Published in final edited form as:

Comput Biol Med. 2014 March ; 46: 1–10. doi:10.1016/j.compbiomed.2013.12.002.

Empirical Evaluation of Consistency and Accuracy of Methods to Detect Differentially Expressed Genes Based on Microarray Data

Dake Yang^a, Rudolph S. Parrish^a, and Guy N. Brock^a

Dake Yang: d0yang03@louisville.edu; Rudolph S. Parrish: rudy.parrish@louisville.edu; Guy N. Brock: guy.brock@louisville.edu

^aDepartment of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, KY 40202

Abstract

Background—In this study, we empirically evaluated the consistency and accuracy of five different methods to detect differentially expressed genes (DEGs) based on microarray data.

Methods—Five different methods were compared, including the *t*-test, significance analysis of microarrays (SAM), the empirical Bayes *t*-test (eBayes), *t*-tests relative to a threshold (TREAT), and assumption adequacy averaging (AAA). The percentage of overlapping genes (POG) and percentage of overlapping genes related (POGR) scores were used to rank the different methods on their ability to maintain a consistent list of DEGs both within the same data set and across two different data sets concerning the same disease. The power of each method was evaluated based on a simulation approach which mimics the multivariate distribution of the original microarray data.

Results—For smaller sample sizes (6 or less per group), moderated versions of the *t*-test (SAM, eBayes, and TREAT) were superior in terms of both power and consistency relative to the *t*-test and AAA, with TREAT having the highest consistency in each scenario. Differences in consistency were most pronounced for comparisons between two different data sets for the same disease. For larger sample sizes AAA had the highest power for detecting small effect sizes, while TREAT had the lowest.

Discussion—For smaller sample sizes moderated versions of the *t*-test can generally be recommended, while for larger sample sizes selection of a method to detect DEGs may involve a compromise between consistency and power.

© 2013 Elsevier Ltd. All rights reserved.

Correspondence to: Guy N. Brock, guy.brock@louisville.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest

The authors have no conflict of interests to declare.

Keywords

microarrays; differential gene expression; moderated t-tests; empirical study; consistency of gene lists; simulation study

1. Introduction

DNA microarrays have allowed investigators to compare gene expression values (measured as relative mRNA abundance) between two or more tissue samples for thousands of genes within the cell. However, because of the high dimensionality of the data with a relatively small number of replicates, microarrays have been referred to as ‘An array of problems’ [1]. Two of the main issues with microarray data are that they can be very noisy (both biological and technical noise), and that they contain a much larger number of mRNA expression measurements relative to the number of samples [2, 3]. Hence, a central question concerning microarrays is the reproducibility of results from multiple studies of the same disease, in particular with regard to the lists of differentially expressed genes (DEGs) [4, 5, 6] and gene-sets for classification studies [7, 8] that are found.

Recent studies have suggested that though the concordance between DEG lists from separate studies may be low, the false discovery rate of subsamples relative to the full data set tends to be low [9]. This suggests that each DEG list comprises mostly ‘true’ DEGs. This finding was reaffirmed when the correlation between DEGs was taken into consideration, in a study which investigated the consistency between DEG lists from two separate studies of the same disease [10]. Due to the high correlation between gene expression measurements, Zhang et al. [10] introduced a new measure for the concordance between two DEG lists, which took into account this correlation. Using the percentage of overlapping genes related (POGR) score, the authors demonstrated that while the DEG lists from two independent studies may not directly overlap with each other, each gene from one list is likely to be correlated with at least one gene from the second list.

An open problem not investigated by either of these two studies, is the influence that the type of test statistic has on the reproducibility of the results. The *t*-test is a popular choice for detecting DEGs in microarray studies, and is well-known to have robust properties (e.g., to non-normality) when the sample size is sufficient (typically, $n \geq 25$). However, the traditional *t*-test has been documented to have problems in microarray studies, particularly for low expression levels when the sample size is small [11, 12]. In this case, a gene with a low expression level but small variance can result in a large absolute *t*-statistic even when the mean difference in expression level is small. These genes will be declared differentially expressed, even if the difference in expression is not biologically meaningful. Conversely, a gene with a high mean difference in expression level may still result in a small *t*-statistic, if the estimate of the variance is unstable and unusually large (e.g., due to outliers).

