Deposited research article

A tool for comparing different statistical methods on identifying differentially expressed genes

Paul Fogel[1*], Li Liu[2*], Bruno Dumas[3], Nanxiang Ge[2]

Addresses: [1]Paul Fogel Consultant, 4 rue Le Goff, 75005 Paris, France. [2]Biometrics and Data Management, Sanofi-Aventis, Mail Stop B-203A, PO Box 6800, 1041 Route 202-206, Bridgewater, NJ 08873, USA. [3]Yeast Genomics, Functional Genomics, Sanofi-Aventis,13 Quai Jules Guesde, 94403 Vitry sur Seine Cedex, France. *These authors contributed equally to this work.

Correspondence: Li Liu. E-mail: Li.Liu@aventis.com

Posted: 8 December 2004

Genome ***Biology*** 2004, **6:**P2

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/6/1/P2

Received: 7 December 2004

This is the first version of this article to be made available publicly.

A tool for comparing different statistical methods on identifying differentially expressed genes

Paul Fogel [1]*, Li Liu [2]*§, Bruno Dumas [3], Nanxiang Ge[2]

[1]Paul Fogel Consultant, 4 rue Le Goff, 75005 Paris, France
[2]Biometrics and Data Management, Sanofi-Aventis, Mail Stop B-203A, PO Box 6800, 1041 Route 202-206, Bridgewater, NJ 08873, USA
[3]Yeast Genomics, Functional Genomics, Sanofi-Aventis,13 Quai Jules Guesde, 94403 Vitry sur Seine Cedex, France

*These authors contributed equally to this work
§Corresponding author

Email addresses:
Paul.Fogel@wanadoo.fr
Li.Liu@aventis.com
Bruno.Dumas@aventis.com
Nanxiang.Ge@aventis.com

Abstract

Background

Many different statistical methods have been developed to deal with two group comparison microarray experiments. Most often, a substantial number of genes may be selected or not, depending on which method was actually used. Practical guidance on the application of these methods is therefore required. We developed a procedure based on bootstrap and a criterion to allow viewing and quantifying differences between method-dependent selections. We applied this procedure on three datasets that cover a range of possible sample sizes to compare three well known methods, namely: t-test, LPE and SAM.

Results

Our visualization method and associated *variability conformation rate* (VCR) criterion show that standard t-test is appropriate for large sample sizes to allow accurate variance estimates. LPE borrows strength from neighboring genes to estimate the variances and is therefore more appropriate for small sample sizes whenever gene variances are similar for similar gene intensity levels. SAM has both advantages of considering gene specific variance like t-test and adjusting multiple tests by permutation based false discovery rate. However, for small sample sizes and in cases of numerous expressed genes, the distribution based on permutated datasets may not approximate the null distribution well, resulting in an inaccurate false discovery rate. Moreover, genes with low variances may be filtered because of the fudge factor.

Conclusion

We proposed using VCR to assess different statistical methods available for analyzing microarray data and developed a bootstrap method - on which our criterion is based - to estimate the 2-d distribution of treated vs. control gene intensity levels, under the null hypothesis that there is no difference between the treatment and control group. The biological evaluation of selected genes according to one or another method confirmed that this criterion is indeed appropriate to help identifying the most suitable method.

Background

Microarray technology has become a widely used tool in drug discovery and is becoming a powerful tool in drug development. One of the most widely used statistical designs in microarray experiment is two-group comparison: disease tissue versus normal, drug treated versus non-treated, etc. Associated with the large amount of data generated with microarray experiments, there are now many published statistical methods for analyzing such experiments, e.g. standard two sample t-test, SAM [12], LPE [7], GEA [8], PFOLD [10]. Accompanying such an array of methods available to practitioners are the questions: when to use which methods? What are the pros and cons for different methods? Is there any consistency between different methods? We illustrate the issue using the following example.

In a study of the relationship between the activation status of the adoptively transferred T-cells and the migration and retention process of the CD8+ T-cells in the

lungs (see below in the Results section), 2677 genes were selected through the SAM test using a 5% False Discovery Rate. We applied two other methods, t-test and LPE to the same dataset and selected the first 2677 genes with respect to the P-values given by the respective methods. A simple Venn diagram suggests the dramatic difference one might get when applying these three different methods (Figure 1). The difference is even more striking when the selected genes are represented on a scatterplot of averaged treated vs. control expression levels (Figure 2). Given such dramatic difference in the gene list generated, it is important to provide a criterion to help deciding when to use which method.

In this paper, by examining the results of applying these three commonly used methods to three representative data sets, we aim to provide practical guidance on their application. To achieve this, we developed a visualization method based on bootstrap allowing one to view the difference with respect to the genes identified by different methods.

Comparison criterion

Under the null hypothesis that there is no difference between the treatment and control group, let's first assume that a 2-d null distribution of treated vs. control gene intensity levels can be estimated (details of the estimation will be given later in the Methods). Contours of the 2-d null distribution can then be added to the scatterplot of Figure 2, at various alpha-levels (Figure 3). As will be later confirmed from a biological point of view (Discussion), selected points that fall beyond the outer contour, which have a very low probability density under the null distribution, correspond to genes that are most likely truly regulated. On the contrary, selected points that fall within the inner contour correspond to genes that are unlikely regulated, i.e. false positives. Thus, it is possible to compare different selection methods using the number of points that fall beyond the outer contour of the 2-d null distribution, the best selection being the one which yields the highest number of such points.

To facilitate the comparison of the methods, we define the *variability conformation rate* (VCR) for each method *m* at a given false discover rate α of SAM and a given contour height h:

$$VCR(m \mid \alpha, h) = \frac{K_{m,\alpha,h}}{K_\alpha}$$

with K_α being the total number of genes identified by SAM as having FDR less than α and $K_{m,\alpha,h}$ being the total number of genes out of the top K_α genes for method *m* lying outside of the contour of 2-d null distribution with height *h*. VCR provides us with a quantitative metric to evaluate the methods. In the above example, among the 2677 selected genes, VCR for LPE, T-test and SAM are 77%, 62% and 60%, respectively, suggesting that the LPE method is in this particular case performing best.

Results

In this section, we illustrate our method and describe the comparison results using three real examples.

