df_{between}}$, $Variance_{within} = MeanSquare_{within}(MS_{within}) = \frac{SS_{within}}{Df_{within}}$,

$$F = \frac{MS_{between}}{MS_{within}}. \quad (2)$$

2.3.3 Pearson's Correlation Test ($corr$)

Correlation is such a degree to which two variables are covaried. It is either positive or negative. It is assumed that the data are from a bivariate normal population. Correlation coefficients are applied in statistics to measure how strong a relationship is made between two variables. Pearson's correlation is a correlation coefficient commonly applied in linear regression. Pearson's correlation [35] coefficient is an estimation of the intensity of the linear association between variables. It may be possible to have non-linear associations. If any association exhibits linearity, it is needed to examine the data closely to determine. Suppose, there are two groups, where group 1 includes n samples, with expression value x, with mean $\overline{x}$ and standard deviation s_x, and group 2 has n samples, with expression value y, with mean $\overline{y}$ and standard deviation s_y. Pearson's correlation coefficient (commonly denoted as ρ) between two variables is described as the covariance of the two variables divided by the product of their standard deviations (i.e., $\rho = \frac{cov(x,y)}{s_x s_y}$, where $cov(x, y) = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$). This test can predict whether two variables are connected. It is not capable to indicate that the variables are not connected. If one variable depends on another, i.e., there is a causal relation, then it is always possible to get some kind of correlation between the two variables. However, if both the variables are depending upon a third variable, then they will produce a sizable correlation with no causal dependency between these. We can compute a t-statistic for the correlation coefficient (ρ) using $t = \frac{\rho}{s_\rho}$, where $s_\rho = \sqrt{\frac{1-(\rho)^2}{n-2}}$ and $df = n - 2$ as there is only one df for each group.

2.4 Nonparametric Tests

Nonparametric tests do not assume any particular distribution. Some of the popular nonparametric tests which have a major role for identifying differentially expressed genes/miRNAs, are described below.

2.4.1 Permuted T-Test ($perm$)

A permutation t-test (also called an exact test, a randomization test or re-randomization test) is a kind of t-test in which a permutation or an rearrangement is done in the labels on the observed data points of each individual gene/miRNA. Normally, 100 permutations are applied here by default. If number of permutations exceed above 1,000, it needs to be set to 1,000. The next steps of this test are similar to t-test. Permutation t-test is a subset of nonparametric statistics. The inversion of this test to recognize confidence regions/intervals needs even more computation. Permuted t-test [36] is very useful when distribution of data is unknown. The permutation t-test has the same pitfall as the Student's t-test (i.e., the "Behrens Fisher" problem).

2.4.2 Wilcoxon Ranksum Test (RST) [37]

Whenever the groups are not normally distributed, specially for small sample sizes, then the result of t-test might not be valid. RST is an alternative test which can be utilized for such case. Since it operates on rank-transformed data, it appears to be a robust choice for microarray data, which are often non-normal and may have outliers. To perform the rank sum test, the combined samples are first ranked. Thereafter, the summation of the ranks is computed for group 1 (i.e., $T_1 = \sum ranks_{group_1}$), and the summation of the ranks for group 2 (i.e., T_2). If the number of samples for two groups (viz., n_1 and n_2, respectively, where $n_1 <= n_2$) are equal, the RST statistic T is the minimum of T_1 and T_2 (i.e., $T = min(T_1, T_2)$). If these are unequal, then we have to find T_1 equal to the sum of the ranks for the smaller sample. Thereafter, we compute $T_2 = n_1(n_1 + n_2 + 1) - T_1$. T is the minimum of

T_1 and T_2. Sufficiently small values of T cause rejection of the null hypothesis that the sample means are equal. Significance levels are tabulated for small values of n_1 and n_2. Here, p-value of each gene/miRNA is calculated from $min(2 * min(T_1, T_2), 1)$ in case of two-sided test. For large n_1 and n_2, the z-statistic is given below:

$$z = \frac{(|T - mean_{w_1}| - 0.5)}{\sqrt{var_{w_1}}}, \tag{3}$$

where $var_{w_1} = n_2 * mean_{w_1}/6 = n_1 * n_2 * (n_1 + n_2 + 1)/12$ and $mean_{w_1} = n_1 * (n_1 + n_2 + 1)/2$. RST is equivalent to the Mann-Whitney test. RST is very slower than t-test.

2.4.3 Modified Wilcoxon Ranksum test ($ModRST$ or $MRST$)

First of all, modified RST prepares a list of ranks of the gene/miRNA expression values for each gene/miRNA across all experiments in ascending, and then tests for equality of means of the two ranked samples. For data sets where both n_1 and n_2 exceed 8, normal approximation of the p-values can be used (Walpole and Myers, 1993 [38]). According to Troyanskaya et al. [39]: $z = \frac{(u_1 - mean_{u_1})}{\sqrt{var_{u_1}}}$. where $var_{u_1} = n_1 * n_2 * (n_1 + n_2 + 1)/12$ and $mean_{u_1} = n_1 * n_2/2$. Here, $u_1 = T_1 - n_1 * (n_1 + 1)/2$ and $T_1 = \sum ranks_{group_1}$.

2.4.4 Significance Analysis of Microarray (SAM)[15]

Avoiding the pitfall of small variance of t-stat is a great challenge. For this, SAM uses a statistic which is somewhat similar to t-stat. At low expression levels, the absolute value of t_g may be very high because of small values in se_g. The limitation of the t-test is that due to the low expression levels, genes/miRNAs with small sample variances have a big chance of being predicted as differentially expressed genes/miRNAs. Thus, SAM chooses to add a small positive constant s_0 (stated as "fudge factor") to solve this limitation. The SAM statistic by Tusher et al. [29] is

$$t_{sam} = \frac{(\overline{x}_{1g} - \overline{x}_{2g})}{se_g + s_0}, \tag{4}$$

where se_g is the standard error of the groups' mean, i.e., $se_g = sPooled * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Here, $sPooled$ is the pooled estimate of the population standard deviation (i.e., $sPooled = (\sqrt{\frac{(n_1-1)*s_{1g}^2 + (n_2-1)*s_{2g}^2}{df}})$. Here, df is degree of freedom of the test. It is stated as $df = (n_1 + n_2 - 2)$. It is assumed that the variance of the two groups are equal. The principle used when setting the value of s_0 is that the variability of t_{sam} should be independent of the level of se_g. This is achieved by estimating the variability of t_{sam} as a function of se_g in windows across the data. The median absolute distance (MAD) is used to estimate the variability and 100 windows with equal number of data points are used by default. With MAD_k equal to the computed variability of t_{sam} (i.e., υ_k^α) within the kth window, s_0 is chosen so that the coefficient of variation of $MAD_1, \ldots, MAD_{100}$ is as small as possible. The approach of determining s_0 is demonstrated in Algorithm 1 (Chu, Narasimhan, Tibshirani, and Tusher -2002).

Algorithm 1 Sub-approach of Computing s_0

1: Calculate the 100 percentiles q_k, where $k = 1, ..., 100$ of the se_g values.
2: **for** $\alpha = 0$ to 1 **do**
3: Determine $t_{sam}^\alpha \models \frac{(\bar{x}_{1g} - \bar{x}_{2g})}{se_g + s^\alpha}$, where s^α refers to the α- quantile of the se_g values, and also $s^0 = q_0 = \min_{\{i=1,\ldots,m\}}\{se_g\}$.
4: Compute $\upsilon_k^\alpha = 1.4826 \cdot MAD\{t_{sam}^\alpha | se_g \epsilon [q_{k-1}, q_k)\}$, where $k = 1, ..., 100$.
5: Calculate the coefficient of variation $CV(\alpha)$ of the υ_k^α values.
6: $\alpha = \alpha + 0.05$.
7: **end for**
8: Put $\hat{\alpha} = argmin_{\alpha \epsilon \Re}$, and $s_0 = s^{\hat{\alpha}}$.

2.4.5 Linear Models for Microarray Data (Limma)

In microarray data analysis, sometimes there are few number of arrays/samples, resulting in highly variable estimates of the standard deviation for each gene. In this situation, if there are large number of genes in the data, then the problem of dimensionality arises. To handle such problem, an empirical Bayes approach is considered in Limma test. It is implemented in the R-package Limma (Smyth) [19]. It depends on the methods proposed by Lönnstedt and Speed [40]. The empirical approach assumes a priori knowledge on the unknown gene/miRNA-specific variances to be a inverse-gamma distribution (Γ^{-1}-distribution) (i.e., before data of each gene/miRNA is observed).

Suppose, there are m number of genes and n number of samples. The linear model is in the following: $E(y_g) = X\alpha_g$, where y_g denotes the expression vector of each gene g (here, $g = 1, \ldots, m$) across all arrays/samples, X is a design matrix of full column rank and α_g is the coefficient vector. Certain contrasts of the coefficients which are assumed to be of biological interest for a particular gene g are stated as $\beta_g = C^T\alpha_g$. Although the methodology can analyse any number of contrasts, only two samples have been compared so that β_g may be stated as the log fold change $(\hat{lx}_{1g} - \hat{lx}_{2g})$, where $\hat{lx}_{1g}$ and $\hat{lx}_{2g}$ denote the means of the two groups' logged expression values. Now, the contrast estimators (i.e., $\hat{\beta}_g$) are assumed to be normally distributed; the residual sample variances (i.e., s_g^2) are assumed to follow scaled chi-squared distribution (i.e., χ_{df}^2) and gene-specific variance (i.e., σ_g^2) to follow Γ^{-1}-distribution. Therefore, this statistic is as: $\hat{\beta}_g | \sigma_g^2 \sim N(\beta_g, \sigma_g^2)$ (where N is stated as the normal distribution), $s_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2$, and

$$\sigma_g^2 \sim \Gamma^{-1}\left(\frac{d_0}{2}, \frac{d_0 s_0^2}{2}\right). \tag{5}$$

Under this hierarchical model, the posterior sample variance (i.e., $\tilde{s}_g^2$) for the gene g becomes: $\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$, where d_0 ($< \infty$) and s_0^2 refer to the prior degrees of freedom and variance respectively, and d_g (> 0) and s_g^2 denote the experimental degrees of freedom and the sample variance of a particular gene g, respectively. Since we have used only two sample comparisons, so d_g is always equal to $n - 2$. The two prior

parameters, d_0 and s_0^2 are stated as the degrees of freedom and variance of the prior distribution, respectively. The prior parameters d_0 and s_0^2 are determined by fitting the logged sample variances to a scaled F distribution. These two parameters are estimated by equating empirical to expected values in case of the first two moments of $log(s_g^2)$. Since the moments of $log(s_g^2)$ are finite in case of any degrees of freedom, and distribution of $log(s_g^2)$ is more nearly gaussian than s_g^2, thus $log(s_g^2)$ is utilized instead of s_g^2. The sub-procedure of estimating these two parameters are described in the Algorithm 2.

Algorithm 2 Sub-approach of Computing d_0 and s_0^2

1: Suppose, $z_g = log(s_g^2)$. Since each $log(s_g^2)$ follows a scaled F-distribution, thereby z_g is distributed as a constant plus Fisher's z-distribution (according to Johnson and Kotz, 1970, page 78).
2: The z_g is roughly normally distributed and it has finite moments of all orders which consist of $E(z_g) = log(s_0^2) + \psi(d_g/2) - \psi(d_0/2) + log(d_0/d_g)$ and $var(z_g) = \psi'(d_g/2) + \psi'(d_0/2)$, where $\psi()$ and $\psi'()$ are the digamma and trigamma functions respectively.
3: Let, $e_g = z_g - \psi(d_g) + log(d_g/2)$. Therefore, $E(e_g) = log(s_0^2) - \psi(d_0/2) + log(d_0/2)$ and $E\{(e_g - \bar{e})^2 G/(G-1) - \psi'(d_g/2)\} \approx \psi'(d_0/2)$.
4: Solve $\psi'(d_0/2) = mean\{(e_g - \bar{e})^2 G/(G-1) - \psi'(d_g/2)\}$ and then estimate d_0.
5: If there is an estimate for d_0 to be given, then s_0^2 is calculated by $s_0^2 = exp\{\bar{e} + \psi(d_0/2) - log(d_0/2)\}$. This estimated s_0^2 is basically less than the mean of the s_g^2 in identification of the skewness of the F-distribution. The gene for which $d_g = 0$, can lead to $\tilde{s}_g^2 = s_0^2$.

The moderated t-statistic in Limma can be defined as

$$\tilde{t}_g = \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \frac{\hat{\beta}_g}{\tilde{s}_g}, \tag{6}$$

where $n = n_1 + n_2$. The associated probability of the two-sample moderated t-statistic under the null hypothesis is determined by reference to the t-distribution with $d_0 + d_g$ degrees of freedom.

The conditional distribution of the test statistic might be affected by the nonspecific filters. It is already known that the Limma t-statistic is depended on an empirical Bayes approach which models the gene-specific error variances (i.e., $\sigma_1^2, \ldots, \sigma_m^2$) with a Γ^{-1}-distribution. In Limma, overall variance filtering is integrated in Limma test-statistic. Since the within-class variance estimator (s_g^2) and the overall variance are correlated, then the filtering on overall variance will deplete the set of genes with low s_g^2.

The moderated t-statistic and overall variance filtering in Limma select the genes which have more variance in the data. Here, $\tilde{t}_g$ shrinks the within-class variance estimates (s_g^2) towards a common $\hat{s}_0^2$, that has the effect of decreasing the denominator of the t-statistic for the genes having large variance. Thus, it has been directed to bias the t-statistics away from zero. In case of the genes having small variance, the denominator of the t-statistic is increased, and the corresponding t-statistic is trying to bias towards zero. Overall variance filtering completely eliminates the genes having low variance from the consideration.