Due to the huge data volume and inherent variation in microarrays, several statistical methods have been proposed to address these problems [11, 12, 13]. One of the earliest methods to appear was the significance analysis of microarrays (SAM) [11]. To solve the problem of unstable variances in gene expression measurements, SAM modifies the

standard t -statistic by adding a small ‘fudge factor’ to the variance in the denominator. This modified test statistic is compared to an expected value under the null hypothesis, which is determined by permutations of the gene expression measurements. Differences between observed and expected values which are above a threshold are considered statistically significant, where the threshold is determined by the desired false discovery rate. Smyth [12] proposed a similarly derived empirical Bayes (eBayes) approach, which shrinks the estimated sample variances towards a pooled estimate. As an alternative to these two approaches, Pounds and Rai [14] proposed a method based on assumption adequacy averaging (AAA), which is robust to violations of normality. This approach incorporates an orthogonal test of normality of the gene expression measurements, and uses the resulting empirical Bayes posterior probability (EBP) estimate from this test to inversely weight the EBP values obtained from the t -test and non-parametric Wilcoxon rank-sum test [15]. Lastly, as an extension to the empirical Bayes method of Smyth [12], McCarthy and Smyth [16] proposed to incorporate a biologically meaningful threshold into the test of differential expression (t -tests relative to a threshold, or TREAT). All of these approaches are designed to have robust performance, particularly for experiments with small numbers of arrays.

Although these methods have been proposed as improvements in detecting DEGs, few studies have empirically compared their performance [17]. Hence, in this paper we investigate both the consistency and power in determining DEGs between five different methods (traditional t -test, SAM, eBayes, AAA, and TREAT), based on three different empirical studies. In the first study we evaluate the effect of sample size reduction on maintaining a consistent list of DEGs using subsets from a single data set. In the second study, we evaluate the consistency for each method when comparing DEG lists obtained from two independent studies of the same disease. Lastly, we conduct a simulation study which evaluates the power and error rate of each method based on an approach which closely models the multivariate distribution of the original microarray data.

2. Methods

We performed three separate experiments to evaluate the consistency (reproducibility), sensitivity (power), and error rate (false discovery rate) of five different methods (t -test, SAM, eBayes, AAA, and TREAT) for determining DEGs in microarray data. The first experiment evaluated the self-consistency of each method based on subsets of the same microarray data set. The goal here was to evaluate how well each method performed at maintaining the same ordering of DEGs as the sample size is decreased. Secondly, we evaluated the consistency of DEG rankings for each method based on subsets of two different data sets concerning the same disease. The goal here was to evaluate whether certain methods outperformed others at maintaining a consistent list of DE genes across different studies of the same disease. Lastly, to ensure that the DEG lists returned by each method are appropriate and detecting truly DEGs, we conducted a simulation study to evaluate the power and false discovery rate of each method. Our simulation approach is based on a method for generating data which closely resembles the multivariate distribution of gene expression values observed in the original microarray data [18].

2.1. Methods for detecting differentially expressed genes

We selected five methods (*t*-test, SAM, eBayes, AAA, and TREAT) for determining DEGs in microarray data. In each case, the goal is to detect which genes are differentially expressed between two classes of samples (e.g., normal and diseased). A table summarizing the distinguishing characteristics of each of the methods is given in Table 1. Supplementary File A provides technical details concerning each of the methods, and references to software.

2.2. POG and POGR scores

To measure the consistency between two DEG lists, we use the percentage of overlapping genes (POG) metric and its extension to incorporate correlated gene expression changes, the POGR score [10]. The POG has been previously used to measure the reproducibility of DEG lists between different platforms [19], as well as from independent studies of the same disease [10]. The POGR score is a natural extension of the POG score which considers not only those genes which are shared between the two lists, but also those genes which are highly correlated with each other. The nPOG and nPOGR scores are normalized versions which account for the positive correlation of both scores with the length of the gene lists. They are analogous to the chance-corrected kappa coefficient [20]. A technical description of all the scores is given in Supplementary File A.