Examples

Yeast

In parallel experiments, CA10/pCD63 (an acetyl pregnenolone producing strain) and Fy 1679-28c (an non producing strain) were submitted to a fermentation process. The process classically comprises three phases: batch phase, fed batch phase and stationary phase. CA10/pCD63 is described in Duport *et. al*. [3]. Fy 1679-28c is described in Thierry *et al*. [11]. The transcription profiles in stationary phase (the production phase) were compared using Affymetrix technology with two duplicated points at the beginning and the end of the stationary phase. The data obtained from the Affymetrix software MAS 4.0 were transferred to the Gecko software [10] with minor modification. The marginally present or absent calls were replaced by present or absent calls respectively.

T-cell Immune Responses Microarray Study

In this study, Hafezi-Moghadam and Ley [4] studied the relationship between the activation status of the adoptively transferred T-cells and the migration and retention process of the CD8+ T-cells in the lungs. Affymetrix murine chip, MG-U74vA, was used to study the three groups of immune exposure: naïve (no exposure), 48h activated, and CD8+ T-cell clone D4 (long term mild exposure). Each group has three replicates. Signal intensity values were obtained from MAS 5.0. In this paper, we compare two groups, naïve and 48h activated.

Breast Cancer Study

Huang *et. al.* [5] investigated the association between the lymph node metastasis, cancer recurrence and gene expression data. We used a subset of patients with one to three positive lymph nodes and studied the recurrence three years after primary surgery. The data set provided expression profiles for 52 cases in this lymph node category (34 non-recurrent, 18 recurrent). We identified the differentially expressedgenes between recurrent and non-recurrent patients.

Generation of a 2-d null distribution: Bootstrap results

(See Methods for details on the Bootstrap procedure)

In the Yeast and T-cell Immune Responses studies, for which the number of replicates is low (=3), we used a bin size of 10 to allow resampling within a reasonably large sample (20^3=8000)

On the contrary, in the Breast Cancer study, it was possible to use the smallest possible bin size (2) thanks to the very large number of replicates, which allowed resampling within a sample of size 4^34.

The Breast Cancer study was also used for validation purpose; Bootstrapped controls based on 17 real controls selected randomly played the role of a learning dataset to calculate the contours of the 2-d null distribution of the average of 17 controls vs. the average of 17 other controls. These contours were further drawn on the plot of averaged real controls that were left out of the learning dataset vs. the averaged real controls that were used to generate the bootstrapped ones. This comparison clearly shows that both distributions almost perfectly overlap (Figure 4).

Generating differential analysis results and comparing difference

We applied three methods, t-test, LPE, SAM to the three datasets to identify differentially expressedgenes. For t-test and LPE, the log2-transformed expression intensities were used. For SAM, both the log2-transformed expression intensities and the untransformed data were used to study the difference.

To make all the tests comparable, for a given false discovery rate, we first counted the number of expressedgenes based on SAM for transformed data. Then we selected the same number of expressedgenes from other tests based on their p-values.

For the T-cell immune responses microarray study, given a false discovery rate of 5%, 2677 genes were selected by SAM. At the same time, we selected the first 2677 most significant genes from t-test, LPE based on the p-values. The identified genes from different methods are plotted in Figure 5. Larger version of Figure 5 can be found in the additional files (additional figures 1-4). As we can see, the genes identified by LPE followed the variability plot very well; Genes identified by SAM fell outside two 45 degree parallel lines; Genes identified by t-test and SAM with raw data were more similar, and followed the variability plot less well than LPE. Table 1 summarized the number of points outside of the estimated contour of the 2-d null distribution at various alpha levels.

Overall, the percentage of identified genes outside the contour is higher based on LPE. As the density level of the contour get bigger, for example, 0.1, the percentage of genes outside the contour from different methods get closer. Similar conclusions can be drawn from the yeast data (Table 2) and the breast cancer data (Table 3). Additional Table 1 gives the number and percentage of overlapped genes identified by t-test, LPE, SAM, and SAM using untransformed intensities for the yeast data, which also suggests that SAM using raw data and t-test are more similar than LPE.

Summary of results

We compared t-test, LPE, SAM using the proposed visualization tool based on bootstrap, and the results from three datasets illustrated the difference of the genes identified by each method.

Tables 1-3 summarize the VCR for all the three different methods on three different data sets. One consistent trend is that the LPE tends to have larger VCR measures than the other two methods.

We summarized the advantages and disadvantages of each method in Table 4, and provided practical suggestions.

Standard t-test considers gene specific variance, and it is a good choice if the sample size is large. However, if the sample size is small, the variance estimate may be inaccurate. T-test does not perform the multiple test adjustment.

LPE borrows strength from neighboring genes to estimate the variances, and it is a good choice if the sample size is small and the gene variances are similar for similar gene intensity levels. However, if we know that there are quite a number of genes with gene-specific variances, this method is not a good choice. LPE does not perform multiple test adjustment.

SAM considers gene specific variance, and adjusts the multiple tests by permutation based false discovery rate. However, if the sample size is small and there are many expressedgenes, the distribution based on permutated datasets may not approximate

the null distribution well, and thus the permutation based false discovery rate may be inaccurate. SAM filtered some genes with low variances because of the fudge factor.

Discussion

The three datasets used in this study cover a range of possible sample sizes: three replicates in each group in Ley's data set; eight samples in the yeast data set and more than twenty samples from the breast cancer data set. Such a variety of sample sizes, along with the VCR criterion, allowed us a comprehensive evaluation of the methods being considered. However, we need also to consider this evaluation from a biological perspective, i.e. determine whether genes lying outside of the contour of a 2-d null distribution are indeed the most relevant ones. To do this, we looked more specifically at the yeast example and compared selected genes according to one or another method in terms of biological relevance, to see whether the same conclusion was reached than while using the VCR criterion.

The transcription profiles of two different strains were compared: wild type strain Fy 1679-28c and the production strain CA10/pCD63, which is a recombinant strain. CA10/pCD63 was selected for its ability to produce steroids and to grow on glucose instead of galactose and its capacity of deregulating the promoters that drive the recombinant protein coding sequences. Genetically, *URA3*, *TRP1* and *LEU2* genes are present in the production strain while absent in the wild type strain and *ERG5* gene is present in the wild type but has been disrupted in the production strain. Phenotypically, the CA10/pCD63 strain differs by the deregulation of the galactose biosynthesis (*GAL* and *GCY1*, genes) pathway. Moreover, it is expected that the *ERG* genes be deregulated in order to compensate for the steroid excretion. In summary, at minimum the two transcription programs should differ in galactose metabolism and possibly in sterol biosynthesis and steroid detoxification.