Here, the Limma model has used two hyperparameters: d_0 and s_0^2 which are formulated from the set of standard gene-level variance estimates s_g^2. The underdispersion of the s_g^2 can lead to an estimate of $\hat{d}_0$ which becomes ∞. The underdispersion is constructed by the filtering-attached misfit between the conditional distribution of the s_g^2 and the Limma model. There are two consequences for estimating $\hat{d}_0 = \infty$. First, it makes the gene-level error variance estimates to be totally ignored, and the moderated t-statistics for each gene will have the same denominator, which creates an analysis based on fold change rather than a t-statistic that is not expected. Second, when the true values of both d_g and d_0 are low, but $\hat{d}_0 = \infty$, then the correct null distribution is to be heavy-tailed, but the moderated $\tilde{t}_g$ are inappropriately compared to a standard gaussian/normal distribution (see Fig. 3a). As a result, the p-values are calculated using an inappropriate null distribution, determining too many true-null p-values that are closed to zero (i.e., the loss of type I error rate control). Under this circumstance, the violation of the Γ^{-1}-distribution has been illustrated.

2.4.6 Shrink-t

The moderated t-statistic developed by Rhein and Strimmer [42] is based on the James-Stein ensemble shrinkage estimation rule. After the use in the genes/miRNAs-specific variance estimators $s_1^2, \ldots, s_G^2$, the results of the rule in adjusted estimators are stated as $\tilde{s}_g^2 = \hat{\lambda} s_0^2 + (1 - \hat{\lambda}) s_g^2$, where $\hat{\lambda}$ is called as the estimated pooling parameter, i.e., $\hat{\lambda} = min(1, \frac{\sum_{g=1}^{G} \widehat{var(s_g^2)}}{\sum_{g=1}^{G} (s_g^2 - s_0^2)^2})$. The estimator s_0^2 is the median of $s_G^2, ..., s_G^2$, and $\widehat{var(s_g^2)}$ is measured by $\frac{(n_1+n_2)^3}{(n_1+n_2-1)^3} \sum_{j=1}^{2} \sum_{i=1}^{n_j} (\frac{(x_{ijg} - \overline{x}_{jg})^2}{n_1 n_2} - \frac{n_1+n_2-2}{(n_1+n_2)^2} s_g^2)^2$. The shrink-t statistic is then defined as: $\tilde{t}_g = \frac{D_g}{\sqrt{\tilde{s}_g^2}}$, where D_g is the difference of means between two groups.

2.4.7 Softthreshold-t (Soft-T)

This testing is stated as L_1 penalized t-statistics which was proposed by Wu [7]. Suppose, group 1: n_1 treated samples, with mean $\overline{x}_{1g}$, standard deviation s_{1g} and corresponding degree of freedom (df1) = $n_1 - 1$; and group 2: n_1 controlled samples, with mean $\overline{x}_{2g}$, standard deviation s_{2g} and corresponding degree of freedom (df2) = $n_2 - 1$. Thus, pooled standard deviation for each gene/miRNA is defined as: $s12 = \sqrt{\frac{(ss1+ss2)}{(df1+df2)} * \left(\frac{1}{(df1+1)} + \frac{1}{(df2+1)}\right)}$, where $ss1 = rowSums((x_{1g} - \overline{x}_{1g})^2)$, and $ss2 = rowSums((x_{2g} - \overline{x}_{2g})^2)$. Now, Δ, the shrinkage parameter is described as: $\Delta = \overline{x}_{1g} - \overline{x}_{2g}$. Therefore, the numerator of this t-stat is defined as

$$numetator = \begin{cases} \Delta - sgn(\Delta) * \lambda, & \text{if } abs(\Delta) > \lambda, \\ 0, & \text{if } abs(\Delta) <= \lambda, \end{cases}$$

where, $sgn(.)$ is the sign function: $sgn(z) = 1$ if $z > 0$, $sgn(z) = -1$ if $z < 0$ and $sgn(z) = 0$ when $z = 0$. Also we

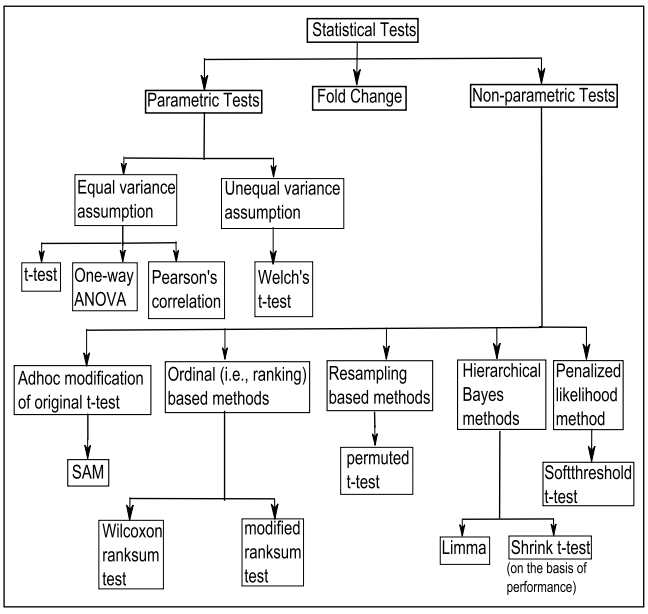


Fig. 1. The classification tree of different statistical tests.

take $\lambda = 2n_1 m_g sd12\lambda$ which denotes the penalty factor, where $m_g = \sqrt{\frac{1}{n_1} - \frac{1}{(n_1+n_2)}}$. The t-stat is then defined as

$$t = \frac{numerator}{\sqrt{\left(s12 + \frac{\lambda^2}{(df1+df2)}\right)}}. \quad (7)$$

2.4.8 Other Nonparametric Tests

The Kruskal-Wallis test (KW test) [43] was proposed by Kruskal and Wallis. It is a non-parametric test which is applied to decide whether samples follow from the same distribution or not. It is used for comparing more than two samples that are independent. The parametric equivalence of the Kruskal-Wallis test is ANOVA 1. Another popular test is Ideal discriminator (i.e., ID) method [39], which is a resampling-based approach. ID is stated as a theoretical gene/miRNA which is maximally expressed in all group 1 samples and minimally expressed in all group 2 samples. The strategy selects the genes/miRNAs which have the highest Pearson's correlation to the ID and justifies the significance by comparing Pearson correlation score of each gene/miRNA for true data to the best Pearson's correlation score from 50,000 random column permutation of data. The Kolmogorov-Smirnov test (KS test) [44] is another useful nonparametric test applied for the equality of continuous, one-dimensional probability distributions which can be applied to compare samples of two groups. The KS statistic determines a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. It is too much sensitive to differences in both shape and location of the empirical cumulative distribution functions of the two samples. The asymptotic p-value becomes very much accurate for large sample sizes, such that $(n_1 * n_2)/\ (n_1 + n_2) >= 4$, where the sample sizes of two groups are n_1 and n_2, respectively.

The classification tree of the tests is presented in Fig. 1.

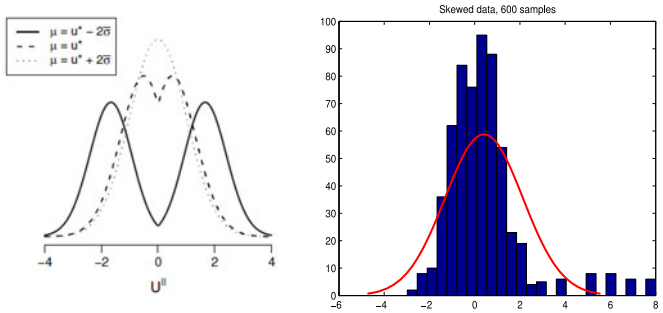


Fig. 2. (a) The conditional null distribution of the test statistic U_g^{II} for non-differentially expressed genes, where μ is greater than a threshold u^* (denoted by dotted curve), is identical to unconditional null distribution. As μ is coming close to the threshold u^* (denoted by dashed curve) and then passes below u^* (denoted by solid curve), the non-normality and excess-enlargement in the tails have been formed; (b) Simulation example of skewed (unsymmetric) distribution: data distribution having 600 samples (550 in one group and 50 in another group), mean of data = 0.4088, standard deviation = 1.7168, Kolmogorov-Smirnov test p-value = 3.0030e-004 which is less than 0.05, so, normality assumption incorrect; thus recommend to use non-parametric test.

TABLE 1
Type I Error and Type II Error

			Condition (Originally)		
			Condition Positive	Condition Negative	
			H_0 is False	H_0 is True	
Predicted	Positive	#Reject H_0	True Positive (TP)	False Positive (FP) (Type I error)	Positive predictive value $= \frac{TP}{TP+FP}$
	Negative	#Non-reject H_0	False Negative (FN) (Type II error)	True Negative (TN)	Negative predictive value $= \frac{TN}{FN+TN}$
			Sensitivity $= \frac{TP}{TP+FN}$	Specificity $= \frac{TN}{FP+TN}$	Accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$

2.5 Testing Errors and Performance Metrics

In this section, we have described different types of testing errors followed by the performance metrics used for evaluating different test methodologies.

In statistics, two popular error types are type I and type II errors [45], [46] which describe particular flaws in a testing method. Suppose, there are m number of genes (or null hypotheses) and n number of samples (or conditions) and number of rejected null hypotheses is R. According to Table 1, a type I error occurs when one rejects the null hypothesis H_0 when it is originally true. The probability of this error is the level of significance of the test of hypothesis, and is denoted by $alpha$. There are four categories of Type I error rates (Shaffer [47]): (i) per-comparison error rate (PCER) as the expected value of the number of Type I errors divided by the number of hypotheses, (i.e., $PCER = E(FP)/m$); (ii) per-family error rate (PFER) as the expected number of Type I errors, (i.e., $PFER = E(FP)$); (iii) family-wise error rate (FWER) as the probability of at least one Type I error (i.e., $FWER = Pr(FP \geq 1)$); and (iv) false discovery rate (FDR) of Benjamini and Hochberg as the expected proportion of Type I errors among the rejected hypotheses

(i.e., $FDR = E(FP/(TP+FP))$, if $R > 0$). Basically, in case of the multiple testing process, $PCER \leq FWER \leq PFER$. Therefore, $PFER$ procedure is more conservative than $FWER$, and also $FWER$ is more conservative than $PCER$; i.e., $PFER$ generates larger number of false positives than $FWER$ and similarly $FWER$ generates larger number of false positives than $PCER$.

A type II error occurs when one rejects the alternative hypothesis H_1 when the hypothesis is true. The probability of a type II error is denoted by $beta$. The power of a test is $(1 - beta)$; i.e., the probability of choosing the alternative hypothesis H_1 when the alternative hypothesis is correct.

There are two important statistical measures, viz., sensitivity and specificity. These are called as classification functions. Sensitivity measures the proportion of actual positives which are correctly identified. Specificity measures the proportion of negatives which are correctly identified. These two measures are closely related to the ideas of type I and type II errors. In Table 1, type I error, type II error, sensitivity, specificity and other related terms are defined.

The Matthews Correlation Coefficient (MCC) [48] is a quality-based measure for two-class classifications. It takes into account true and false positives and negatives. This MCC is stated as a balanced measure which can be used even if the classes are of very different sizes. This MCC is also very useful in case of imbalanced data set. It returns a value between -1 and $+1$. A coefficient of $+1$ signifies a prediction which is perfect, 0 implies no better than random prediction and -1 indicates total disagreement between prediction and observation. This statistic is also known as the phi-coefficient. The MCC can be formulated as

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{8}$$

2.6 Post-Test Processes: Different Corrections on p-Values in Different Statistical Tests

Multiple testing correction implies the repetition of calculation of the probabilities obtained from a statistical test which is repeated multiple times. For retaining a family-wise error rate α (i.e., FWER) in an analysis involving more than one comparison, the error rate for each comparison must be more stringent than α. These multiple testing corrections adjust p-values which are obtained from different statistical tests to make correction for the occurrence of FPs. At the time of testing, each gene/miRNA is considered independently from one another. The false positives is proportional to the number of tests performed and the p-value cutoff. According to GeneSpring, there are mainly four categories of multiple testing corrections:

2.6.1 Bonferroni Correction [49]

Bonferroni procedure is a useful in multiple testing corrections for strong control of the $FWER$ at the level α. This procedure rejects the null hypothesis H_g whose unadjusted p-value is less than or equal to α/m, where m denotes total number of miRNAs/genes or null hypotheses. The single-step Bonferroni adjusted p-value (i.e., $\tilde{p}_g$ for each gene g) is obtained by: $\tilde{p}_g = min(m\tilde{p}_g, 1)$. Now, the strong control of the $FWER$ follows Booles inequality. Suppose, m_0 denotes the number of true null hypotheses and P_g denotes random variable for unadjusted p-value of the gth hypothesis/miRNA/gene. Therefore, according to Booles inequality,

$$FWER = Pr(FP \geq 1) = Pr\left(\bigcup_{g=1}^{m_0}\{\tilde{P}_g \leq \alpha\}\right)$$
$$\leq \sum_{g=1}^{m_0} Pr(\tilde{P}_g \leq \alpha)$$
$$\leq \sum_{g=1}^{m_0} Pr\left(P_g \leq \frac{\alpha}{m}\right) \leq \frac{m_0\alpha}{m},$$

where the last inequality follows from $Pr(P_g \leq x | H_g) \leq x$, where $x \in [0, 1]$.

2.6.2 Bonferroni Step-Down (Holm) Correction [50]

This correction is very similar to the Bonferroni, but a little less stringent. Improvement in power may be achieved by step-down procedures. Here, the genes/miRNAs are ranked in ascending order of p-values. Let, $p_{r_1} \leq p_{r_2} \leq \cdots \leq p_{r_m}$ refer to the ranked observed unadjusted p-values and let $H_{r_1}, H_{r_2}, \ldots, H_{r_m}$ denote the corresponding null hypotheses. According to Holm (in 1979), index $g^* = min\{\, g : p_{r_g} > \alpha/(m-g+1)\}$ and reject the hypotheses H_{r_g}, where $g = 1, \ldots, g^* - 1$. Now, if there exists no such g^*, then reject all hypotheses. The Holm adjusted p-value is obtained by

$$\overline{p}_{r_g} = \max_{t=1,\ldots,g}\{min\big((m-t+1)p_{r_t}, 1\big)\}. \tag{9}$$

Since it is less corrective as the p-value increases, this correction is less conservative than Bonferroni correction.