2.3. Study Design

2.3.1. Consistency of DE detection methods within a single data set—Three data sets of different diseases were used to evaluate the consistency of methods to detect differential expression, based on subsets of the same data (Table 2). Before all the procedures we filtered the raw data to remove invariant transcripts, using the `nsFilter` function in the **genefilter** package [21]. Transcripts were filtered using `var.cutoff` with the value set to 0.5, so that all transcripts with an interquartile range below the median value were removed. This is motivated by the observation that in many tissues only 40% of the genes are expressed. We additionally set the argument `remove.dupEntrez=TRUE`, so that for transcripts mapping to the same Entrez Gene ID, the transcript with the largest variance was retained and the others removed. All of the data sets were obtained from Bioconductor [22] as normalized gene expression data sets, and were named based on the first three letters of the first author's last name.

The **CHI** data from the Ritz Laboratory [23] (Bioconductor package **ALL** [24]) consists of 128 samples from patients having acute lymphoblastic leukemia (ALL). We selected 79 B-cell tumor samples representing two molecular biology subtypes, those found to carry the BCR("breakpoint cluster region")/ABL("Abelson") translocation and those with no cytogenetic abnormalities (Neg). After screening out invariant expression profiles, the final number of probes in the data set was 4,399. The **ALO** data (Bioconductor package **ColonCA** [25]) contains 2,000 expression measurements from 40 colon cancer tumor samples and 22 normal samples [26]. No screening of transcripts was applied as this data set was already pre-processed, however expression values were first $\log_2$ transformed prior to analysis. The **GOL** data [27] (Bioconductor package **golubEsets** [28]) consists of 47 patients with ALL and 25 patients with acute myeloid leukemia (AML) (the combined training and test samples

from that study). The samples were assayed using Affymetrix Hgu6800 chips and data on the expression of 7,129 genes (Affymetrix probes) are available. Negative expression values were set to missing. After screening out invariant probes and removing probes with fewer than three non-missing values in each group, a total of 2,400 probes remained. Expression values were then $\log_2$ transformed, and remaining missing values were imputed using the k-nearest-neighbors (kNN) algorithm in Bioconductor package **impute** [29, 30]).

To evaluate the consistency of the five DE detection methods, we randomly selected a varying number of subsets without replacement from the full data (four subset sizes in total from each data set). Our smallest subset size was based on three replicates per group, which though underpowered is still commonly seen in experimental biology studies. For the **CHI** data set, we chose subsets of size 3, 5, 10 and 25 randomly and without replacement from each of the 37 BCR/ABL and 42 Neg samples (total sample size of 6, 10, 20 and 50 samples). For the **ALO** data (40 tumor samples, 22 normal samples), we chose subsets of total size 6, 12, 27 and 51 in a 2:1 ratio of tumor to control samples. Finally, for the **GOL** data (47 ALL, 25 AML) we chose total subset sizes of 6, 12, 27 and 51 in a 2:1 ratio of ALL to AML patients.

Gene-specific test statistics were calculated for both the subsets and the full data using the five methods, and genes were ranked based on the absolute values of these statistics (in the case of SAM) or the p -values associated with them (all other methods). Next, list lengths of 50, 100, 500, and 1000 were used to calculate the POG, nPOG, POGR and nPOGR scores between the the subsets and the full data set. The entire process was repeated 1000 times for each combination of method, data set, subset length, and list length. For calculating the POGR scores, we used a correlation cutoff based on the 99.9% of the entire distribution of pairwise correlations in the data set. All computations were conducted using the R statistical computing software [31].

2.3.2. Consistency of DE detection methods between data sets—Six data sets from independent studies of three diseases were analyzed to evaluate the consistency of DE detection methods between different data sets concerning the same disease (Table 3). For each disease, we only analyzed the genes that were present on both platforms after screening out invariant and redundant probes. In all cases, the cDNA data were normalized by subtracting the median value from each array, while the Affymetrix GeneChip data were normalized by RMA (Robust Multi-array Analysis) [32] using Bioconductor package **affy** [33].