We first checked that obvious differences corresponding to known genetic modifications were found. The three methods indicate that *LEU2*, *URA3* and *TRP1* transcripts were clearly induced in the production strain while *ERG5* transcript was absent in this same strain, as expected. Furthermore, all methods clearly point out that the two strains differ dramatically by their expression profile - with up to 1/6 of the genes of the genome having different expression level – and allow for detecting profound changes in the galactose (comprising the *GCY1* co regulated gene) biosynthesis pathway, in agreement with the biological selection process; The genes (*GAL1*, *GAL2*, *GAL10*) coding for enzymatic activities are deregulated between 24 to 50 times while the genes coding for transcription factors such as *GAL80* and *GAL3* are deregulated 3 to 6 times. This corresponds to a partial deregulation of the pathway, as induction with galactose is known to bring up to 500-fold induction of the *GAL1* promoter [6].

Since part of the ergosterol synthesis is routed to excrete steroids, *ERG* genes transcription might be modified or even up regulated during the production phase. Apart from the *ERG5* control gene, three other genes of the family namely *ERG1*, *ERG6* and *ERG24* are detected showing a two-fold induction with LPE and t-test for *ERG6* and with LPE and SAM test for *ERG1* and *ERG24*. *CYB5* electron carrier gene transcript is detected by all three methods while LPE and SAM detect the *NCP1* induction. It has been shown [1,13] that during azole treatment (targeting the ergosterol biosynthesis), which is mimicking our steroid excretion, these five genes

(*ERG1*, *ERG6*, *ERG24*, *CYB5* and *NCP1*) can be induced among other genes of the *ERG* family. It is apparent here that LPE is the only method that can discriminate the subtle changes of all five genes. On the contrary, t-test is clearly not performing well, as it detects only two out of these five genes. In this respect, SAM appears much closer to LPE (four detected genes out of five).

In order to further assess the selection power of LPE as compared to SAM, we selected a set of 22 genes that were found up regulated by LPE but not SAM (*ERG6*, *THI11*, *FAA2*, *MSK1*, *TIF35*, *RPL33B*, *YBR090C*, *RPL8B*, *TNA1*, *SSA3*, *RPL12B*, *SNF1*, *GTT1*, *YKL151C*, *YER044C*, *RPS11B*, *NCP1*, *RPL21A*, *YGR043C*, *RPL17A*, *RPS3*, *SMC2*). We used the "Micro Array Global Viewer" (www.transcriptome.ens.fr/ymgv/) [1] to see whether any of these 22 genes could match an already described transcription profile in the database consisting of 1347 yeast dataset conditions. In addition, a randomly selected set of 22 genes was used as a control to insure the specificity of the comparison with the database. Two conditions showed the same set of up regulated genes. One condition found with both the randomly selected set of genes and the LPE specific set of genes was discarded. It corresponds to a non-specific induction of a large spectrum of genes by an antifungal compound of unknown mechanism of action [9]. The second condition corresponds to 17 out of the 22 genes that are induced by 0.4M NaCl stress in a *HOG1* independent fashion. This could point out the fact that yeast strains are submitted to a high osmolarity in fermentors due to the continuous base feeding in order to maintain a neutral pH. It indicates that the production strain shows a small but significant induction of a *HOG1* independent pathway.

The same kind of experiment was also performed with LPE specific and down regulated genes namely: *QCR8*, *ACO1*, *MDH1*, *INH1*, *COX8*, *CAR1*, *YMR265C*, *SDH1*, *DDR48*, *CPA2*, *ICY2*, *COX9*, *TPO1*, *COX6*, *CYT1*, *ACS2*, *ILV3*, *FUM1*, *IDH2*, *ORT1*, *OAC1*, *CWP1*. Among the 1347 transcription profiles, a few conditions were matching the down regulation of this set of genes. Interestingly, two temperature sensitive mutants corresponding to cell cycle arrested cells, namely cdc15 and cdc24, matched the above set of genes. It is not clear why the production strain should be more arrested in its cycle than the control strain. Both strains are arrested in their cell cycle since they are in stationary phase. Finally, a majority of genes (13 out of 22) of this LPE specific and down regulated list localized to mitochondria. Interestingly, five of the encoded proteins namely: *ACO1* (Aconitate hydratase), *IDH2* (Isocitrate dehydrogenase), *MDH1* (Malate dehydrogenase), *SDH1* (Succinate dehydrogenase), *FUM1* (Fumarate hydratase) can be clearly co-regulated as they belong to the tricarboxylic acid cycle (Krebs cycle) [2]. Thus, the LPE method points out a down regulation of the transcription of the genes involved in this cycle. This regulation should slow down the production of the corresponding enzymes and acetylCoA consumption in the cycle, thus improving acetylCoA availability for sterol biosynthesis. It is worth noting that the *ACS2* (acetylCoA synthase) gene appears also down regulated. Most of the other half of the genes are involved in electron transport machinery i.e. *QCR8*, *COX6*, *COX8*, *COX9*. All in all, the LPE method appears to specifically pick up genes that are in the same pathways.

Conclusions

In this paper, we tackled a very practical problem: how to understand the different statistical methods available for analyzing microarray data and how they differentiate

in terms of performance. We proposed a criterion (VCR) to assess different statistical methods and developed a bootstrap method to estimate the null distribution of treated vs. control gene intensity levels on which our criterion is based. Finally, the biological evaluation of selected genes according to one or another method strengthened our first conclusion - drawn from a pure statistical point of view - that the LPE method is a better choice when the sample size is small. This suggests that VCR is indeed an appropriate criterion to assess different methods.

Methods

Generation of a 2-d null distribution: Bootstrap procedure

The 2-d null distribution can be estimated using 2-d non-parametric distribution of one averaged subset of controls vs. another averaged subset of controls, each subset being of the size of the treated set. This is possible whenever the experimental design contains twice as many controls as treated conditions. However, most experimental designs tend to be balanced. We therefore present a simple bootstrap procedure that allows creating as many "virtual" controls as needed, in order to obtain a non parametrical 2-d null distribution. We will see that this procedure guarantees that the 2-d null distribution is similar to the one that would be achieved with real controls.