2.6.3 Westfall and Young Permutation [51]

In both Bonferroni and Holm methods, each p-value is corrected independently. In many times, the unadjusted p-values may be dependent on each other. In this case, for DNA microarray expression study, some groups of genes have a tendency to be highly correlated. Westfall and Young [51] proposed a method as less conservative multiple testing procedure than the other two for determining the adjusted p-values, where the dependency among test statistics is considered. Here, p-values are determined for each gene/miRNA based on the data set. The p-values are then ranked. The permutation method produces a pseudo-data set by dividing the data into artificial control and treatment groups. The p-values for all genes/miRNAs are calculated on the pseudo-data set. The successive minima of the new p-values are taken and then compared with the original ones. This technique is repeated a large number of times. The proportion of re-sampled data sets in which the minimum pseudo p-value is less than the original p-value, is

stated as adjusted p-value. The single-step *min P* adjusted p-values are defined by

$$\tilde{p}_g = Pr\Big(\min_{1\le l\le m} P_l \le p_g | H_0^C\Big), \tag{10}$$

where H_0^C refers to the complete null hypothesis, and P_l denotes the random variable for the unadjusted p-value of the l-th hypothesis. The alternative way is to determine single-step *max T* adjusted p-values defined as $\tilde{p}_g = Pr(\max_{1\le l\le m} |T_l| \ge |t_g||H_0^C)$, where T_l refers to the non-identically distributed test statistics of the l-th hypothesis (i.e., t-statistics of the lth hypothesis with different degrees of freedom) This method generates the greatest power among the methods for $FWER$ control. But, it is very slow processing method as permutations are included here. Therefore, it is not normally useful.

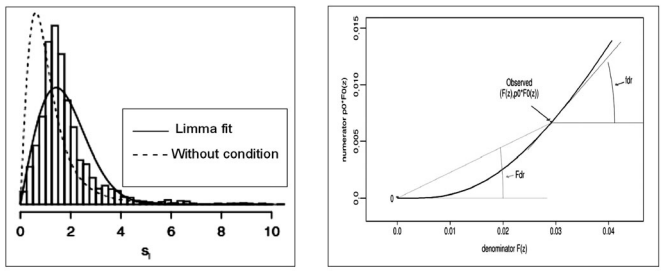
2.6.4 Benjamini and Hochberg False Discovery Rate (BHFDR/BHFdr) Correction [52], [53]

This correction is the least stringent among the all four corrections. It suffers from more false positives. At first, the p-values of genes/miRNAs are ranked in ascending order. The largest p-value remains the same. The second largest p-value is multiplied by the total number of genes/miRNA in gene list or miRNA list divided by its rank and if less than 0.05, it is significant: Corrected p-value = p-value$*(n/n-1) < 0.05$, if so, gene/miRNA is significant. For the third largest one: corrected p-value = p-value$*(n/n-2) < 0.05$, if so, gene/miRNA is significant, and so on.

The four methods are listed in order of their stringency, with the Bonferroni being the most stringent, and the Benjamini and Hochberg Fdr being the least. The stringency signifies that less false positive genes are coming out. Now, for better intuition, we have to go through the Fdr details. Actually, Fdr is basically the ratio of #False-discovery/#total-discovery; i.e., FP/(TP+FP) [for the notations, see Table 1]. It was first started with Benjamini and Hochberg [54]. This was derived for the strong control of the Fdr for independent test statistics. Suppose, $p_{r_1} \le p_{r_2} \le \cdots \le p_{r_m}$ refers to the observed ordered unadjusted p-values. To control of the Fdr at level α, g^* is defined as $max\{g : p_{r_g} \le (g/m)\alpha\}$ and reject hypotheses H_{r_g} where $g = 1, \ldots, g^*$. If there is no such g^*, then reject no hypothesis. The adjusted p-values are

$$\tilde{p}_{r_g} = \min_{k=g,\ldots,m} \Big\{ min\Big(\frac{m}{k} p_{r_k}, 1\Big)\Big\}. \tag{11}$$

For large-scale situations, the Fdr based on an **empirical Bayes method** [55] is very useful, where a versatile technique for both size and power are considered instead of strong Bayesian or frequentist assumptions. Let, there are N null hypotheses to consider simultaneously, each with its own test statistic. Suppose, the N cases (i.e.,"genes" for microarray studies) are subdivided into two types of class-labels/hypotheses, either null or nonnull with prior probability p_0 or $p_1 = 1 - p_0$, and with z-values having density $f_0(z)$ or $f_1(z)$ (where $p_0 = Pr\{null\}$ if $f_0(z)$ is null and $p_1 = Pr\{nonnull\}$ if $f_1(z)$ is nonnull). Let, $F_0(z)$ and $F_1(z)$ be the cumulative distribution functions corresponding to $f_0(z)$ and $f_1(z)$, respectively. According to them, $\overline{Fdr}(z) = p_0 F_0(z)/\overline{F}(z)$, where empirical cdf (of z values) $\overline{F}(z) = \#\{z_i \le z\}/N$; p_0 is chosen as 1 and a control level q is set as 0.1. Then declare those genes as nonnull having z-values z_i which satisfies $z_i \le z_0$; [here, z_0 is the maximum value of z satisfying $\overline{Fdr}(z_0) \le q$]. The expected number of null genes is be not greater than q. Storey [53] pointed out the necessity of some Bayesian interpretation.



(a) (b)

Fig. 3. (a) The overall variance filtering preferentially discards the genes with small variance s_g^2, due to the relative-shifting of the gene-level standard deviation estimates s_g^2 of the genes passing the filter to the unconditional distribution or standard normal distribution of the data (dashed curve). Here, when an estimate of d_0^2 is equal to ∞, then the Γ^{-1}-distribution (solid curve) model cannot able to produce a good fit to the s_i passing the filter and instead of that, the test-statistics are inappropriately compared to a unconditional distribution (see dashed curve) [23]. (b) Relationship between Fdr(z) to fdr(z) [41]: Heavy curve plots numerator of Fdr (i.e., $p_0 F_0(z)$) versus denominator (i.e., F(z)), where fdr(z) is slope of tangent and Fdr slope of secant.

Densities are more natural than tail areas for Bayesian fdr methodology. From the mixture density:

$$f(z) = p_0 f_0(z) + p_1 f_1(z). \tag{12}$$

Bayes rule provides $fdr(z) \equiv Pr\{null|Z = z\} = p_0 f_0(z)/f(z)$, for the probability of a gene of the null group where z-score is given. Here $fdr(z)$ is the **local false discovery rate** (Efron et al.) [56]. The relationship between $Fdr(z)$ and $fdr(z)$ is: $Fdr(z) = E_f\{fdr(Z)|Z \le z\}$, where E_f indicates the expectation with respect to the mixture density $f(z)$, $Fdr(z)$ is the mixture average of $fdr(Z)$ for $Z \le z$. Basically, $fdr(z)$ decreases when $|z|$ increases. $Fdr(z)$ will be smaller than $fdr(z)$. If we consider to label all genes with z_i which is less than some negative value z_0 as nonnull, then $fdr(z_0)$, the false discovery rate at the boundary point z_0, will be greater than $Fdr(z_0)$. Then the average false discovery rate goes beyond the boundary. The Benjamini-Hochberg Fdr control system denotes an upper bound on the secant slope (see Fig. 3b).

In the empirical Bayes approach of Robbins and Stein (1955), an appropriate prior distribution is to be estimated from the data (see Efron (2003) [55]). For local fdr calculations, N must be large (i.e., N $\ge$ 100), but the z_i is not necessary to be independent. For estimating the local false discovery rate $fdr(z) = p_0 f_0(z)/f(z)$, we assume the null distribution: $f_0(z) = \varphi(z) \equiv \frac{1}{\sqrt{2\Pi}} e^{-(1/2)z^2}$. Here, the z-values have been found by transforming the usual two-sample t statistic comparing treated and control expression labels for the gene i, to a standard normal scale through $z_i = \Phi^{-1}(F_{100}(t_i))$, where Φ and F_{100} are the cdfs of standard normal and t_{100} distributions respectively. When p_0 is taken, the prior probability of a gene becomes null. Benjamini and Hochberg [54] took $p_0 = 1$. An upper bound for $Fdr(z)$ has been provided.

This leaves us with only the denominator $f(z)$ to calculate. From the definition of Eq. (12), as $f(z)$ is the marginal density of all N z_is, we can utilize all the data to calculate $f(z)$. The local fdr algorithm is available as a R function *locfdr* from the CRAN library at the following link: http://cran.r-project.org/web/packages/locfdr/index.html.

Let, the z-values have been binned, producing the bin-counts $y_k = \{z_i \text{ in bin } k\}$, where $k = 1, 2, \ldots, K$ (K is total number of bins). The nonnull counts $y_k^{(1)} = [1 - \widehat{fdr}_k]y_k$, which is the estimated fdr value at the center of bin k (where $\widehat{fdr}_k = \widehat{fdr}(x_k)$). As (1-$\widehat{fdr}_k$) approximates the nonnull probability for a gene in bin k, the $y_k^{(1)}$ is a significant estimate for the expected number of nonnulls.

Power diagnostics are estimated from comparisons of $\widehat{fdr}(z)$ with the nonnull histogram (i.e., it is equal to the expected nonnull fdr). There is an indication of high power if $\widehat{fdr}_k$ is relatively small where $y_k^{(1)}$ is large. The definition of the power diagnostic [41] is: $\widehat{Efdr}^{(1)} = \sum_{k=1}^{K} y_k^{(1)} \widehat{fdr}_k / \sum_{k=1}^{K} \hat{y}_k^{(1)}$. To obtain high power, it is necessary that $\widehat{Efdr}^{(1)}$ should be small, (near 0.2), so that a typical nonnull genes will come out. Suppose, any data set has $\widehat{Efdr}^{(1)} = 0.72$, therefore it indicates low power. Efron [41] discusses about the power analysis for microarray data. The details about the empirical Bayes Fdr method is described in Efron [57]. There is another Fdr control method suggested by Benjamini and Yekutieli (i.e., BY) in 2001 [58].

2.7 Data Dependency

2.7.1 Surrogate Variable Analysis (SVA) [59]

It has been observed that gene expression levels can be affected by multiple factors including genetic, environmental, demographic and technical factors. Expression heterogeneity (i.e., EH) can decrease power, and can produce unwanted dependency among genes and sources of spurious signal to many genes because of unmodeled factors (i.e., $UMFs$) (viz., age, gender, etc.). It also increases the variability of the ranking of genes in expression levels and distorts the null distribution. Thus, SVA has been applied to overcome the problems because of EH. If there is only one surrogate variable (i.e., SV) in an analysis, then the accuracy of it can be determined by correlation. But, there are more than one UMF or more than one SV, then multiple regression should be applied and coefficient of determination (i.e., R_2) must be estimated. SVA can correctly rectify the genes which are affected by the UMF. If the UMF and the primary variable are correlated to each other, the null p-values become very small and it tends towards zero. If these are uncorrelated, the null p-values become very high and are biased towards one. SVA calculates properly the unobserved factor, and it also generates a correct Uniform distribution of null p-values. SVA can be useful for time course, disease class, and gene expression. It can able to enhance the consistency, reproducibility and biological accuracy in genome-wide expression analyses by making adjustment for surrogate variables.

Suppose, $X_{m\times n} = (x_1, \ldots, x_m)^T$ is a normalized $m \times n$ expression matrix with n number of arrays/samples for m number of genes, where each $x_i = (x_{i1}, \ldots, x_{in})^T$ refers to the vector of normalized expression for gene i. Let us assume that $y = (y_1, \ldots, y_n)^T$ is a vector of length n denoting the primary variable of interest. Now, generality model $x_{ij} = \mu_i + f_i(y_j) + e_{ij}$, where μ_i is the baseline level of expression, $f_i(y_j) = E(x_{ij}|y_j) - \mu_i$ provides the relationship between the measured variable of interest and gene i, and e_{ij} is random noise with mean zero. Fo example, for a dichotomous outcome $y_j \in \{-1, 1\}$, we can utilize the linear model $x_{ij} = \mu_i + \beta_i y_j + e_{ij}$; and calculate μ_i and β_i by least squares. After that, we can perform a standard test of whether $\beta_i = 0$ or not for each gene i. This hypothesis test is totally equivalent to performing a test of differential expression between the two class labels. Suppose, there are L biologically meaningful $UMFs$ in a microarray analysis, and $g_l = (g_{l1}, \ldots, g_{ln})$ is an arbitrarily complex function of the lth factor towards all n arrays/conditions, where $l = 1, 2, \ldots, L$. Hence, the expression model for gene i on array j is: $x_{ij} = \mu_i + f_i(y_j) + \sum_{l=1}^{L} \gamma_{li} g_{lj} + e_{ij}^*$, where γ_{li} is a gene-specific coefficient for the lth UMF. If unmodeled factor l does not influence the expression of gene i, then $\gamma_{li} g_{lj} = 0$. As there are n number of arrays and the expression of each gene can be represented by at most n linearly independent factors, so any dependence between genes can be modeled using $L \leq n$ vectors. For this formulation, the inter-gene dependent e_{ij} is replaced by $\sum_{l=1}^{L} \gamma_{li} g_{lj} + e_{ij}^*$. Here, e_{ij}^* is the true gene-specific noise which is now independent across genes. In other sense, the error e_{ij} is divided into two sub-parts, where the first sub-part (i.e., $\sum_{l=1}^{L} \gamma_{li} g_{lj}$) denotes the dependent variation across genes because of $UMFs$; and the second sub-part (i.e., e_{ij}^*) denotes gene-specific independent fluctuations. As the direct calculation of the unmodeled g_l is not possible, thus it is replaced by an orthogonal set of vector h_k (where, $k = 1, \ldots, K$; $(K \leq L)$ which has a same linear space as g_l. The vector h_k is also called as SV. The coefficient γ_{li} is also replaced by the mutually orthogonal coefficient λ_{ki}. Hence, the new model should be: $x_{ij} = \mu_i + f_i(y_j) + \sum_{k=1}^{K} \lambda_{ki} h_{kj} + e_{ij}^*$. The SVA algorithm can be divided into two subparts which are described in the followings:

(a) Detection of unmodeled factors: At first, we should make the estimates $\hat{\mu}_i$ and $\hat{f}_i$ by fitting the model $x_{ij} = \mu_i + f_i(y_j) + e_{ij}$ and determine $r_{ij} = x_{ij} - \hat{\mu}_i - \hat{f}_i(y_j)$ to eliminate the effect of the primary variable on expression. Make the $m \times n$ residual matrix R, where r_{ij} is the (i, j)th element of R. Thereafter, calculate the singular value decomposition of $R = UDV^T$. Let, d_l be the lth eigenvalue, which is the lth diagonal element of D, for $l = 1, \ldots, n$; df is the degree of freedom of the model fit $\hat{\mu}_i + \hat{f}_i(y_j)$. Now, for eigengene $k = 1, \ldots, n - df$ the observed statistic becomes: $T_k = \frac{d_k^2}{\sum_{l=1}^{n-df} d_l^2}$. Thereafter, a matrix R^* has been formed by permuting each row of R independently for omitting any structure in the matrix. Suppose, r_{ij}^* be the (i, j)th entry of R^*. Therefore, fit the model $r_{ij}^* = \mu_i^* + f_i^*(y_j) + e_{ij}^*$ and determine the corresponding residuals $r_{ij}^0 = r_{ij}^* - \mu_i^* - f_i^*(y_j)$ to make the $m \times n$ model-structured null matrix R_0; and determine the singular value decomposition of the centered and permuted expression matrix $R_0 = U_0 D_0 V_0^T$. So, for eigengene k, the corresponding null statistic becomes:

$T_k^0 = \frac{d_{0k}^2}{\sum_{l=1}^{n-df} d_{0l}^2}$, where d_{0l} is the lth diagonal element of D_0. Repeat the steps of permutation by B times to get null statistics T_k^{0b} for $b = 1, \ldots, B$ and $k = 1, \ldots, n - df$. Thereafter, calculate the corresponding p-value for eigengene k as

$$p_k = \frac{\#\{T_k^{0b} \geq T_k; b = 1, \ldots, B\}}{B}. \quad (13)$$

When eigengene k' (here, $k' > k$) is significant, then automatically k is significant. Therefore, the monotonicity among the p-values has been conservatively forcefully forced. Thus, $p_k = max(p_{k-1}, p_k)$, where $k = 2, \ldots, n - df$.

(b) Construction of SVs: First, we should make the estimates $\hat{\mu}_i$ and $\hat{f}_i$ by fitting the model $x_{ij} = \mu_i + f_i(y_j) + e_{ij}$ and determine $r_{ij} = x_{ij} - \hat{\mu}_i - \hat{f}_i(y_j)$ to eliminate the effect of the primary variable on expression. Make the $m \times n$ residual matrix R, where r_{ij} is the (i, j)th element of R. Thereafter, calculate the singular value decomposition of $R = UDV^T$. Suppose, $e_k = (e_{k1}, \ldots, e_{kn})^T$ is the kth column of V, where $k = 1, \ldots, n$. The e_k are the residual eigengenes and show orthogonal residual expression heterogeneity signals independent of the signal because of the primary variable. Thereafter, $\hat{K}$ has been set as the number of significant eigengenes obtained from the above algorithm. Here, *significant eigengenes* refers to the eigengenes which have a greater proportion of variation than expected by chance. Thus, $k = 1, \ldots, \hat{K}$ for each significant eigengene e_k. Regress e_k on the x_i, where $i = 1, \ldots, m$ and then determine a p-value testing for an association between each genes expression and the residual eigengene to measure the strength of the association. Calculate the number of genes associated with the residual eigengene, as $\hat{m}_1 = [(1 - \hat{\pi}_0 Xm)]$, where π_0 is the proportion of genes whose expression is not truly associated with e_k, and $\hat{\pi}_0$ is an estimate to be formed according to [60]. Let $s_1, \ldots, s_{\hat{m}_1}$ be the indices of the genes where there are $\hat{m}_1$ smallest p-values from this test. Now the size of the expression matrix is reduced to $\hat{m}_1 Xn$. Thus, the new matrix $X_r = (x_{s_1}, \ldots, x_{s_{\hat{m}_1}})^T$. Thereafter, calculate the eigengenes of X_r (i.e., e_j^r for $j = 1, \ldots, n$). Suppose, $j^* = argmax_{(1 \leq j \leq n)} corr(e_k, e_j^r)$ and set $\hat{h}_k = e_{j^*}^r$; i.e., the estimate of the surrogate variable ($\hat{h}_k$) becomes equal to the eigengene of the reduced matrix ($e_{j^*}^r$) which is most correlated with the corresponding residual eigengene (e_k). This is the ideal technique to estimate SV. The model is

$$x_{ij} = \mu_i + f_i(y_j) + \sum_{k=1}^{K} \lambda_{ki} \hat{h}_{kj} + e_{ij}^*, \quad (14)$$

which represents as an estimate of the ideal model $x_{ij} = \mu_i + f_i(y_j) + \sum_{k=1}^{K} \lambda_{ki} h_{kj} + e_{ij}^*$. Here, singular value decomposition is applied in the SVA algorithms. As the singular value decomposition gives the uncorrelated variables which decompose the data through an additive linear way having the aim of minimizing the sum of squares, it seems to be the most appropriate decomposition. The R package of the SVA method is available online at: http://www.genomine.org/sva/.

2.7.2 Other Methods for Dependency Removal [61]

There are many methods except SVA. These are $LEAPP$ [61], $EIGENSTRAT$ [61], etc., to distinguish latent variable from the primary variable. It has been noticed that high throughput multiple hypothesis testing is very difficult in the presence of systemic effects and other latent variables. Those variables can change the level of tests and make correlations among multiple genes/tests. They alter the ordering of significance levels among genes and make a poor rankings. Therefore, it is necessary to isolate the latent variables from the primary. It has been noticed that $LEAPP$ produces better ranking of genes than SVA and $EIGENSTRAT$. Now, we describe $LEAPP$ in short. Suppose, $Y \epsilon \mathbb{R}^{m \times n}$ denotes a response matrix and $g \epsilon \mathbb{R}^n$ refers to a variable of interest or primary variable, where m and n denote #gene and #samples respectively. The linear association through the $m \times n$ matrix γg^T should be considered, where g is the variable, γ be a vector of N coefficients. Here, γ will be sparse, when most of the genes are not related to g. It is needed for adjustment due to covariates X (e.g., sex) per sample which is other than g. The total covariate term is βX^T, where β has coefficients. The product form UV^T should be considered due to the latent variables. Here, both terms U and V are not observable. An sample-wise constant noise with the variance has been considered to be different for each gene. The model for $LEAPP$ is: $Y = \gamma g^T + \beta X^T + UV^T + \Sigma E$, where U and V refer to the latent nonrandom rows (e.g., genes) and latent independent rows (e.g., samples); here, $E \sim \mathcal{N}(0, I_m) \otimes I_n)$ which is noise, and $\Sigma = diag(\sigma_1, \ldots, \sigma_m)$ which denote standard deviations. The above function is rewritten as

$$Y_{ij} = \gamma_i g_j + \beta_i^T X_j + U_i^T V_j + \sigma_i \varepsilon_{ij}, \quad (15)$$

where $1 \leq i \leq m$, $1 \leq j \leq n$. Here, β_i and U_i denote the ith rows of β and U, respectively, as column vectors. Similarly, X_j and V_j are the jth rows of X and V, σ_i is the ith diagonal element of Σ and ε_{ij} is the ij element of E. $LEAPP$ algorithm is depended on a series of reductions. However, the key-point is to split the whole data into two parts, one of which is totally free from the primary variable. Thereafter, some characteristics of the latent variable model should be calculated from the primary-free data. The estimated quantities in the part of the data, that have the primary variable to identify genes related to the primary variable, should be applied. Moreover, in general, ordering (i.e., ranking) of genes is better in LEAPP than other two due to the better calibration of p-values. An R package for $LEAPP$ is available at: http://cran.r-project.org/web/packages/leapp/.

2.7.3 Batch Effect

It is found from different literatures that the microarray expression analysis suffers from the **problem of unwanted variation (i.e.,** $Uvar$**)**. The biological factors of interest (i.e., FOI) are the researcher, the influence of observed gene expression levels (i.e., $GExs$). One example of $Uvar$ is a batch effect, which is happened when some of the samples are processed differently than the remaining. The relevant factors for the **batch effect** might be different laboratory, different day, different technician, etc., (see [59], [62], [63], [64]). To eliminate the $Uvar$, two techniques can be useful. First one is global adjustment procedure (viz., Quantile-Normalization/QN). Second

technique is to use negative control genes (i.e., G_{-ct}), which is totally application specific. The G_{-ct} are those whose $GExs$ are known a priori not to be truly differentially expressed to the FOI. Conversely, positive control genes (i.e., G_{+ct}) are genes whose expression levels are known a priori to be truly associated with FOI. For example, $CyclinE$ is a positive control if the FOI is the presence or absence of ovarian cancer. G_{-ct} are harder to identify with certainty. Housekeeping (HK) genes are example of G_{-ct}. The variation in the expression levels of the G_{-ct} is to be $Uvar$. Lucas et al. [65] has utilized G_{-ct} for making adjustment of $Uvar$. It is called as **Remove Unwanted Variation, 2-step (RUV-2)** [66]. For applying any adjustment procedure, it should be known whether it is helping or hurting. Sometimes, it may be actually ambiguous. If $Uvar$ is roughly orthogonal to the FOI, then $Uvar$ will show itself as additional noise that obscure the true association between $GExs$ and FOI. Thus, an effective adjustment increases the number of discovered genes. But, if $Uvar$ is correlated with the FOI, then it prepares unauthentic associations between the FOI and $GExs$. So, an effective adjustment method will decrease the number of discovered genes. In fact, if the rate of misclassification is higher at with using the adjustment method than without adjustment, then the adjustment method is hurting. It may be equally happened that the adjustment method is working correctly, but, the resulting batch effects may effect the classification if experimental samples were processed in batches. There are three techniques [66] (for evaluating the quality of an adjustment) which are described below:

(a) *Control Genes or Gene Rankings*: G_{+ct} are useful to evaluate the quality in Differential Expression (i.e., DE) analysis. At first, p-values are determined for genes and rank of the genes are calculated in ascending order of their p-values. G_{+ct} must be toward the top of the list, and so we can choose some number of them (e.g., the top 25). If any adjustment substantially increases the number of top-ranked G_{+ct}, it seems to be effective. Here, the ranks of the p-values are used, not the p-values. Any good adjustment might increase or decrease the p-values of G_{+ct} based on the characteristics of the $Uvar$. The ambiguity can be resolved by this ranking. But, sometimes it is better to consider the p-values themselves than ranking p-values. But, for very small number of G_{+ct}, if their rankings do not change substantially after the adjustment, then the p-values of both positive and negative controls (i.e., $Pval_{+ct}$ and $Pval_{-ct}$ respectively) may be checked. If the $Pval_{+ct}$ substantially decrease and the $Pval_{-ct}$ do not, then adjustment seems to be helpful. If the $Pval_{+ct}$ and $Pval_{-ct}$ both decrease substantially, then the adjustment is an artifact. Similarly, if the $Pval_{+ct}$ and $Pval_{-ct}$ both increase, then the result is ambiguous. A better scheme is to utilize two different groups of G_{-ct}, one for making adjustment and other for the use in assessing the quality of the adjustment. For example, spike-in controls can be used for making an adjustment, and HK genes may be utilized for assessing the quality of the adjustment or vice versa. (b) *The p-value distribution*: In a DE analysis, the distribution of the p-values of the genes which are unassociated with the FOI, is uniformly distributed over the unit interval, whereas the p-values of the genes associated with the FOI becomes nearly zero. Therefore, a histogram of the p-values becomes ideally be nearly uniform, with a spike near zero. In fact, it is uncommon, as $Uvar$ tends to bring up dependence across measured gene expression levels. The adjustment for the $Uvar$ remove this dependence resulting the p-value histograms closer to the *ideal* [59], [62]. (c) *Relative Log Expression (RLE) Plot*: RLE plots are boxplots which are used to obtain the overall quality of a data set and also identify bad chips. If the chip has good quality, then the plot should be in the center position around zero and its width (interquartile range) should be around 0.2 or less.

2.8 Generation of Non-Normal Data Sets [67], [68]

Different approaches have been developed for preparing non-normally distributed data. The approaches discovered by Fleishman in 1978; and Vale and Maurelli are basically used to prepare multivariate non-normal random numbers in 1983. Skewness and kurtosis are two measures of the degree of asymmetry/inhomogeneity and degree of peakedness, respectively. Mattson proposed an approach to generate data on the latent variables with controlled kurtosis and skewness of the observed variables in case of inhomogeneity of data distributions among different groups in 1997. Population skewness (viz., γ_1) and kurtosis (viz., γ_2) are defined as $\gamma_1 = \mu_3/(\mu_2^{3/2})$ and $\gamma_2 = (\mu_4/\mu_2^2) - 3$, respectively. According to him, three different approaches can be applied to transform a univariate random variable as input to Mattsons methodology. For the first one (by Burr in 1942), non-normal distributions have been simulated using the equation given below:

$$Y = \frac{[(1-u)^{\frac{1}{k}} - 1]^{\frac{1}{c}}}{\sigma} - \frac{\mu}{\sigma},$$

where u is uniformly (0,1) distributed, μ, σ are the location and scale parameters respectively, c and k denote the shape parameters. The last four parameters are used for selected values of kurtosis and skewness. In the second one, Fleishman (1978) developed a polynomial transformation with specific values of kurtosis and skewness:

$$Y = a + bX + cX^2 + dX^3,$$

where Y is a non-normal variable, X is a random deviate which is normally distributed with zero mean and unit variance $(N(0;1))$, may be used to obtain non-normal distributions. The constants a, b, c, and d may be chosen such that Y has a distribution with specified moments of the first four orders, i.e., the mean, variance, skewness, and kurtosis. (For details, see [67].) In the third approach, a generalized Lambda distribution, which is applied (by Ramberg in 1979), is mentioned below:

$$Y = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2}, \tag{16}$$

where u is uniformly (0, 1) distributed, λ_1 and λ_2 are the location and scale parameters respectively, λ_3 and λ_4

signify shape parameters. The four parameters (viz., λ_1, λ_2, λ_3, and λ_4) are used for selected values of kurtosis and skewness (see [68]).