The prostate cancer cDNA microarray data (**LAP**) consisted of 103 total samples, 62 tumor samples and 41 normal prostate samples [34]. Of these, we were able to successfully read in 54 tumor and 32 normal prostate samples. The Affymetrix data (Affymetrix U95Av2 arrays) consisted of 52 tumor and 50 non-tumor samples (**SIN**) [35]. A total of 2,072 genes were found to be in common between the two platforms after screening, which we considered to be the complete set of genes in both cases. Subsets of size 3, 5, 10 and 25 samples were drawn randomly and without replacement from both the tumor and normal tissue samples for each of the two data sets (total sample size of 6, 10, 20, and 50 samples).

The lung cancer cDNA microarray data consisted of 18 total samples, 13 squamous cell lung cancer and 5 normal lung specimens (GAR) [36]. The lung cancer Affymetrix data (U95Av2 oligonucleotide arrays) [37] consisted of 21 squamous cell lung carcinomas and 17 normal lung specimens (BHA). The intersection of the two-data sets post-screening was 1,893 genes. Since there were few normal specimens for the cDNA data, the subsets were drawn asymmetrically from the tumor and normal samples. Subsets of size 6 (3 normal, 3 tumor), 8 (2 normal, 6 tumor), 12 (3 normal, 9 tumor), and 16 (4 normal, 12 tumor) were drawn randomly and without replacement from the cDNA data. For the Affymetrix data, subsets of total size 6, 12, 20, and 30 were selected in a 1:1 ratio of tumor to normal samples.

The first data set for Duchenne muscular dystrophy (DMD) contained 24 samples from 12 DMD patients and 12 unaffected controls (HAS) [38], while the second consisted of 36 samples from 22 DMD patients and 14 controls (PES) [39]. The two data sets were based on Affymetrix HG-U95Av2 and HG-U133A GeneChips, respectively. After screening, there were 3,158 genes in common between the two data sets. Subsets of total size 6, 10, 14 and 20 were selected in a 1:1 ratio of DMD to normal samples for each of the data sets.

After selecting subsets from the common set of gene expression measurements for the two data sets corresponding to the same disease, we again used the five methods (*t*-test, eBayes, SAM, AAA, and TREAT) to determine the DEG lists for each subset. Genes were subsequently ranked on the basis of the absolute values of their test statistics (SAM) or the *p*-values associated with them (all other methods), and list lengths of 50, 100, 500, and 1000 were used to determine the POG, nPOG, POGR and nPOGR scores between the two subsets corresponding to the same disease. We repeated the above procedure 1000 times for each combination of method, data set, subset length, and list length. The correlation cutoff for the POGR scores were based on the 99.9% of pairwise correlations, separately in each data set.

2.3.3. Simulation study—We conducted a simulation study To evaluate the power of each method to detect true DEGs and ensure that the proper error rate (false discovery rate) was maintained in each case. Prior studies [40] have noted that results based on simulated data are sensitive to assumptions regarding data covariance and error structure, which for microarray data is frequently poorly characterized. To address this issue, we used a simulation approach specifically designed to mimic the multivariate distribution of gene expression in the original microarray data [18]. In brief, two systems of transformations (Box-Cox transformation and Johnson system of distributions) are used to transform the expression measurements for each gene to normality, and then a modified Cholesky algorithm is used to estimate the gene-wise covariance matrix and ensure that it is positive definite. We selected the three data sets used in the single data consistency study (CHI, ALO, and GOL) as a basis for simulating microarray data, using the ‘Neg’ samples from CHI, the normal samples from ALO, and the ALL samples from GOL. For each simulation, we selected 1500 genes at random from the original reduced data, and artificially generated a total of 60 DEGs with varying effect sizes between the two study groups (10 genes each with effect sizes 0.5, 1.0, 1.5, 2.0, 3.0, and 4.0). Sample sizes of 3, 6, 12, 25, and 50 per group were evaluated, and a total of 500 simulations were conducted for each scenario. *P*-values from each method were adjusted based on the Benjamini-Hochberg [41] method to

maintain an overall false-discovery rate of 0.05. Overall power, empirical false discovery rate, and power for each effect size were all estimated based on the mean of the 500 simulations.