For the sake of simplicity, we consider the case of duplicates controls (the general case is described below in Theoretical grounds). Let (X,Y) be duplicate expression log intensities of a particular gene. Assume $X = \mu + \varepsilon_X$ and $Y = \mu + \varepsilon_Y$ where μ follows the probability distribution $\mathrm{g}(\mu)$ and $(\varepsilon_X, \varepsilon_Y)$ is a couple of independent error terms that follow the probability distribution $\mathrm{h}(\varepsilon)$. $\mathrm{g}(\mu)$ is associated with gene diversity within the chip, different genes being possibly expressed at different levels. $\mathrm{h}(\varepsilon)$ is associated with experimental variability. We assume normal distributions for g and h:

$$\mathrm{g}(\mu) = \sigma_a f_N\left(\frac{\mu - \mu_0}{\sigma_a}\right)$$

$$\mathrm{h}(\varepsilon) = \sigma_e f_N(\varepsilon / \sigma_e)$$

where $f_N(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$

The bootstrap procedure is based on the following main result (see proof below in Theoretical ground):

Define $X_z \equiv X$ given $\frac{X+Y}{2} = z$

- The expectation of X_z is z
- The variance of X_z is $\sigma_e{}^2 / 2$

Procedure:

1. Rank genes with respect to the average of the duplicate $Z = \frac{X+Y}{2}$
2. Bin ranked genes into bins of size *s*
    a. The size is chosen small enough to ensure that within each bin the average can be considered as constant: $Z \approx z$ (see the results section for a discussion on the size *s*).

b. It seems reasonable to assume that within each bin, the real expression levels μ_i are close enough to ensure that the error terms $(\varepsilon_{Xi}, \varepsilon_{Yi})$ have the same distribution (see the results section for a validation of this assumption on real data)

3. Let (x_i, y_i) and (x_j, y_j) be duplicate observations of two genes within the same bin. Since $z_i \approx z_j$, the four conditional variables $X_{zi}, X_{zj}, Y_{zi}, Y_{zj}$ follow the same distribution. Thus, given $Z = z$, all x_i's and y_i's have expectation z and variance $\sigma_e^2 / 2$. New x and y values with expectation z and variance σ_e^2, noted x'_i and y'_i, can be obtained by:
   a. Re-sampling the x_i's and y_i's with replacement
   b. Applying the variable transformation $\begin{cases} X_z' = z + \sqrt{2}(X_z - z) \\ Y_z' = z + \sqrt{2}(Y_z - z) \end{cases}$
4. The same process is repeated for each bin

Remark 1:
For a particular gene with expression level μ_0, X_z' and Y_z' are biased with bias $z - \mu_0$. However, the 2-d null distribution formed by the (X_z', Y_z')'s is still similar to the original one formed by the (X, Y)'s, as (X_z', Y_z') is unbiased for those particular genes with expression level $\mu \approx z$ (such genes exist most likely given the high number of genes).

Remark 2:
Consider $2K$ controls from which we can form arbitrarily K different $(X_k, Y_k), 1 \le k \le K$. Thanks to the bootstrap process, any (X_{k0}, Y_{k0}) will allow creating K bootstrapped (X_k', Y_k'). However:

- The average of virtual pairs $\frac{1}{K}\sum_{1 \le k \le K}(X_k', Y_k')$ will tend to (z, z), where z is the value taken by $Z = \frac{X+Y}{2}$ on the original pair used to generate the virtual ones.
- The average of real pairs $\frac{1}{K}\sum_{1 \le k \le K}(X_k, Y_k)$ will tend to (μ, μ), where μ is the expression level of the gene under consideration.

In other words, while the bootstrap process allows finding the 2-d null distribution, it does *not* improve the estimation of the expression level of individual genes.

Generation of a 2-d null distribution: Theoretical grounds

For the sake of simplicity, we will first consider the case of duplicates. The extension to the general case will follow. We will now prove the main results:

Define $X_z \equiv X$ given $\frac{X+Y}{2} = z$

1. The expectation of X_z is z
2. The variance of X_z is $\sigma_e^2 / 2$

Demonstration:

Let $Z \equiv \frac{X+Y}{2}$, $T \equiv \frac{Y-X}{2}$. Then:

- $X = Z - T$ and $Y = Z + T$
- Due to the normality assumption, Z and T, which are orthogonal, are independent
- $Z = \mu + \varepsilon_Z$ and $T = \varepsilon_T$ where μ follows the probability distribution $\mathrm{g}(\mu)$ and $(\varepsilon_Z, \varepsilon_T)$ is a pair of independent error terms that follow the probability distribution $\mathrm{h}'(\varepsilon)$ with $\mathrm{h}'(\varepsilon) = \frac{\sigma_e}{\sqrt{2}} f_N\left(\sqrt{2}\frac{\varepsilon}{\sigma_e}\right)$.

As the former error distribution $\mathrm{h}(\varepsilon)$ will no longer be needed, we will refer to the distribution of T as $\mathrm{h}(\varepsilon)$ instead of $\mathrm{h}'(\varepsilon)$.

Z is the sum of the two random distributions g and h. Thus:

$$\mathrm{f}(X_z = x) = \frac{\mathrm{f}(Z = z, Z - T = x)}{\mathrm{f}(Z = z)} = \frac{\int g(\mu)h(z-\mu)h(z-x)\,d\mu}{\int g(\mu)h(z-\mu)\,d\mu} = h(z-x) \qquad (1)$$

Let us calculate the two first moments of X_z:

$$\mathrm{E}(X_z) = \int xh(z-x)dx = \int (x-z)h(z-y)dx + \int zh(z-x)dx = z$$

$$\mathrm{E}(X_z^{\,2}) = \int x^2 h(z-x)dx$$

$$= \int (x-z)^2 h(z-x)dx + \int z^2 h(z-x)dx + 2\int z(x-z)h(z-x)dx = \sigma_e^{\,2}/2 + z^2$$

The variance follows immediately:

$$\mathrm{var}(X_z) = \left(\sigma_e^{\,2}/2 + z^2\right) - z^2 = \sigma_e^{\,2}/2$$

In exactly the same way, we can define Y_z, which has the same properties as X_z.

Now, considerer the newly transformed variables:

$$\begin{cases} X_z' = z + \sqrt{2}(X_z - z) \\ Y_z' = z + \sqrt{2}(Y_z - z) \end{cases} \qquad (2)$$

These two variables have expectation z and variance $\sigma_e^{\,2}$.

Extension to n-uplates:

Consider now the n-uplate $(X_1, X_2, \cdots, X_n)$. We have the following result:

Define $X_{1,z} \equiv X_1$ given $\frac{X_1 + \cdots + X_n}{n} = z$

1. The expectation of $X_{1,z}$ is z
2. The variance of $X_{1,z}$ is $\frac{n-1}{n}\sigma_e^{\,2}$

Demonstration:

Let $Z = \frac{X_1 + \cdots + X_n}{n}$ and $T = Z - X_1$.