Two strategies, SAS (by Clark & Woodward in 1992) and PRELIS (by Joreskog & Sorbom in 1996), have been developed to compute the z-scores for kurtosis and skewness (see [68]). According to SAS methodology,

$$sample\_skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(x_i - \overline{x})^3}{s^3}, \quad (17)$$

and

$$sample\_kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \frac{(x_i - \overline{x})^4}{s^4} - 3\frac{(n-1)(n-1)}{(n-2)(n-3)}. \quad (18)$$

Now a days, different Matlab functions (e.g., 'gamrnd' function for 'gamma' distribution) and R functions are available online for generating the non-normally (e.g., skewed, bimodal, chi-square, gamma, exponential) distributed data.

3 Data Sets for Experiments

For experiments, we have used Lu's [22] miRNA expression data sets of 217 miRNAs for each of the colon, breast and uterus samples. For the colon data set (denoted by D_c), 10 samples are treated and five samples are control. For the breast data set (denoted by D_b), six samples are treated and three samples are control. For the uterus data set (denoted by D_u), such numbers are 10 and nine respectively. These three data sets are available at: http://www.nature.com/nature/journal/v435/n7043/pdf/nature03702.pdf.

Besides miRNA expression data sets, we also utilize two mRNA expression data sets. The first one is the Uterine Leiomyoma mRNA expression data set (data set's NCBI ref. id: GSE31699) (denoted by $D1$) belonging 16 uterine leiomyoma tumor and 16 normal myometrial samples where total number of mRNAs are 48,803. The second one is mRNA expression data set of cigarette smokers of lung adenocarcinoma (data set's NCBI ref. id: GSE10072) (denoted by $D2$) having 24 current smoker and 16 never smoker samples for Tumor category. It has a total of 22,283 mRNAs.

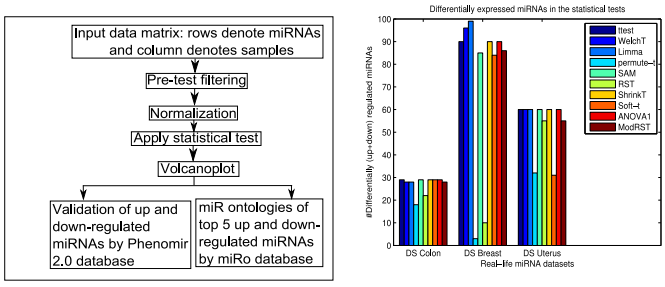
In addition to the five real data sets, we have prepared simulated data sets of different sample sizes (viz., 10, 15, 40, 60 and 100 samples per group individually) taking 1,000 genes under two different assumptions. In the first assumption, gene expressions for the experimental and control/normal groups are taken from the assumption of *normal/gaussian distributions* having same variance (i.e., $sd1 = sd2 = 2$), but different means (i.e., $\mu_1 = \mu_2 + 2$).

The second assumption is that the data of two groups are both drawn from *non-normal distributions* having same variance (i.e., $sd1 = sd2 = 2$), but different means (i.e., $\mu_1 = \mu_2 + 2$). There are other distributions like skewed, bimodal, chi-square, gamma, exponential distributions, etc. Any of these distributions can be considered. We have taken gamma distribution for the first group and exponential distribution for the second group (see supplementary file "$s0\_simulation.m$", for details). Note that that we have selected different distributions for the two groups as it will help to make asymmetry/inhomogeneity not only in data distribution for each group, but also between two groups.

4 Experimental Strategies and Results

Our goal is to identify correctly the differentially expressed up- and down-regulated genes/miRNAs [69], [70]. We have used Matlab and R softwares for our experiments. As input, the expression values of treated and control samples of specified miRNAs are taken from Lu's data sets. To do so, the following approach should be used.

Suppose, a input matrix is given whose rows indicate genes and columns denote samples/arrays. Use some pre-test filtering procedures. Here, we have determined the overall variance of each gene without considering their class-labels/samples and filtered out the genes having low variance (briefly described in Section 2.1.1). We have utilized 'genevarfilter' Matlab function to do so. Another pre-test filtering procedure is fold change filtering. This fold change filtering is merged with p-value filtering in volcanoplot which will be described next. After pre-test filtering, normalization procedure should be applied on the expression values of each miRNA/gene in the data sets. Here, we have utilized zero-mean normalization (described in Section 2.1.2). After normalization, normality test should be applied on the normalized data (See Section 2.1.3). We can use parametric tests on the data which follow normal distribution, and utilize non-parametric tests on the data which do not follow normal distribution as the parametric tests are well-fitted in normally distributed data and the non-parametric tests are well-suited in data with non-normal distribution. Thereafter, different statistical parametric and non-parametric tests can be applied according to the requirement. Matlab software package is utilized for one way ANOVA, Wilcoxon ranksum, modified ranksum and Pearson's correlation test. R Bioconductor package can be used for fold change, t-test, Welch's t-test, Limma, SAM, shrink t, softthreshold t, permute t-test and other tests (described in Fig. 1). The corresponding p-value of each gene/miRNA is estimated for each test. After that, volcanoplot is applied to identify differentially expressed up- and down-regulated miRNAs rankwise separately. We already mentioned that volcanoplot is such a plot where two filtering (viz., fold change filtering and p-value filtering) are integrated together. Here, $log_2(foldchangeratio)$ is plotted in horizontal axis and $\{-log_{10}(pvalue)\}$ is plotted in vertical axis, where the fold change ratio refers to the ratio of mean of expression values of experimental samples to mean of expression values of control samples. After all, multiple testing corrections should be used. For independent data, Bonferroni, Holm and Fdr corrections are suitable as these are considered data of each gene individually. Moreover, for dependent variables, SVA method should be utilized (described in Section 2.7.1). The data dependency can be checked by Pearson's



(a) (b)

Fig. 4. (a) The steps of applying any statistical test on D_c, D_b and D_u. (b) #Differentially expressed miRNAs in the different statistical tests for D_c, D_b and D_u, consecutively. (here, 'DS' refers to 'Data set'.)

correlation test. Subsequently, the corresponding heatmaps of differentially expressed miRNAs are observed individually. The ideal application procedure of different statistical tests on any gene/miRNA expression data to identify differentially expressed miRNAs/genes for microarray data is presented in Fig. 11.

4.1 Experiments and Results for Real miRNA Data Sets

For the three real miRNA data sets, D_c, D_b and D_u, pre-test filtering (i.e., miRNAs having low variance removal), zero-mean normalization, application of statistical tests, multiple testing corrections (if required), identification of differentially expressed miRNAs by volcanoplot through both p-value filtering and fold-change filtering are performed consecutively. The steps used for the three data sets are presented in Fig. 4a. The number of differentially (up + down) regulated genes identified using different statistical tests are shown in Fig. 4b (without applying any correction), Table 2 (with Bonferroni, Holm and FDR corrections), and Table 3 (with local fdr correction). Fig. 5 shows a volcanoplot for D_c using Welch's t-test. All the volcanoplots and clustergrams for the data sets are presented in the supplementary file "RSurvey-supplementary1.pdf" available online. For details, see "RSurvey-supplementary4.pdf" available online.

Besides these, the outcomes (i.e., the number of up- or down-regulated miRNAs for each specified tissue) of each test are also compared with the information of each miRNA of a standard miRNA database, PhenomiR 2.0 (http://mips.helmholtz-muenchen.de/phenomir/). Now,

TABLE 2
Total Number of Differentially (i.e., Up +Down-Regulated) miRNAs in the tests at 0.05 p-value cutoff for D_c, D_b and D_u.

Tests	Bonferroni			Holm			FDR		
	D_c	D_b	D_u	D_c	D_b	D_u	D_c	D_b	D_u
t-test	1	32	10	1	32	10	2	86	50
Welch's t	0	26	6	1	27	6	0	90	50
Limma	0	38	5	0	42	5	0	80	56
Permute	0	0	0	0	0	0	0	0	28
SAM	0	0	0	0	0	0	0	0	50
RST	0	0	13	0	0	13	0	6	48
ModRST	2	6	5	0	6	5	5	69	51
Shrink-t	1	31	9	2	32	9	2	88	49
Soft-t	1	31	0	1	32	0	2	78	27
ANOVA1	1	32	10	1	32	10	2	86	50
FC	-	0	0	-	0	0	-	0	0

(For the each data set, the correlation is not applicable due to inequality of number of samples between two groups.)

TABLE 3
Number of Differentially Expressed miRNAs Using the Different Statistical Tests with Local fdr Correction (Local Fdr Cutoff $= 0.2$) for D_c, D_b and D_u

Statistical Tests	$\#gene_{up}$		$\#gene_{down}$		
	D_c	D_u	D_c	D_b	D_u
t-stat	3	5	23	84	37
Welch's t	4	5	21	91	37
Limma	3	5	23	93	37
permute	2	5	22	0	27
SAM	3	3	23	84	36
Wilcoxon ranksum	4	5	17	9	46
modified ranksum	4	5	27	84	46
Shrink t	3	6	23	98	40
Softthreshold t	3	4	23	84	27
ANOVA 1	3	5	23	84	37

our target is to get the number of true positives (TPs), true negatives (TNs), false positives and false negatives (FNs) individually in the cases of up- and down-regulated miRNAs for the data sets. Using PhenomiR 2.0 database, we have checked whether information of the differentially expressed miRNAs obtained by the different tests exists for the same tissue type (e.g., colon, breast or uterus) and also for the same type regulation (e.g., up- or down-regulation). The calculated positive predictive rate (PPR), negative predictive rate (NPR), sensitivity, specificity and MCC of each test are reported in the supplementary file "RSurvey-supplementary4.pdf" in the case of down-regulation in breast data set. Brief information about the ranks and miR-Ontologies of top 5 up- and down-regulated miRNAs from the tests, and PhenomiR-verified down-regulated miRNAs in the tests for the breast data set can be found in supplementary file entitled as "RSurvey-supplementary2.pdf" available online. For miR-Ontologies, we have used $miRò$ i.e., the miR-Ontology Database (http://ferrolab.dmi.unict.it/miro/).

4.2 Experiments and Results for Real mRNA Data Sets

For the two real mRNA data sets, $D1$ and $D2$, at first pre-test filtering (i.e., 20 percent miRNAs having lowest variance

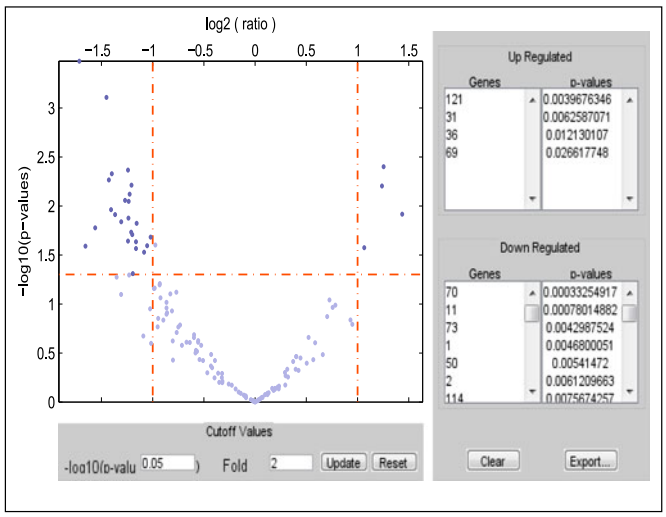


Fig. 5. Volcanoplot for identifying differentially expressed genes using Welch's t-test from D_c.

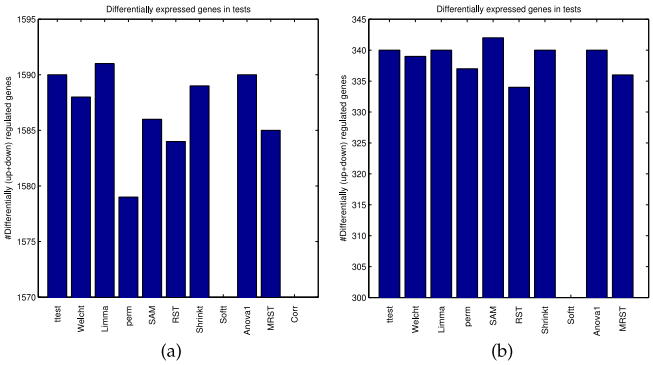


Fig. 6. #Differentially (up + down) regulated genes using different statistical tests for (a) $D1\_N$ and (b) $D2\_N$. For (a), the numbers of differentially regulated genes using Pearson's correlation and softthreshold t-test are very low (i.e., 304 and 119, respectively) as compared to that of others. For (b), the number using softthreshold t-test is also very lower (i.e., 59) than that of others.

removal), zero-mean normalization and Jarque-Bera normality test have been performed. After normality test, each data set has been divided into two sub-data sets, one following normal distribution ($D1\_N$ and $D2\_N$) and the other comprising the remaining mRNAs, i.e., non-normally distributed data ($D1\_NN$ and $D2\_NN$). Thereafter, application of statistical tests, multiple testing corrections (if required), identification of differentially expressed genes by volcanoplot through both p-value filtering and fold-change filtering are followed. The number of differentially (up + down) regulated genes identified using different statistical tests are shown in Figs. 6a, 6b, 7a, and 7b (without applying any correction), and Table 4 (with Bonferroni, Holm and FDR corrections).