3. Results

3.1. Consistency of DE detection methods using subsets from the same data set

For the three data sets mentioned in Table 2, we compared the DEG lists determined using subsets from the same data set to test the consistency of the five methods. In each case, test statistics for differential gene expression between the two tissue types were calculated for each gene based on the full data set and each subset size, and DEG lists of length 50, 100, 500, and 1000 were obtained. For conciseness, we selected a single list length for presentation to illustrate the differences between methods. Figure 1 displays the nPOG and nPOGR scores for all three data sets as a function of subset size, for a list length of 100. Complete results for all scores (POG, POGR, nPOG, and nPOGR) by data set, list length, and subset size are given in Supplementary File B (Supplemental Figures 1 – 4). Points represent mean values while vertical bars indicate the 5th and 95th percentiles from the Monte Carlo experiment (1000 total simulations).

Several general observations can be made. First, both the POG and POGR scores increase with list length and subset size, as expected. The normalized scores in each case are relatively constant with list length, excepting the list length of 1000 which shows a precipitous decline in several cases. Presumably, this is due to the higher percentage of false positives at larger list lengths. Another factor at play is the strong inter-gene correlation. The 99.9th percentiles used for the POGR cutoffs were 0.704 for the GOL data set, 0.744 for the CHI data set, and 0.903 for the ALO data set. For the CHI and ALO data sets, the nPOGR scores for the smallest subset size were all near or at zero, reflecting the fact that the POGR scores for comparing DEG lists in these cases are no better than chance. This is attributable to the high inter-gene correlation in the case of ALO, and the relatively low POG scores in the case of CHI.

In comparing the five different methods, the moderated *t*-test methods (SAM, eBayes, and TREAT) generally have larger values when compared to the traditional *t*-test and AAA methods, with TREAT having the overall highest consistency values. This separation between the moderated test statistics and the other two methods is greatest when the subset size is small, with the consistency of the methods rapidly converging to each other as the subset size increases. The separation between methods is also greatest for the shorter list lengths. The five methods were ranked from 1 (highest) to 5 (lowest) across all combinations of data set, subset size, and list length. Figure 2, bottom panel displays the stacked bar chart showing the percentage distribution of ranks for each method, for both nPOG and nPOGR scores. The bars are color-coded by rank from highest rank (1 = light orange) to lowest rank (5 = dark orange), and height represents the percentage of that rank for the corresponding method. Methods are ordered from left to right according to overall average ranking from highest to lowest. In each case, TREAT had the highest overall consistency among the five methods with an average rank of 1.04 for the nPOG scores and 1.40 for the nPOGR scores. This was followed by SAM (average rank of 2.27 for nPOG and

2.52 for nPOGR) and eBayes (average rank of 2.71 for nPOG and 2.81 for nPOGR). The *t*-test (average rank of 4.23 for nPOG and 3.92 for nPOGR) and AAA method (average rank of 4.75 for nPOG and 4.35 for nPOGR) had the lowest consistency. However, the difference between methods is small in comparison to the effect of increasing the subset size. Second, the empirical 90% confidence intervals (interval between 5th and 95th percentiles) overlapped between the best and worst performing methods in every case, indicating that the separation between the methods was not substantial in relation to the naturally occurring variability.

3.2. Consistency of DE detection methods from different data sets of the same disease

For the six data sets of the three diseases listed in Table 3, we compared the DEG lists from the two data sets for each disease to test the consistency of the five DE detection methods. In each case, test statistics for differential gene expression between the two tissue types were calculated for each gene based on corresponding subsets for each data set and DEG lists of length 50, 100, 500, and 1000 were obtained. Figure 3 displays the nPOG and nPOGR scores for all three diseases as a function of subset size, for a list length of 100. Complete results for all scores (POG, POGR, nPOG, and nPOGR) by disease, list length, and subset size are given in Supplementary File B (Supplemental Figures 5 – 10). Points represent mean values while vertical bars indicate 90% empirical confidence intervals (5th and 95th percentiles). The POGR and nPOGR scores displayed use the first data set listed for each disease in Table 3 as the reference data set. Results for the opposite direction (second data set treated as reference) are given in Supplementary File B (Supplemental Figures 7 and 10).