Again, Z and T are orthogonal, thus independent. The only difference with the duplicate case is in the variance of ε_Z and ε_T, since $\mathrm{var}(\varepsilon_Z) = \sigma_e^{\,2}/n$ and $\mathrm{var}(\varepsilon_T) = \frac{n-1}{n}\sigma_e^{\,2}$. The rest of the demonstration is exactly the

same as in the duplicate case, except that we now consider two independent distributions h_Z and h_T for ε_Z and ε_T.

>

The transformation 2 becomes:

$$\begin{cases} X_{1,z}' = z + \sqrt{\dfrac{n}{n-1}}\left(X_{1,z} - z\right) \\ \vdots \\ X_{n,z}' = z + \sqrt{\dfrac{n}{n-1}}\left(X_{n,z} - z\right) \end{cases} \tag{3}$$

Note that the larger *n*, the smaller the effect of the final transformation.

Three methods for identifying *differentially* expressedgenes

In this section, we describe three commonly used methods in analyzing microarray data: Two sample t-test, SAM (Significance Analysis of Microarrays) and LPE (Local Pooled Error) .

T-test is a traditional statistical method for testing the difference between two groups. Suppose we have two groups, treatment group and control group. The microarray intensities in the treatment group are $x_1, x_2, \cdots, x_m$, and the intensities in the control group are $y_1, y_2, \cdots, y_n$. To test whether is any difference between the treatment group and the control group, if we assume equal variances for the two groups, we have

$$T_1 = \frac{\bar{x} - \bar{y}}{\sqrt{s_p^2 (\frac{1}{m} + \frac{1}{n})}}$$

with $(m+n-2)$ degrees of freedom, where $s_p^2 = \dfrac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}$, s_x^2 and s_y^2 are the variances for the treatment group and control group.

If we assume unequal variance, we have

$$T_2 = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{s_x^2}{m} + \dfrac{s_y^2}{n}}}$$

with $df = \dfrac{(s_x^2/m + s_y^2/n)^2}{\dfrac{(s_x^2/m)^2}{m-1} + \dfrac{(s_y^2/n)^2}{n-1}}$.

The t-test works well if the sample size is relatively large. If the sample size is small, the estimated variance may be misleading. Jain *et. al.* [7] proposed a method called LPE to identify differentially expressedgenes, which borrowed strength from neighboring genes to estimate the variability. The LPE variance estimate is based on pooling errors within genes and between replicate arrays for genes in which expression values are similar.

The LPE statistic for the median difference is calculated as :

$$Z = \frac{median(x) - median(y)}{\sigma_{pooled}},$$

where

$$\sigma^2_{pooled} = \frac{\pi}{2}[\sigma^2_x(median(x))/m + \sigma^2_y(median(y))/n],$$

$\sigma^2_x(median(x))$ is the estimate of variance of X from the LPE error distribution at each median log-intensity $median(x)$, and $\sigma^2_y(median(y))$ is the estimate of variance of X from the LPE error distribution at each median log-intensity $median(y)$.

Significance Analysis of Microarrays (SAM) is proposed by Tusher *et. al.* [12], and it assigns a score to each gene based on the changes in gene expression relative to the standard deviation of repeated measurements. Genes with scores greater than a threshold are deemed potentially significant. The percentage of such genes identified by chance is the false discovery rate (FDR), which was estimated by permutation. For the two groups comparison, the score for the i^{th} gene is defined as

$$d_i = \frac{\bar{x} - \bar{y}}{s + s_0},$$

where s is the standard deviation of repeated measurements, which is the same as the denominator of the t-test for comparing two groups assuming equal variances. s_0 is a small positive constant, which was added to ensure that the variance of d_i is independent of gene expression. Thus, we can compare the values of d_i across all genes and compute FDR.

Abbreviations

FDR: false discovery rate
GEA: global error assessment
LPE: local pooled error
SAM: significance analysis of microarrays
VCR: variability conformation rate

Acknowledgements

We thank our colleagues Drs. Steve Binysh and Michel Poncet for their wise comments and suggestions.

References

1. De Backer MD, Ilyina T, Ma XJ, Vandoninck S, Luyten WH, Vanden Bossche H: **Genomic profiling of the response of Candida albicans to itraconazole treatment using a DNA microarray.** *Antimicrob Agents Chemother* 2001, 45: 1660-70
2. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic genetic control of gene expression on a genomic scale.** *Science* 1997, 278: 680-686
3. Duport C, Spagnoli R, Degryse E, Pompon D: **Self-sufficient biosynthesis of pregnenolone and progesterone in engineered yeast.** *Nat Biotechnol.* 1998, 16: 186-9
4. Hafezi-Moghadam A, Ley K: **Relevance of L-selectin shedding for leukocyte rolling in vivo.** *J. Exp. Med.* 1999, 189: 939–947
5. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT: **Gene expression predictors of breast cancer outcomes.** *Lancet* 2003, 361: 1590–96
6. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.** *Science* 2001, 292: 929-34
7. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK: **Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays.** *Bioinformatics* 2003, 19: 1945-1951
8. Mansourian R, Mutch DM, Antille N, Aubert J, Fogel P, Le Goff JM, Moulin J, Petrov A, Rytz A, Voegel JJ, Roberts MA. **The global error assessment (GEA) model for the selection of differentially expressed genes in microarray data.** *Bioinformatics* 2004, doi:10.1093/bioinformatics/bth319
9. Savoie CJ, Aburatani S, Watanabe S, Eguchi Y, Muta S, Imoto S, Miyano S, Kuhara S, Tashiro K: **Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades.** *DNA Res.* 2003, 10: 19-25
10. Theilhaber J, Bushnell S, Jackson A, Fuchs R: **Bayesian estimation of fold-changes in the analysis of gene expression: the PFOLD algorithm.** *J Comput Biol.* 2001, 8: 585-614
11. Thierry A, Fairhead C, Dujon B: **The complete sequence of the 8.2 kb segment left of MAT on chromosome III reveals five ORFs, including a gene for a yeast ribokinase.** *Yeast* 1990, 6: 521-34
12. Tusher, Tibshirani, Chu: **Significance analysis of microarrays applied to the ionizing radiation response.** *PNAS* 2001, 98: 5116-5121
13. Venkateswarlu K, Kelly DE, Manning NJ, Kelly SL: **NADPH cytochrome P-450 oxidoreductase and susceptibility to ketoconazole.** *Antimicrob Agents Chemother* 1998, 42: 1756-61

Figures

Figure 1 - Venn diagram (Ley data)

Using a 5% False Discovery Rate, 2677 genes were selected through the SAM test. We applied two other methods, t-test and LPE onto the same dataset and selected the first 2677 genes with respect to the P-values given by the respective methods. The Venn diagram shows the dramatic difference one might get when applying these three different methods.