4.3 Experiments and Results for Simulated Data Sets

As stated in Section 3, two categories of simulated data sets, one following normal distribution and the other following non-normal distribution, have been generated for different sample sizes. Pre-test filtering (i.e., removal of 10 percentile of genes having low variance), zero-mean

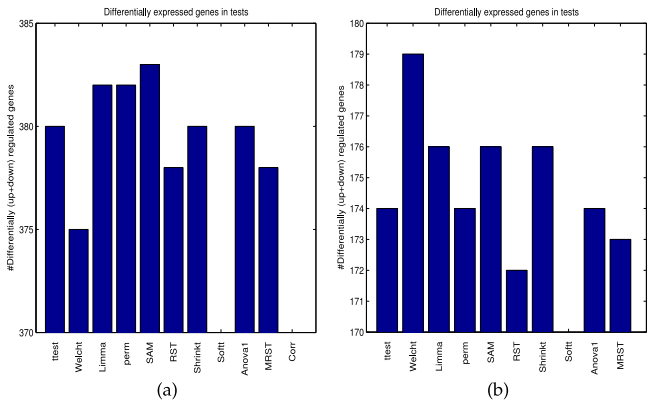


Fig. 7. #Differentially (up + down) regulated genes in different statistical tests for (a) $D1\_NN$ and (b) $D2\_NN$. For (a), the numbers of differentially regulated genes using Pearson's correlation and softthreshold t-test are very low (i.e., 46 and 40, respectively) as compared to that of others. For (b), the number using softthreshold t-test is also very lower (i.e., 21) than that of others.

TABLE 4
Total Number of Differentially Expressed Genes in the Statistical Tests at 0.05 p-Value Cutoff for $D1\_N$, $D2\_N$, $D1\_NN$ and $D2\_NN$. (For $D2\_N$ and $D2\_NN$, the Correlation is not Applicable due to Inequality of Number of Samples between two Groups.)

Tests	Bonferroni				Holm				FDR			
	D1_N	D2_N	D1_NN	D2_NN	D1_N	D2_N	D1_NN	D2_NN	D1_N	D2_N	D1_NN	D2_NN
t-test	65	13	1	2	65	13	1	2	1127	169	87	4
Welch's t	49	12	0	2	49	12	0	2	1095	156	0	2
Limma	0	1	0	0	0	1	0	0	360	21	0	0
Permute	423	31	77	11	423	31	77	11	423	31	77	11
SAM	61	7	0	0	61	7	0	0	406	28	58	0
RST	15	3	5	1	15	3	5	1	397	27	63	3
ModRST	15	3	5	1	15	3	5	1	397	27	63	3
Soft-t	0	0	0	0	0	0	0	0	29	0	0	0
ANOVA1	65	13	1	2	65	13	1	2	1127	169	87	4
Corr	0	-	0	-	0	-	0	-	21	-	3	-

(No differentially expressed gene is found by shrink t-test for all the sub-data sets with the three corrections.)

normalization, application of statistical test, multiple testing p-value correction (if required) and making the volcanoplot to identify differentially expressed genes through both p-value filtering and fold change filtering are consecutively performed. Thereafter, the power of each test is calculated. This procedure has been applied for different sample sizes (viz., $10, 15, \ldots, 100$). Finally, a comparative study has been done on the basis of the power of each test. Fig. 8 shows the steps of our experimental approach on the simulated data. Here, before applying any multiple testing correction, Pearson's correlation is used to identify the dependent data/genes. We have obtained very few dependent genes for each simulated data (approximately 0.034-0.048 percent in all the studies). The rest of the genes are independent. Bonferroni, Holm and FDR based p-value corrections are utilized on the independent genes.

The comparison of the power of the statistical tests is presented in Fig. 9. As seen in Figs. 9a and 9b, For small number of samples (up to 15 for each group), some power difference among the tests can be observed for both the normality and non-normality assumptions. When sample size increases, the power of test increases, but the power difference is minor. For such cases, any test either parametric or non-parametric is recommended.

In case of normality assumption and small sample sizes (viz., 10 and 15 in Fig. 9a), the performances of t-test and SAM are average; but, Limma and Shrink t-test

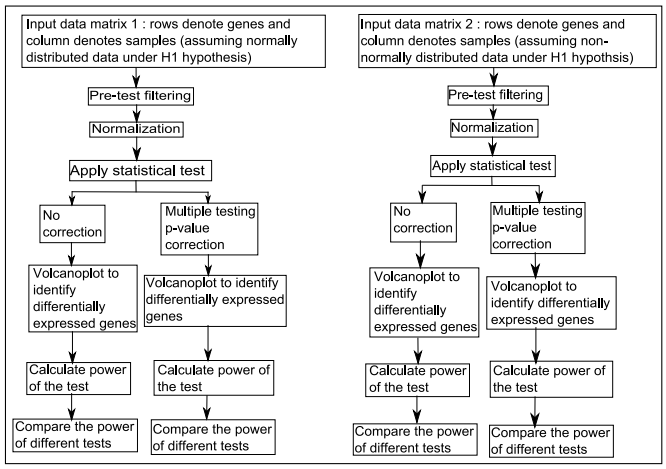


Fig. 8. The steps of applying any statistical test on simulated data in our experiment.

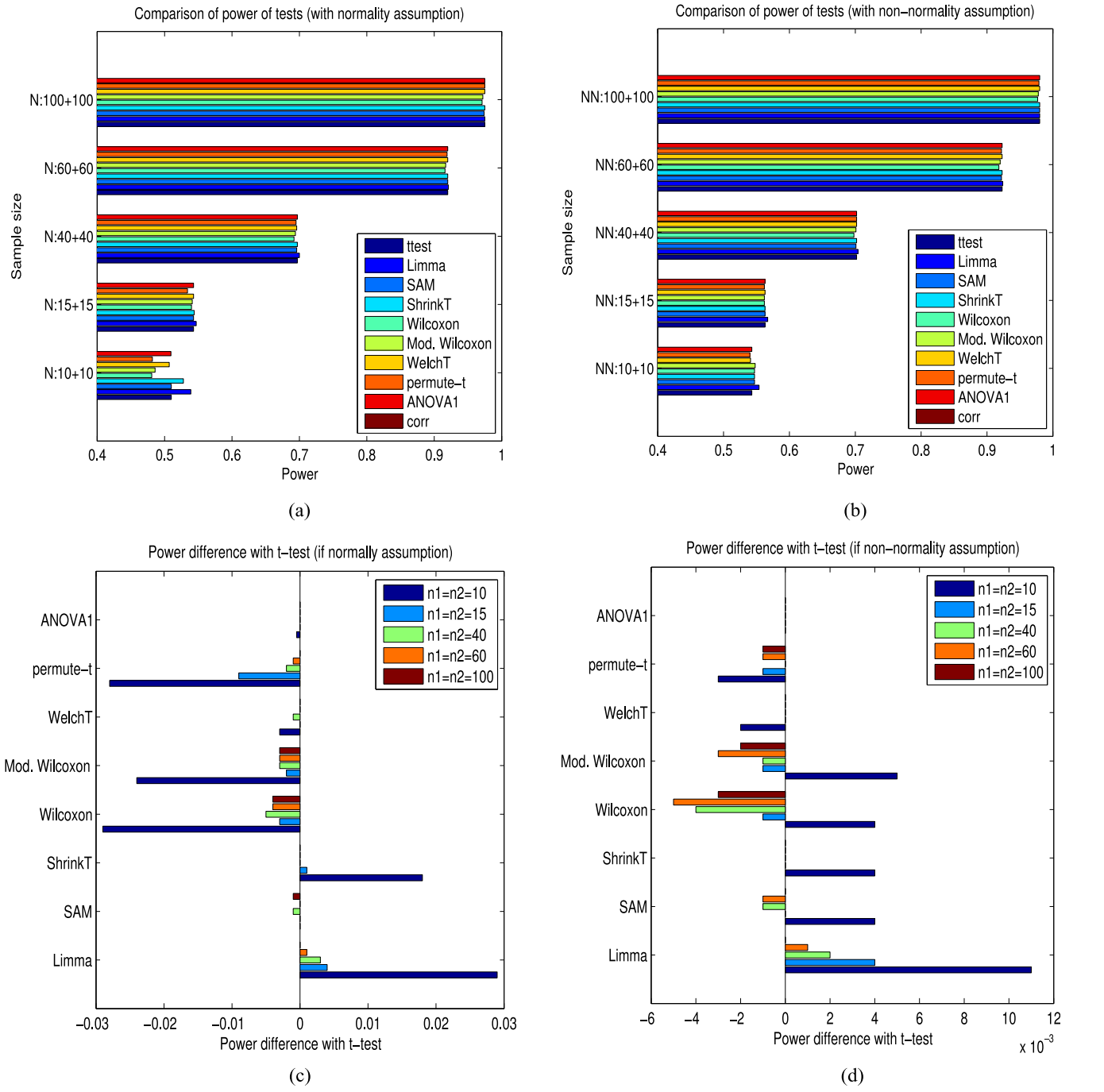


Fig. 9. (a) Comparison of power of the statistical tests on the simulated data for normality assumption (denoted as 'N' in figure) for the different sample sizes (viz., $10, 15, \ldots, 100$) (e.g., '$10 + 10$' refers to 10 samples for group 1 and 10 samples for group 2.); (b) comparison of power of different tests on the simulated data for non-normality assumption (denoted as 'NN' in figure) for the different sample sizes (viz., $10, 15, \ldots, 100$); (c) difference of power with t-test for normality assumption of the data for the different sample sizes and (d) difference of power with t-test for non-normality assumption of the data for the different sample sizes.

produce much better performance. Here, Wilcoxon's ranksum, modified ranksum and permuted t-test produce poor performance here. However, for non-normality assumption, performances of both ranksum tests and permute t-test are quite satisfactory. For the same number of samples, Limma and SAM provide better power for the non-normality assumption of data than the normality assumption. These signify that the non-parametric tests have less power than parametric tests for normally distributed data. For 10 samples per group, Limma produces best performance for both the data distribution assumptions. The performances of t-test and one way ANOVA are almost similar for all situations.

The summary of the different tests used here are represented in Table 5 on the basis of the power of these.

5 Discussion

Based on our experiments, some significant observations have been made which are described below:

TABLE 5
Summary of Different Statistical Tests (Here, Pearson's Correlation and Softthreshold t-Test Are Not Mentioned As Both Provide Extremely Poor Performance for All the Situations)

tests	Power				Simplicity
	Small samples		Large samples		
	normal	non-normal	normal	non-normal	
t-test	+	+	++	++	++
Welch's t	+	-	++	+	++
ANOVA 1	+	+	++	++	++
permute t	-	-	+	++	-
SAM	+	+	+	++	-
Limma	++	++	++	++	+
Wilcoxon	-	+	+	+	++
mod. Wilcoxon	-	+	+	+	++
Shrinkage t	+	+	++	++	+

"Here, '+', '++' and '−' denote good, better and poor performances, respectively."

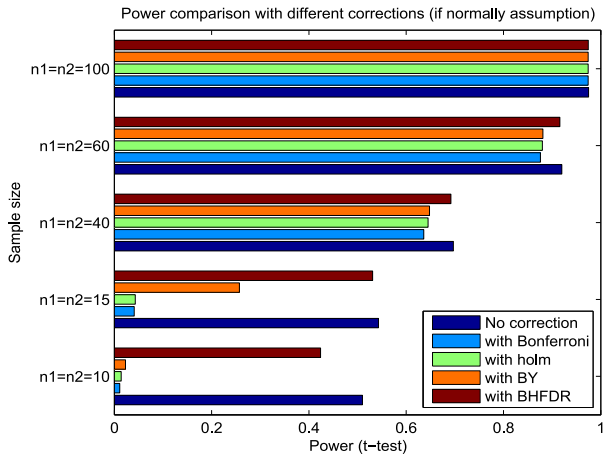


Fig. 10. Barplots: Comparison of power of t-test with corrections and without correction on the simulated data for normality assumption. Observation: BHFDR has more power to identify differentially expressed transcripts than other multiple testing correction methods specially for small sample sizes.

Observation 1) Normality test should be utilized before applying any statistical test for identifying differentially expressed transcripts specially for small number of samples. Otherwise, p-value may be misleading due to the assumption of incorrect distribution.

Observation 2) Limma and Shrink t-test are both performing well in all conditions specially for small sample sizes, but Limma is the best performer in all cases (for small and large sample sizes) for simulated data (e.g., they generally perform better than or equal to t-test presented in Figs. 9a, 9b, 6a, 6b, and 4b).

The performances of t-test and ANOVA 1 are quite similar as can be seen from Fig. 9 (for the simulated data sets) and Figs. 6a, 6b, 7a, 7b and 4b (for the real data sets).

For small sample size, performance of Welch's t-test is poorer than standard t-test for both normality and non-normality assumptions of data distribution in most of the cases, especially when the variance of the data of two groups are the same (e.g., Figs. 9a and 9b, for simulated data sets).

Our experiments show that SAM generally performs well, even for small samples sizes [see Figs. 4b (for D_c and D_u), 6b, 7a and 7b]. However, it has been pointed out in [71] that SAM may sometimes fail for small sample sizes.

The sample standard error correction in SAM is not model-motivated where unstable variance estimate is corrected using Error fudge Factors. Therefore, it is very difficult to know when it will fail.

It can be observed from our results that the performance of SAM and Limma are comparable. However, as mentioned earlier, the performance of SAM may not be consistent for small sample sizes and it may produce more false discoveries [71]. Moreover, the smoothing in Limma (done using a Bayesian approach) is more than that in SAM (which is controlled by a fudge factor s_0) [71].

The performance of permuted t-test is satisfactory in case of non-normal distributions for all types of sample sizes (e.g., Figs. 7a, 7b and 9b). But, in case of normal distributions, it works poorly (viz., Fig. 6a), especially for small sample sizes.

To summarize, it can be stated that for small number of samples (e.g., $10, 11, \ldots, 40$), Limma is very useful for both the distribution assumptions. Besides this, Shrink t-test, SAM and permuted t-test can also be applied for only non-normal distributions.

Observation 3) Nonparametric tests are less powerful than parametric tests that assume normal distribution (see the performances of Limma, SAM, Wilcoxon's ranksum, modified ranksum and permuted t-test for the normal data distribution assumption in Fig. 9a); here, p-values have a tendency to be higher, making it more difficult to detect real differences.

Observation 4) For large number of samples (say, > 40), the differences in powers of the tests are very low (viz., Figs. 9a and 9b). Thus, any test, either parametric or non-parametric, can be used.

Observation 5) The performances of Pearson's correlation and softthreshold t-test are much poorer than those of the others for both the distributions and all sample sizes.