In contrast to the single data results, consistency between the two data sets depended on the disease under consideration. For the prostate cancer (PC) data sets, the nPOG scores were very low for all subset sizes irrespective of the DE detection method, indicating low consistency between the two data sets. In contrast, for both the lung cancer (LC) and DMD data sets both higher scores and a greater separation between methods is observed. In particular, the eBayes, Treat, and SAM methods all clearly outperform the *t*-test and AAA methods, in particular for smaller subset sizes and list lengths. The distinction between methods exceeds the naturally occurring variability (no overlap between 90% empirical confidence intervals) for the smallest subset size and list lengths in each case. Even at larger subset sizes, there is a marked distinction between the means for the three moderated methods versus the *t*-test and AAA.

Among the three moderated methods, TREAT had slightly higher consistency relative to SAM and eBayes for the nPOG scores. For the nPOGR scores, the result was dependent on which data set was selected to obtain the reference DEG list. When the first data set in Table 3 was selected (smaller of the two data sets), TREAT had consistently higher scores relative to SAM and eBayes (Supplemental Figure 9 in Supplementary File B). However, when the second data set was selected as the reference, both SAM and eBayes had higher scores compared to TREAT in many circumstances (Supplemental Figure 10 in Supplementary File B).

Further in contrast to the single data results, the normalized POG and POGR scores for the LC and DMD data sets were not consistent across the different list lengths. The nPOG and nPOGR scores showed differential patterns with list length, as the nPOG increased with list length while the nPOGR decreased. The differences in pattern can be explained by inspecting the plots for the raw POG and POGR scores. The POGR scores are higher than the POG scores for both data sets at the smaller list lengths, but the two scores are nearly identical at list lengths of 500 and 1000. Thus, at these larger list lengths the methods are not discovering any additional genes that are highly correlated with the genes from the second DEG list. The 99.9th pairwise correlation percentiles used for the POGR cutoffs were fairly high and consistent for the two LC data sets (0.863 for GAR and 0.854 for BHA) and the two DMD data sets (0.869 for HAS and 0.850 for PES). However, the cutoffs for the two PC data sets differed distinctly (0.737 for LAP and 0.942 for SIN).

The five methods were ranked across all combinations of data set, subset size, and list length. Figure 2, top panel displays the bar chart showing the distribution of ranks for each method, for both nPOG and nPOGR scores. The bars are color-coded by rank from highest rank (1 = light orange) to lowest rank (5 = dark orange), and height represents the percentage of that rank for the corresponding method. Methods are ordered from left to right according to overall average ranking from highest to lowest. TREAT had the overall highest consistency among the five methods for both the nPOG score (average rank of 1.48) and the nPOGR score (average rank of 1.84). This was followed by the other two moderated methods eBayes (average rank of 2.56 for nPOG and 2.38 for nPOGR) and SAM (average rank of 2.75 for nPOG and 2.52 for nPOGR). The *t*-test (average rank of 3.79 for nPOG and 3.94 for nPOGR) and AAA method (average rank of 4.42 for nPOG and 4.32 for nPOGR) again had the lowest average rankings.

3.3. Simulation study

False discovery rates for each of the five methods were maintained at or below the nominal level of 0.05 (Supplementary File B, Supplemental Figure 11), with the exception of the smallest sample size (3 per group) where several methods exceeded the 0.05 level. Overall power for smaller sample sizes (3 and 6 per group) was highest for eBayes and TREAT, with little difference in power for sample sizes of 12 per group or more. TREAT had the lowest power for larger sample sizes, which was attributable to the lower power of TREAT to detect the smallest effect size of 0.5 (see bottom panel of Supplemental Figure 12 in Supplementary File B). This is expected, since TREAT is specifically designed to detect DEGs with fold-changes above a certain threshold (here, set to the default $\log_2(1.1)$). The power of SAM was somewhat lower relative to eBayes, which for the CHI and GOL data sets may be attributable to the lower error rate observed for SAM. For the largest sample sizes (25 and 50 per group), the AAA method has the largest overall power. Again, this is mainly evident in the smallest effect size.

4. Discussion

In this article, we compared the consistency and accuracy (power and error rate) of five distinct methods (*t*-test, eBayes, SAM, AAA, and TREAT) for determining DEGs. The

calculation of the test statistic and p -value for each method captures different aspects of the changes in expression measurements for each gene, resulting in potentially different lists of DEGs. The traditional t -test still remains a popular choice, and we compared the t -test with four methods that have been custom-tailored to gene expression studies to evaluate the empirical impact of choice of methodology on the consistency and accuracy of the resulting list of DEGs.