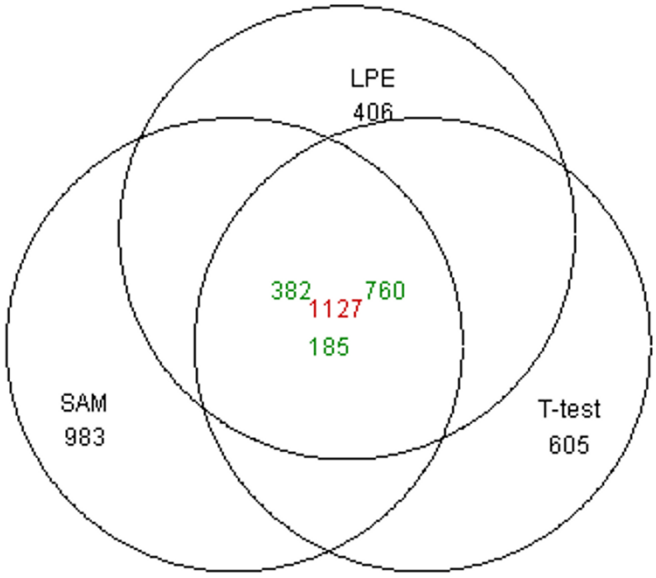


Figure 2 - Scatterplot of treated vs. controls (without variability cloud)

The up/down regulated genes selected by the three methods using 5% false discovery rate are colored red. As we can see, the difference is even more striking when the selected genes are represented on a scatter plot of treated vs. control averaged expressions.

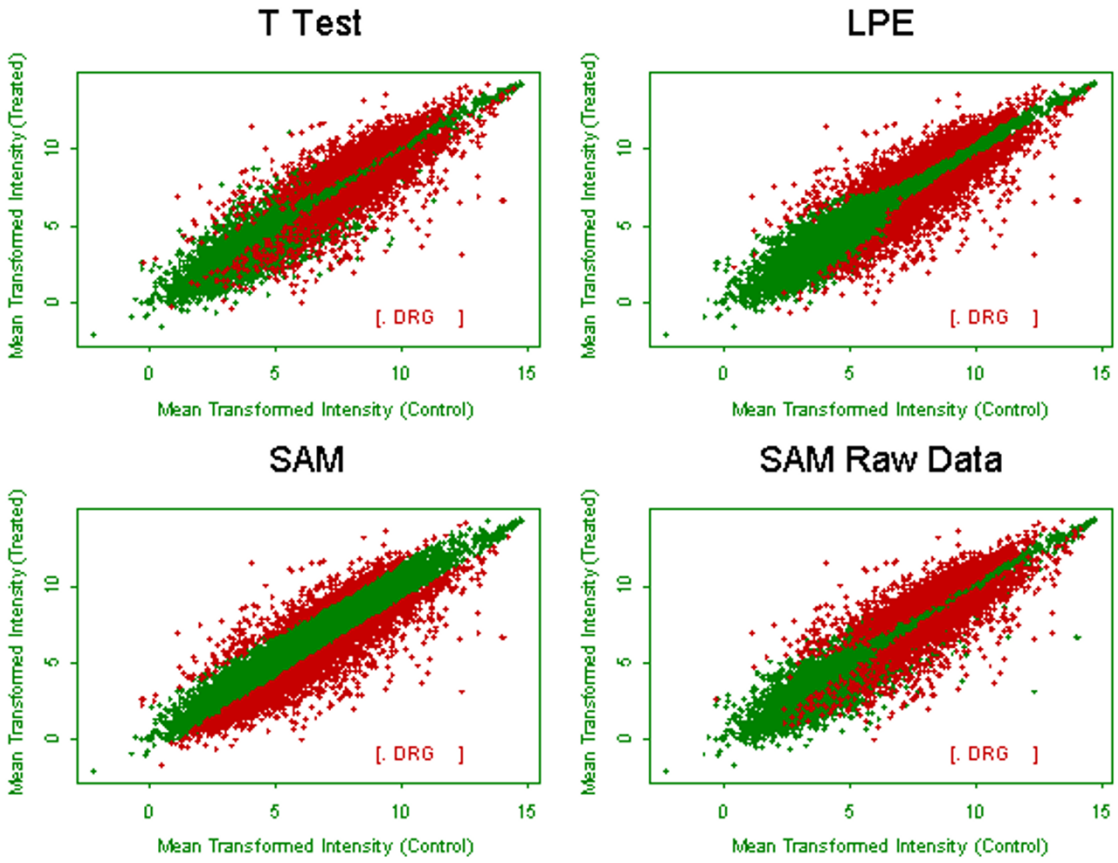


Figure 3 - Scatterplot of treated vs. controls (without variability cloud and with contours of the 2-d null distribution)

The up/down regulated genes selected by the three methods using 5% false discovery rate are in red color. Various alpha-levels of the contours of the 2-d null distribution are added. Genes outside the contours are more likely to be regulated genes.