For the simulated data, the range of power of Pearson's correlation method is [0.023-0.048], which is very low as compared to that of the others (see Figs. 9a and 9b). Again, for the $D1\_N$ (mRNA) data set, while most of the other methods identify around 1,500 differentially expressed genes, Pearson's correlation approach provides only about 304 genes (see Fig. 6a and Table 4). Similarly, for the $D1\_NN$ data set, the number of differentially expressed genes are 46, which is much lower than that of others (see Fig. 7a and Table 4).

Figs. 6a, 6b, 7a and 7b demonstrate the poor performance of softthreshold t-test for both distributions for all sample sizes. Similar behaviour has been found for the simulated data sets.

Wilcoxon ranksum and modified ranksum tests perform inconsistently. Generally, they produce much poorer performance than t-test especially for small sample sizes for both the distributions (presented in Figs. 6a, 6b, 7a, 7b and 9a). But, for non-normally distributed data, it is better than the case of normally distributed data (viz., Fig. 9b).

To summarize, softthreshold t-test, Pearson's correlation, Wilcoxon ranksum and modified ranksum tests are generally not useful for identifying differentially expressed genes/miRNAs except some special situations (described in Table 7).

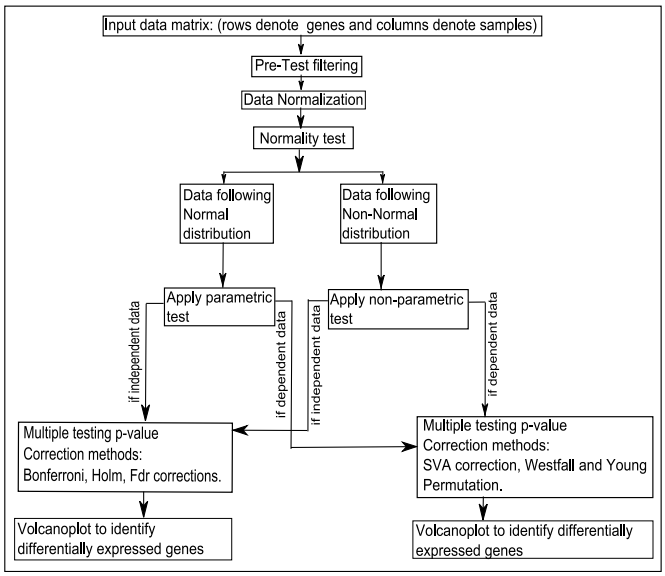


Fig. 11. The ideal approach of applying any statistical test on microarray data.

Observation 7) FDR produces more power to identify differentially expressed transcripts than the other correction methods specially for small sample sizes (viz., Fig. 10, and Tables 2 and 4). For large sample sizes (viz., ≥ 100 according to Fig. 10), all correction methods produce almost the same power.

In addition to the above observations, some more points need to be mentioned. Independent pre-test filtering increases the power of identifying differentially expressed genes. Filtering on overall gene-wise variance is better than filtering on overall mean as the latter removes many significant genes, while the former retains them. If the data are dependent sample-wise (due to batch-effect or other unwanted variations), the p-values of the genes will be incorrect. SVA and LEAPP corrections are ideal for solving sample-wise dependency due to batch-effect.

Nonparametric methods can be useful for dealing with unexpected outlying observations, that might be problematic with a parametric approach. In Table 6, some assumptions are listed for parametric/nonparametric tests for large/small number of samples. Nonparametric tests have bias towards hypothesis testing than estimation-effects. Obtaining nonparametric estimates and related confidence intervals are not straightforward. Tied values might be problematic here if these are common. Thus, adjustments to the test statistic is necessary.

In variance tests [71], using their own variance estimates, each method computes a t-statistic, either based on equal variances (SAM, t-test or t-statistic, Limma, Shrinkage-t) or unequal variances (Welch's t-test). If data comes from two normal populations with the same variances, the two-sample t-test is as powerful as or more powerful than Welch's t-test. However, Welch's t-test is approximation-based and its performance with small number of samples may be questionable. Limma is based on a model where the variances of the residuals vary from gene to gene and are assumed to be drawn from a chi-square distribution. In case of fold change, variability

TABLE 6
Choice of a Parametric (i.e., Param) or Nonparametric (i.e., Nonparam) Test

Test	For large samples ($>= 100$)	For small samples ($<= 10$)
param	• Robust. • p-value is approximately correct even if population is not normally distributed.	• Not robust. • If the population is not normally distributed, the p-value may be misleading.
nonparam	• Powerful. • If the population is normally distributed, the p-value will be approximately identical to the p-value obtained from a parametric test. • With large sample sizes, nonparametric tests are almost as powerful as parametric tests.	• Less powerful. • If the population is normally distributed, the p-value will be higher than the p-value obtained from a t-test. • If the population is non-normal, then it can be used to some extent. But, it is not really good enough as it looses power.

check between two populations is simply ignored. The property "number of replicates" [72] is another important condition. It is observed that when there are only two replicates of miRNAs, the Regularized t-test (Welch's t-test) performs the best. Except for two replicates, the Shrinkage t-test performs the best. It is also noticed that when there are 50 or more replicates, the choice of the test method becomes less important, as all of the methods perform similarly. The usefulness [7], [16], [19], [37], [39] and pitfalls [14], [15], [32], [73] of different statistical tests are listed in the Table 7.

Different new testing methodologies have been proposed in recent times like LEMMA [74], LPE [17], VarMixt [75], RVM [76], SMVar [77], WAME [78], cyber t-test [17], BRB t-statistic [17], etc. These methods are not discussed in this paper as the performances of VarMixt, LEMMA, WAME, BRB t-statistic are quite similar to that of Limma, while SMVar and RVM provide similar performance with Wilcoxon ranksum test (see [79]).

6 Conclusions

In this paper, we provide a comprehensive survey of different parametric and non-parametric approaches for identifying differentially expressed transcripts. A number of statistical tests have been applied to identify differentially expressed miRNAs/genes at 0.05 p-value cutoff for four real-life miRNA expression data sets and two mRNA expression data sets. For the computational verification of the resulting differentially expressed miRNAs, evidence is collected from PhenomiR 2.0 database, and a list of miRNAs categorized as true positives is created.

Subsequently, we have prepared different simulated data sets of different sample sizes (from 10 to 100 per group/population), and thereafter, the power of each test has been calculated individually. The comparative simulated study might lead to formulate robust and comprehensive judgements about the performance of each test on the basis of assumption of data distribution. We observe that the selection of the differentially expressed up- or down-regulated transcripts totally depends heavily on the choice of the statistical testing methodology. The performances of these testing methods are affected by certain preliminary conditions like sample size, distributional assumption, the variance structure, number of replicates of genes/miRNAs, etc. Hence, for obtaining the reliable testing results to identify

TABLE 7
Advantages and Pitfalls of Different Statistical Tests for Detecting Differentially Expressed miRNAs (Here, 'corr', 'ranksum' and 'soft-t' Denote Pearson's Correlation, Wilcoxon Ranksum and Softthreshold T-Sstat, Respectively)

Test		Advantages	Limitations
	FC	• Simple bio-interpretation. • Useful for data with very few samples (say 1-2).	• Variability is ignored. Outliers affect largely. • More FPs.
Parametric	t-test	• Work properly with large samples if data come from two normal populations with same variances.	• Work average with small samples. • Incur a higher FPs with increasing variance. • Irrelevant for comparing more than two groups.
	Welch's t	• Assumes unequal variances of 2 groups. • Compare means of 2 independent groups.	• Its performance in small sample sizes may be questionable.
	ANOVA 1	• Compares means of 3 or more groups. • Work well with large sample sizes in normal data. • Similar performance with t-test.	• Inappropriate if the different columns represent different variables instead of different groups. • Work average with small sample sizes.
	corr	• Commonly used in linear regression. • This Coefficient measures both the degree and direction of the correlation between two variables.	• Valid for linear relationship between the variables. • Is unduly affected by the values of extreme items.
Nonparametric	permuted t	• Goal is to get confidence intervals. • Use it if 2 groups are distribution free. • Simulate the null distribution repeatedly randomly reassigning group labels.	• Taking more time to compute test statistic than t-test.
	ranksum	• Robust for non-normal data having outliers. • Useful for ordinal-based data. • The difference between the 2 groups's medians (or any other measure of location) is obtained by inverting the test.	• Is less likely to detect a location shift than two-sample t-stat. • Performance is not satisfactory specially for small sample sizes.
	SAM	• Avoids small variance problem. • Useful for small sample sizes. • Uses permutation for correlations in genes and avoids parametric assumptions about the distribution of individual genes. • Assumes equal variance, independence of genes. • Reports local FDR and miss rates. • Uses **a fudge factor s_0 for sample variance correction.** • Correlates expression data with time.	• **Not consistently** performing well for small sample sizes. • sample variance correction technique is **not model-motivated**.
	Limma	• Useful especially with small number of samples. • **Estimates variance of Array(Group) prior** to fitting the model, using the information of all genes and filters out the genes having low variance. • Uses **an empirical Bayes model** to correcting sample variance. • Can be used to compare two or more groups. • Uses the *degrees of freedom* associated with each variability term as weighting factors. • Can be used for multifactorial designs (e.g. genotype and treatment).	• It makes the assumption that all the populations have the same standard deviation.
	shrink-t	• *Stabilizing t-stat* with very small replicates. • Uses when the over-fitting problem of 'large number of genes and small number of samples' arises. • Performs best for a lot of replicates of genes (except 2 replicates).	• Its performance is not satisfactory in case of 2 replicates of gene.
	soft-t	• Uses when the over-fitting problem of 'large number of genes and small number of samples' arises.	• Estimators have low variances. • The **weak consistency** of penalized L1 estimators.

differentially expressed transcripts in microarray data analysis, we first need to extract the actual characteristics of the given data and thereafter apply the most appropriate testing method on that data.

The traditional methodologies of microarray expression data often apply either statistical test to identify differentially expressed transcripts, or a clustering algorithm to measure groups of transcripts that behave almost similarly. These strategies may fail to focus the groups of differentially co-expressed transcripts [80] which is now being found to provide valuable biological insights. A pair of biomolecules is said to be differentially co-expressed under two conditions if their co-expression values differ significantly over these conditions. Appropriate statistical tests for measuring differential co-expression are not many, and further work is required in this direction.

7 Supplementary Materials

The supplementary materials are available at: https://www.dropbox.com/sh/a504ga06g9r4ta6/IZWVhWyUUr and http://www.isical.ac.in/~bioinfo_miu/supplementary%20file.rar.

References

[1] J. Phillips and V. Corces, "CTCF: Master Weaver of the Genome," *Cell*, vol. 137, pp. 1194-1201, 2009.

[2] R. Ramanathan, S. Varma, J. Ribeiro, T. Myers, T. Nolan, D. Abraham, J. Lok, and T. Nutman, "Microarray-Based Analysis of Differential Gene Expression between Infective and Noninfective Larvae of Strongyloides stercoralis," *PLoS Neglect Trop D*, vol. 5, no. 5, p. 1039, 2011.

[3] U. Maulik, A. Mukhopadhyay, and S. Bandyopadhyay, "Combining Pareto-Optimal Clusters Using Supervised Learning for Identifying Co-Expressed Genes," *BMC Bioinformatics*, vol. 10, article 27, pp. 1-16, 2009.

[4] U. Maulik, S. Bandyopadhyay, and A. Mukhopadhyay, "Multi-Class Clustering of Cancer Subtypes through SVM Based Ensemble of Pareto-Optimal Solutions for Gene Marker Identification," *PLoS One*, vol. 5, no. 11, p. e13803, 2010.

[5] A. Mukhopadhyay, U. Maulik, and S. Bandyopdhyay, "On Biclustering of Gene Expression Data," *Current Bioinformatics*, vol. 5, no. 3, pp. 204-216, 2010.

[6] U. Maulik and A. Mukhopadhyay, "Simulated Annealing Based Automatic Fuzzy Clustering Combined with ANN Classification for Analyzing Microarray Data," *Computers & Operations Research*, vol. 37, no. 8, pp. 1369-1380, 2010.

[7] B. Wu, "Differential Gene Expression Detection and Sample Classification Using Penalized Linear Regression Models," *Bioinformatics*, vol. 22, pp. 472-476, 2006.

[8] Z. Joseph, A. Gitter, and I. Simon, "Studying and Modelling Dynamic Biological Processes Using Time-Series Gene Expression Data," *Nature Reviews Genetics*, vol. 13, pp. 552-564, 2012.

[9] S. Mallik, A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "Integrated Analysis of Gene Expression and Genome-Wide DNA Methylation for Tumor Prediction: An Association Rule Mining-Based Approach," *Proc. IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE Symp. Series on Computational Intelligence (SSCI '13)*, pp. 120-127, Apr. 2013.

[10] S. Mallik, A. Mukhopadhyay, and U. Maulik, "Integrated Statistical and Rule-Mining Techniques for DNA Methylation and Gene Expression Data Analysis," *J. Artificial Intelligence and Soft Computing Research*, vol. 3, no. 2, pp. 1-17, 2013.

[11] J. Hacia, J. Fan, O. Ryder, L. Jin, K. Edgemon, G. Ghandour, R. Mayer, B. Sun, L. Hsie, C. Robbins, L. Brody, D. Wang, E. Lander, R. Lipshutz, S. Fodor, and F. Collins, "Determination of Ancestral Alleles for Human Single-Nucleotide Polymorphisms Using High-Density Oligonucleotide Arrays," *Nature Genetics*, vol. 22, pp. 164-167, 1999.

[12] W. Alvord, J. Roayaei, O. Quinones, and K. Schneider, "A Microarray Analysis for Differential Gene Expression in the Soybean Genome Using Bioconductor and R," *Briefings in Bioinformatics*, vol. 8, no. 6, pp. 415-431, 2007.

[13] W. Pan, "A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments," *Bioinformatics*, vol. 12, pp. 546-554, 2002.

[14] A. Vickers, "Parametric versus Non-Parametric Statistics in the Analysis of Randomized Trials with Non-Normally Distributed Data," *BMC Medical Research Methodology*, vol. 5, article 35, pp. 1-12, 2005.