Based on our results, the moderated versions of the t -test (SAM, eBayes, and TREAT) had higher consistency relative to both the t -test and AAA, which combines the t -test with a non-parametric approach. The differences were particularly pronounced for comparisons between two different data sets for the same disease, and remained evident up to subset sizes of 10 per group and DEG list lengths of 500. Differences for studies involving subsets from the same data set were less pronounced but consistent. Consistency of DEG rankings based on p -values from traditional t -statistic has been previously demonstrated to be low [40], because the per-gene estimate of variability can be unstable. Moderated versions of t -statistics, which employ variance stabilizing methods, result in improved consistency.

Among all the methods, we found that TREAT had the highest overall consistency in each scenario. TREAT tests whether differences in gene expression are above a given threshold and essentially eliminates genes with low $\log_2$ ratios from the DEG list. This result agrees with conclusions from the MicroArray Quality Control (MAQC) project [42, 40], which found that DE gene rankings based on fold-changes (FCs) produced the most consistent gene lists. In fact, Shi *et al.* [40] recommended coupling FC ranking with a non-stringent p -value cutoff to generate reproducible DEG lists. Generally, we feel this is a reasonable approach which experimental biologists commonly practice, by naturally ranking the most important DEGs based on FC or biological relevance.

When considering the power of the various approaches, SAM, eBayes, and TREAT had a clear advantage the t -test and AAA method for smaller sample sizes (6 or less per group). However, for larger sample sizes (25 or more per group) the AAA method had the highest power for detecting the smaller effect sizes. In contrast, TREAT had the lowest power for detecting the smaller effect sizes due to its intentional design to detect DEGs with FCs above a certain threshold. This difference is also observed in the number of DEGs identified in the full data sets used in this study (c.f. Supplemental Table 1 in Supplementary File B), where TREAT identified the fewest number of DEGs in nearly every case (the only exceptions being the cDNA data sets LAP and GAR, which had higher coefficients of variation (CV) relative to the Affymetrix data sets). Of course, lowering this threshold will increase the power of TREAT, which is identical to eBayes when the $\log_2$ threshold is set to zero. Our comparisons used the default threshold of $\log_2(1.1)$ throughout.

In our research, we used the POG and POGR metrics to evaluate the consistency of DEG lists. As noted by Zhang *et al.* [10], the DEG lists from small-scale microarray studies may only contain a small portion of the total number of differentially expressed genes, and the POGR score is helpful for measuring how well different methods retain genes that are biologically important. Normalized versions of the scores (nPOG, nPOGR) [10] were intended to stabilize the scores with respect to list length. However, the normalized scores

were not always constant for larger list lengths (500 and 1000), where the fraction of ‘true’ DEGs decreased. The nPOGR score was more strongly affected by increasing list length relative to the nPOG score, and was also strongly impacted by the degree of inter-gene correlation in the samples (the scores decreased for larger list lengths if the degree of correlation was high). Hence, despite the biological motivation for using the nPOGR score, the nPOG score is more stable for making comparisons between methods. Both scores could potentially be improved by considering the rank of the consistent genes, e.g. by weighting them in a fashion inversely proportional to the rank.

Subset size had a dramatic impact on consistency scores when considering subsets from the same data set. However, the impact of increasing subset size was less substantial when comparing two different data sets concerning the same disease. In the first case, the increasing scores are a simple reflection of the asymptotic convergence of the results to the population values. However, when comparing two different data sets, differences in experimental factors between the two studies (e.g., platform and site differences) may limit the level of agreement between the studies, irrespective of sample size. Among the three disease we investigated, the DMD data compared two Affymetrix data sets while the other two diseases (PC and LC) compared one Affymetrix data set with one cDNA data set. Consistency between studies was highest for DMD and quite low for PC, with consistency between the LC studies closer to that of DMD. This suggests that while platform similarity plays a role in consistency between studies, other factors may be more influential. In comparing the LC and PC data sets, the CVs for the PC cDNA data (*LAP*) were lower than the CVs for the LC cDNA data (*GAR*) (c.f. Supplemental Table 1 in Supplementary File B). CVs for the Affymetrix data sets were considerably smaller, and lower for LC relative to PC. However, the magnitude of FCs detected among the DEGs was higher for LC relative to PC in both data sets, with a correspondingly greater number of DEGs detected despite considerably larger sample sizes for the PC data. Hence, the greatest determinant of consistency between studies is likely the magnitude of biological effect under investigation.