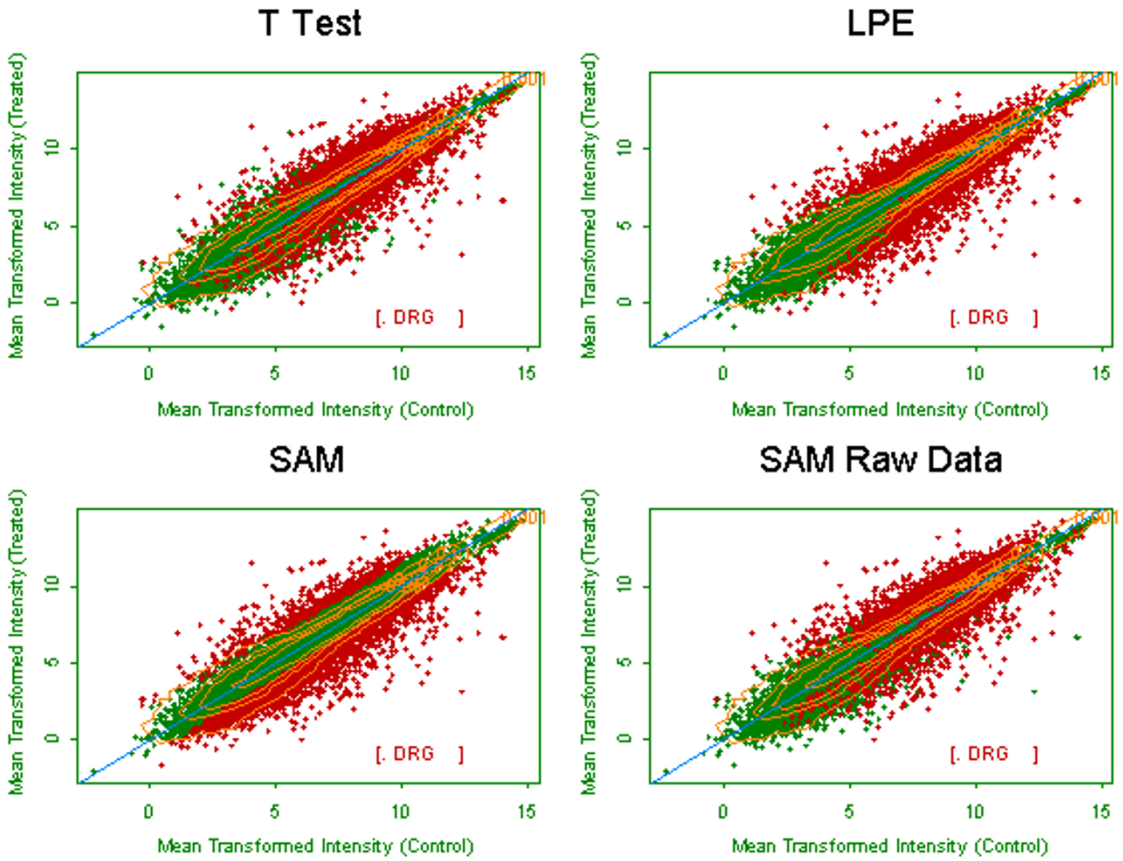


Figure 4 - Real controls vs. Virtual Controls of the Breast Cancer Data

Bootstrapped controls based on 17 real controls selected randomly were used to draw the contours of the 2-d null distribution. The scatterplot of the average of 17 real controls vs. the average of another 17 real controls was added to the 2-d contours drawn on the bootstrapped controls. This comparison clearly shows that both distributions almost perfectly overlap.

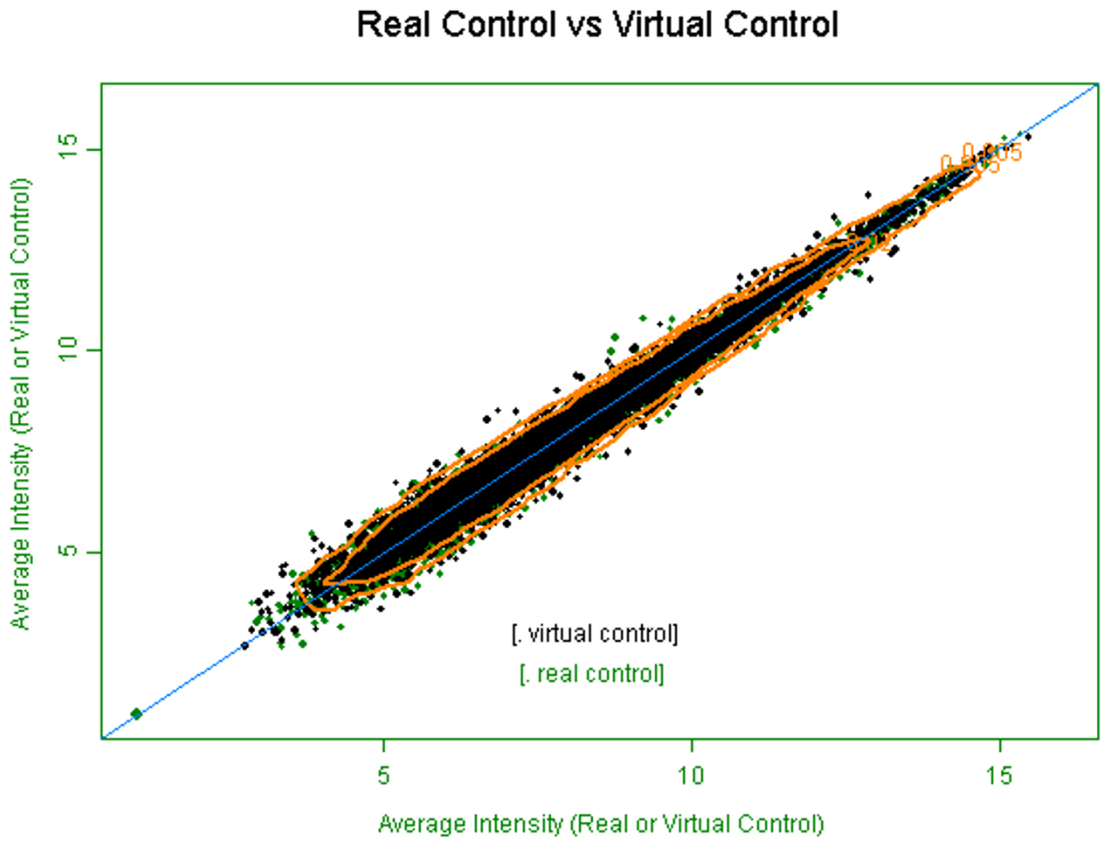
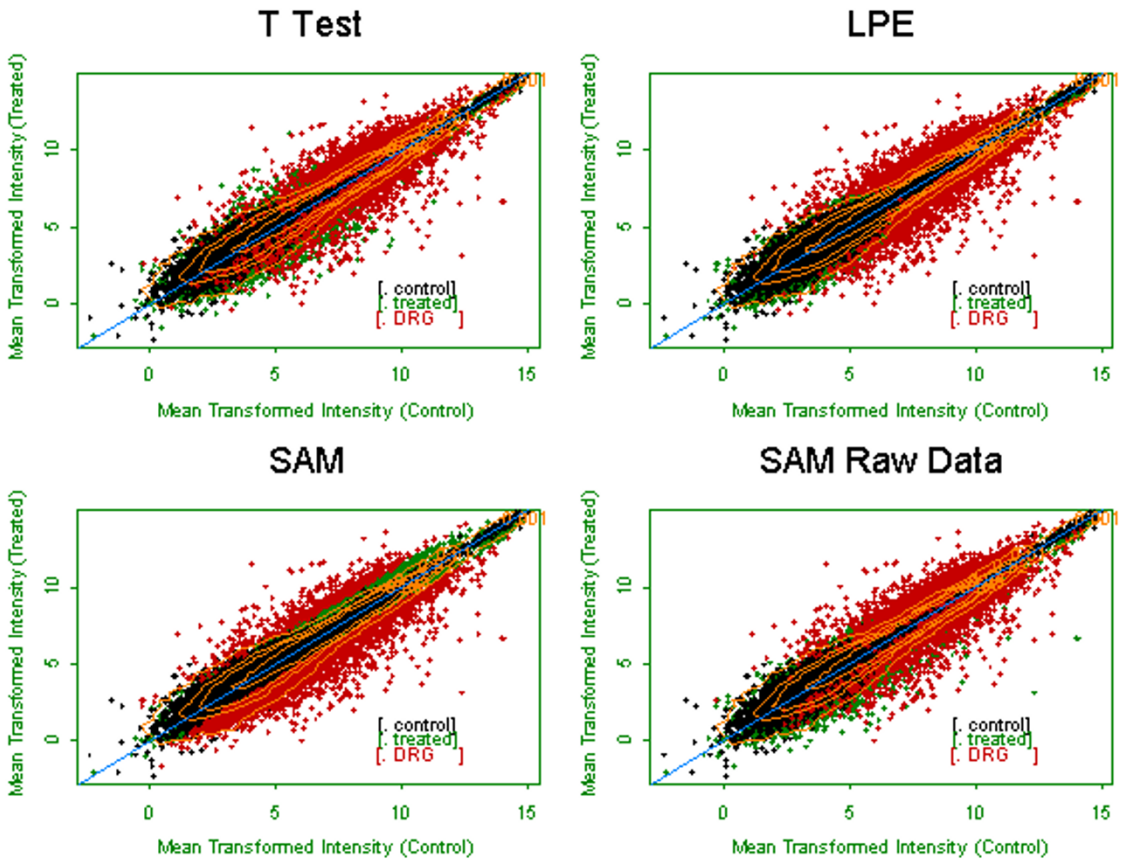