[15] S.Y. Kim, J.W. Lee, and I.S. Sohn, "Comparison of Various Statistical Methods for Identifying Differential Gene Expression in Replicated Microarray Data," *Statistical Methods in Medical Research*, vol. 15, pp. 3-20, 2006.

[16] J. Sreekumar, K.K. Jose, "Statistical Tests for Identification of Differentially Expressed Genes in cDNAMicroarray Experiments," *Indian J. Biotechnology*, vol. 7, pp. 423-436, 2008.

[17] C. Murie, O. Woody, A. Lee, and R. Nadon, "Comparison of Small n Statistical Tests of Differential Expression Applied to Microarrays," *BMC Bioinformatics*, vol. 10, article 45, pp. 1-18, 2009.

[18] X. Cui, J. Hwang, J. Qiu, N. Blades, and G. Churchill, "Improved Statistical Tests for Differential Gene Expression by Shrinking Variance Components Estimates," *Biostatistics*, vol. 6, no. 1, pp. 59-75, 2005.
[19] G. Smyth, "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 3, 2004.
[20] O. ElBakry, M. Ahmad, and M. Swamy, "Identification of Differentially Expressed Genes for Time-Course Microarray Data Based on Modified RM ANOVA," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 9, no. 2, pp. 451-466, Mar./Apr. 2012.
[21] U. Maulik and A. Mukhopadhyay, "Towards Improving Fuzzy Clustering Using Support Vector Machine: Application to Gene Expression Data," *Pattern Recognition*, vol. 42, no. 11, pp. 2744-2763, 2009.
[22] J. Lu, G. Get, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando, J.R. Downing, T. Jacks, H.R. Horvitz, and T.R. Golub, "MicroRNA Expression Profiles Classify Human Cancers," *Nature*, vol. 435, pp. 834-838, 2005.
[23] R. Bourgon, R. Gentleman, and W. Huber, "Independent Filtering Increases Detection Power for High-Throughput Experiments," *Proc. Nat'l Academy Sciences USA*, vol. 107, no. 21, pp. 9546-9551, 2010.
[24] T. Jayalakshmi and A. Santhakumaran, "Statistical Normalization and Back Propagation for Classification," *Int'l J. Computer Theory and Eng.*, vol. 3, no. 1, pp. 1793-8201, 2011.
[25] B.M. Bolstad, R.A. Irizarry, M. Astrand, and T. Speed, "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias," *Bioinformatics*, vol. 19, no. 2, pp. 185-193, 2003.
[26] W. Huber, A.V. Heydebreck, H. Sultmann, A. Poustka, and M. Vingron, "Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression," *Bioinformatics*, vol. 18, no. suppl. 1, pp. S96-S104, 2002.
[27] T. Thadewald and H. Buning, "Jarque-Bera Test and Its Competitors for Testing Normality," *J. Applied Statistics*, vol. 34, no. 1, pp. 87-105, 2007.
[28] N. Razali and Y. Wah, "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests," *J. Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21-33, 2011.
[29] V.G. Tusher, R. Tibshirani, and G. Chu, "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proc. Nat'l Academy Sciences USA*, vol. 98, pp. 5116-5121, 2001.
[30] S.E. Choe, M. Boutros, A.M. Michelson, G.M. Church, and M.S. Halfon, "Preferred Analysis Methods for Affymetrix Genechips Revealed by a Wholly Defined Control Dataset," *Genome Biology*, vol. 6, no. 2, article R16, pp. 1-16, 2005.
[31] Y. Chen and H. Hu, "An Overlapping Cluster Algorithm to Provide Non-Exhaustive Clustering," *European J. Operational Research*, vol. 173, pp. 762-780, 2006.
[32] B. Welch, "The Significance of the Difference between Two Means when the Population Variances Are Unequal," *Biometrika*, vol. 29, pp. 350-362, 1938.
[33] A. McCluskey and A. Lalkhen, "Statistics IV: Interpreting the Results of Statistical Tests," *Continuing Education in Anaesthesia, Critical Care & Pain*, vol. 7, no. 6, pp. 208-212, 2007.
[34] K. Rothman, "Curbing Type I and Type II Errors," *European J. Epidemiology*, vol. 25, no. 4, pp. 223-224, 2010.
[35] J. Aldrich, "Correlations Genuine and Spurious in Pearson and Yule," *Statistical Science*, vol. 10, no. 4, pp. 364-376, 1995.
[36] M. Anderson, "Permutation Tests for Univariate or Multivariate Analysis of Variance and Regression," *Canadian J. Fisheries and Aquatic Science*, vol. 58, pp. 626-639, 2001.
[37] Wilcoxon, Frank, "Individual Comparisons by Ranking Methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80-83, 1945.
[38] R. Walpole and R. Myers, *Probability and Statistics for Engineers and Scientists*, vol. 5, *Macmillan, 1993.*
[39] O.G. Troyanskaya, M.E. Garber, P.O. Brown, D. Botstein, and R.B. Altman, "Nonparametric Methods for Identifying Differentially Expressed Genes in Microarray Data," *Bioinformatics*, vol. 18, pp. 1454-1461, 2002.
[40] I. Lonnstedt and T. Speed, "Replicated Microarray Data," *Statistical Sinica*, vol. 12, pp. 31-46, 2002.
[41] B. Efron, "Size, Power, and False Discovery Rates," *The Annals of Statistics*, vol. 35, no. 4, pp. 1351-1377, 2006.
[42] R.O. Rhein and K. Strimmer, "Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, article 9, 2007.
[43] Kruskal, Wallis, "Use of Ranks in One-Criterion Variance Analysis," *J. Am. Statistical Assoc.*, vol. 47, no. 260, pp. 583-621, 1952.
[44] A. Kolmogorov, "Sulla Determinazione Empirica di Una Legge di Distribuzione," *Giomale dell' Instituto Italiano degli attuari*, vol. 4, pp. 83-91, 1933.
[45] Y. Pawitan, S. Michiels, S. Koscielny, A. Gusnanto, and A. Ploner, "False Discovery Rate, Sensitivity and Sample Size for Microarray Studies," *Bioinformatics*, vol. 21, pp. 3017-3024, 2005.
[46] S. Dudoit, J. Shaffer, and J. Boldrick, "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, vol. 18, no. 1, pp. 71-103, 2003.
[47] J. Shaffer, "Multiple Hypothesis Testing: A Review," *Ann. Rev. Psychology*, vol. 46, pp. 561-584, 1995.
[48] S. Lise, C. Archambeau, M. Pontil, and D. Jones, "Prediction of Hot Spot Residues at Protein-Protein Interfaces by Combining Machine Learning and Energy-Based Methods," *BMC Bioinformatics*, vol. 10, article 365, 2009.
[49] E. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilita," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 3-62, 1936.
[50] S. Holm, "A Simple Sequential Rejective Multiple Test Procedure," *Scandinavian J. Statistics*, vol. 6, pp. 65-70, 1979.
[51] P. Westfall and S. Young, "Resampling-Based Multiple Testing," . John Wiley & Sons, 1993.
[52] Y. Benjamini and Y. Hochberg, "On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics," *J. Educational and Behavioral Statistics*, vol. 25, pp. 60-83, 2000.
[53] J. Storey, "A Direct Approach to False Discovery Rates," *J. Royal Statistical Soc. Series, B*, vol. 64, pp. 479-498, 2002.
[54] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. the Royal Statistical Soc., Series B*, vol. 85, pp. 289-300, 1995.
[55] B. Efron, "Robbins, Empirical Bayes, and Microarrays," *The Annals of Statistics*, vol. 31, no. 2, pp. 366-378, 2003.
[56] B. Efron, R. Tibshirani, J. Strey, and V. Tusher, "Empirical Bayes Analysis of a Microarray Experiment," *J. the Am. Statistical Assoc.*, vol. 96, pp. 1151-1160, 2001.
[57] B. Efron, "Microarrays, Empirical Bayes and the Two-Groups Model," *Statistical Science*, vol. 23, pp. 1-22, 2008.
[58] Y. Benjamini and D. Yekutieli, "The Control of the False Discovery Rate in Multiple Testing under Dependency," *The Annals of Statistics*, vol. 29, pp. 1165-1188, 2001.
[59] J. Leek and J. Storey, "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis," *PLoS Genetics*, vol. 3, no. 9, pp. 1724-1735, 2007.
[60] J. Storey and T.R., "Statistical Significance for Genome-Wide Studies," *Proc. Nat'l Academy of Sciences USA*, vol. 100, no. 16, pp. 9440-9445, 2003.
[61] Y. Sun, N. Zhang, and A. Owen, "Multiple Hypothesis Testing Adjusted for Latent Variables, with an Application to the Agemap Gene Expression Data," *J. Applied Statistics*, vol. 6, no. 4, pp. 1664-1688, 2013.
[62] J. Leek and J. Storey, "A General Framework for Multiple Testing Dependence," *Proc. Nat'l Academy of Sciences USA*, vol. 105, no. 48, pp. 18718-18723, 2008.
[63] J. Leek, R. Scharpf, H. Bravo, D. Simcha, B. Langmead, W. Johnson, D. Geman, K. Baggerly, and R.A. Irizarry, "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data," *Nature Rev. Genetics*, vol. 11, pp. 733-739, 2010.
[64] A. Scherer, *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. John Wiley & Sons, 2009.
[65] J. Lucas, C. Carvalho, Q. Wang, A. Bild, J. Nevins, and M. West, "Sparse Statistical Modelling in Gene Expression Genomics," *Bayesian Inference for Gene Expression and Proteomics*, pp. 155-176, Cambridge Univ. Press, 2006.
[66] J. Gagnon Bartsch and T. Speed, "Using Control Genes to Correct for Unwanted Variation in Microarray Data," *Biostatistics*, vol. 13, no. 3, pp. 539-552, 2012.

[67] H. Luo, "Generation of Non-Normal Data a Study of Fleishmans Power Method," *Working Paper, Dept. of Statistics Uppsala Univ.*, Mar. 2011.
[68] W. Reinartz, R. Echambadi, A. Lee, and W. Chin, "Generating Non-Normal Data for Simulation of Structural Equation Models Using Mattsons Method," *Multivariate Behavioral Research*, vol. 37, no. 2, pp. 227-244, 2002.
[69] R. Simon, E. Korn, and L. McSchane, *Design and Analysis of DNA Microarray Investigations*, pp. 75-84, Springer, 2003.
[70] J. Zou, S. Zas, and M. Aldea, "Expression Profiling Soybean Response to Pseudomonal Syringae Reveals New Defense-Related Genes and Rapid HR-Specific Downregulation of Photosynthesis," *Molecular Plant-Microbe Interactions*, vol. 18, pp. 1161-1174, 2005.
[71] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591-611, 1965.
[72] B. Hertogh, B. Meulder, F. Berger, M. Pierre, E. Bareke, A. Gaigneaux, and E. Depiereux, "A Benchmark for Statistical Microarray Data Analysis that Preserves Actual Biological and Technical Variance," *BMC Bioinformatics*, vol. 11, article 17, pp. 1-14, 2010.
[73] S. Dudoit, Y. Yang, T. Speed, and M. Callow, "Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments," *Statistica Sinica*, vol. 12, pp. 111-139, 2002.
[74] H. Bar, J. Booth, E. Schifano, and M. Wells, "Laplace Approximated EM Microarray Analysis: An Empirical Bayes Approach for Comparative Microarray Experiments," *Statistical Science*, vol. 25, no. 3, pp. 388-407, 2010.
[75] P. Delmar, S. Robin, and J. Daudin, "Varmixt: Efficient Variance Modelling for the Differential Analysis of Replicated Gene Expression Data," *Bioinformatics*, vol. 21, no. 4, pp. 502-508, 2005.
[76] G. Wright and R. Simon, "A Random Variance Model for Detection of Differential Gene Expression in Small Microarray Experiments," *Bioinformatics*, vol. 19, no. 18, pp. 2448-2455, 2003.
[77] F. Jaffrezic, G. Marot, S. Degrelle, I. Hue, and J. Foulley, "A Structural Mixed Model for Variances in Differential Gene Expression Studies," *Genetics Research*, vol. 89, pp. 19-25, 2007.
[78] A. Sjogren, E. Kristiansson, M. Rudemo, and O. Nerman, "Weighted Analysis of General Microarray Experiments," *BMC Bioinformatics*, vol. 8, pp. 387-401, 2007.
[79] M. Jeanmougin, A. Reynies, L. Marisa, C. Paccard, G. Nuel, and M. Guedj, "Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies," *PLoS One*, vol. 5, no. 9, p. e12336, 2010.
[80] S. Bandyopadhyay and M. Bhattacharyya, "A Biologically Inspired Measure for Coexpression Analysis," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 8, no. 4, pp. 929-942, July/Aug. 2011.

Sanghamitra Bandyopadhyay received the PhD degree in computer science in 1998 from Indian Statistical Institute, Kolkata, India, where she currently serves as a professor. She was in Los Alamos National Laboratory, University of New South Wales, Australia, University of Texas at Arlington, University of Maryland, Baltimore County, Fraunhofer Institute AiS, Germany, Tsinghua University, China, University of Rome, Italy and MPI, Saarbruken, Germany. She has received the prestigious S.S. Bhatnagar Award in 2010. She has also received Humboldt fellowship for experienced researchers. She has coauthored five books and more than 200 research papers. Her research interests include pattern recognition, data mining, soft computing and bioinformatics. She is a senior member of the IEEE.

Saurav Mallik received the BTech and MTech degrees in computer science and engineering from West Bengal University of Technology, India in 2009 and 2011, respectively. He is currently a junior research fellow in a DST-sponsored project at the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. His research interests include bioinformatics, data mining, computational biology, gene expression, and methylation analysis.

Anirban Mukhopadhyay received the PhD degree in computer science from Jadavpur University, Kolkata, India, in 2009. He is currently an associate professor with the Department of Computer Science and Engineering, University of Kalyani, India. He received the University Gold Medal from Jadavpur University in 2004. He was with the German Cancer Research Center, Heidelberg, Germany, University of Nice Sophia-Antipolis, France, and the University of Goettingen, Germany. He has coauthored one book and about 90 research papers. His research interests include soft and evolutionary computing, data mining, machine learning and bioinformatics. He is a senior member of the IEEE.