Figure 5 - Scatterplot of treated vs. controls (with variability cloud and contours of the 2-d null distribution)

The up/down regulated genes selected by the three methods using 5% false discovery rate are in red color. The variability plot and various alpha-levels of the contours of the 2-d null distribution are added.



Tables

Table 1 - Number of Genes Outside the Density Curve and the Corresponding Percentage (T-cell data)

Density level	0.001	0.005	0.01	0.02	0.05
T-test	1657	2027	2219	2395	2603
	0.619	0.757	0.829	0.895	0.972
SAM using raw data	1717	2110	2277	2450	2616
	0.641	0.788	0.851	0.915	0.977
SAM	1609	2199	2470	2670	2677
	0.601	0.821	0.923	0.997	1.000
LPE	2065	2513	2616	2658	2673
	0.771	0.939	0.977	0.993	0.999

Using a 5% False Discovery Rate, 2677 genes were selected for the T-cell data. Table 1 lists the number of selected genes that are outside the different levels of density curves and the corresponding percentages.

Table 2 - Number of Genes Outside the Density Curve and the Corresponding Percentage (Yeast Data)

Density level	0.005	0.01	0.02	0.05	0.1
T-test	871	1130	1445	1972	2431
	0.319	0.414	0.529	0.722	0.890
SAM using raw data	923	1183	1508	2040	2479
	0.338	0.433	0.552	0.747	0.908
SAM	963	1224	1586	2225	2695
	0.353	0.448	0.581	0.815	0.987
LPE	1119	1457	1872	2413	2640
	0.410	0.534	0.685	0.884	0.967

Using a 5% False Discovery Rate, 2731 genes were selected for the yeast data. Table 2 lists the number of selected genes that are outside the different levels of density curves and the corresponding percentages.

Table 3 - Number of Genes Outside the Density Curve and the Corresponding Percentage (Breast Cancer Data)

Density level	0.005	0.01	0.02	0.05	0.1
T-test	253	374	495	708	834
	0.289	0.427	0.565	0.808	0.952
SAM using raw data	220	333	457	681	832
	0.251	0.380	0.522	0.777	0.950
SAM	285	433	584	781	864
	0.325	0.494	0.667	0.892	0.986
LPE	303	450	593	743	829
	0.346	0.514	0.677	0.848	0.946

Using a 5% False Discovery Rate, 876 genes were selected for the breast cancer data. Table 3 lists the number of selected genes that are outside the different levels of density curves and the corresponding percentages.

Table 4 - Advantages and Disadvantages of different methods

t-test	Advantage: • Considers gene-specific variance. • If the sample size is large, it will be a good choice.
	Disadvantage: • If the sample size is small, the variance estimate may not be accurate. • It does not deal with multiple testing issue.
LPE	Advantage: •Borrows strength from neighboring genes. • If the sample size is small and gene variances are similar for same intensity levels, it is a good choice.
	Disadvantage: • Does not consider gene-specific variance and assumes the variance depends on the mean intensity. (If we know that there are quite a number of genes with gene-specific variances, this method is not a good choice). • It does not deal with multiple testing issue.
SAM with transformed data	Advantage: • Deals with multiple testing issues using permutation based false discovery rate. • Consider gene-specific variance.
	Disadvantage: • Filters some high intensities genes with low variances because of the fudge factor. • The permutation based false discovery rate may not be accurate if there are many regulated genes and the sample size is small since the permuted dataset may not approximate the null distribution well.
SAM with untransformed data	Advantage: • Deals with multiple testing issues using permutation based false discovery rate. • Consider gene-specific variance.
	Disadvantage: • The permutation based false discovery rate may not be accurate if there are many regulated genes and the sample size is small. • It is not as powerful as the transformed data when the variances in the two groups differ a lot, that is, if the intensities in one group is high, while the intensities in another group is low.
	Remarks: • Filters some low intensity genes with low variance.

Additional files

Additional file 1 – BSTRP_add_table1.pdf (additional Table 1)
Description: BSTRP_add_ta ble1.pdf shows the overlap among different methods for the yeast data.

Additional file 2 – BSTRP3_additional_figure1.png (additional Figure 1)
Additional file 3 – BSTRP3_additional_figure2.png (additional Figure 2)
Additional file 4 – BSTRP3_additional_figure3.png (additional Figure 3)
Additional file 5 – BSTRP3_additional_figure4.png (additional Figure 4)