There are several limitations with our study that are areas of future research. First, since our evaluation of consistency used fixed list lengths, a natural follow-up study would compare consistency of lists based on a significance threshold with the recommendation from Shi *et al.* [40]. Second, in our study design we ignored design factors which could cause dependencies between samples from the same phenotypic group (diseased or control). This may have an impact on the results, and in particular a sampling design which accounts for the design factors in the study may be more appropriate. Additionally, using a more complicated model for the gene expression values, which includes the original design variables in the model, would be worthwhile to investigate. In this case, a comparison between the standard linear model and the hierarchical empirical Bayesian model [12] would be warranted. A final consideration in this regard would be the effect of using different normalization methods, as this can have an important impact on significance testing [43].

5. Conclusion

In this study, we empirically evaluated the consistency and accuracy of five different methods to detect DEGs based on microarray data. We found that TREAT had the highest overall consistency in each scenario, in agreement with prior studies which found gene rankings based on fold-changes produced the most consistent gene lists. For smaller sample sizes (6 or less per group), the moderated versions of the t -test (SAM, eBayes, and TREAT) were superior in terms of both power and consistency relative to the t -test and AAA, which combines the t -test with a non-parametric approach. The differences in consistency were most pronounced for comparisons between two different data sets for the same disease. However, the AAA method had the highest power for detecting small effect sizes in larger samples (25 or more per group). In contrast TREAT, due to its intentional design of eliminating genes with low $\log_2$ ratios from the DEG list, had the lowest power for detecting small effect sizes. Thus, while moderated versions of the t -test can generally be recommended for smaller sample sizes, for larger sample sizes selection of a method to detect DEGs may involve a compromise between consistency and power.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported by DOE Grant 10EM00542, by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health (Grant no. P20GM103453), and by NIH Grant no. R03DE021460. The authors thank the editors and four anonymous reviewers, whose comments greatly improved the content of this manuscript.

References

1. Frantz S. An array of problems. *Nat Rev Drug Discov.* 2005; 4(5):362–363. [PubMed: 15902768]
2. Tu Y, Stolovitzky G, Klein U. Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci USA.* 2002; 99(22):14031–14036. [PubMed: 12388780]
3. Xu P, Brock GN, Parrish RS. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Comput Stat Data Anal.* 2009; 53(5):1674–1687.
4. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics.* 2005; 21(2):171–178. [PubMed: 15308542]
5. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet.* 2005; 365(9458):488–492. [PubMed: 15705458]
6. Boulesteix AL, Slawski M. Stability and aggregation of ranked gene lists. *Brief Bioinform.* 2009; 10(5):556–568. [PubMed: 19679825]
7. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics.* 2010; 26(3):392–398. [PubMed: 19942583]
8. Rajapakse JC, Mundra PA. Multiclass gene selection using paretofronts. *IEEE/ACM Trans Comput Biol Bioinform.* 2013; 10(1):87–97. [PubMed: 23702546]
9. Zhang M, Yao C, Guo Z, Zou J, Zhang L, Xiao H, Wang D, Yang D, Gong X, Zhu J, Li Y, Li X. Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics.* 2008; 24(18):2057–2063. [PubMed: 18632747]

10. Zhang M, Zhang L, Zou J, Yao C, Xiao H, Liu Q, Wang J, Wang D, Wang C, Guo Z. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*. 2009; 25(13):1662–1668. [PubMed: 19417058]
11. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*. 2001; 98(9):5116–5121. [PubMed: 11309499]
12. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004; 3 Article3.
13. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol*. 2002; 23(1):70–86. [PubMed: 12112249]
14. Pounds S, Rai SN. Assumption adequacy averaging as a concept for developing more robust methods for differential gene expression analysis. *Comput Stat Data Anal*. 2009; 53:1604–1612. [PubMed: 20161327]
15. Wilcoxon F. Individual comparisons by ranking methods. *Biometrika*. 1945; 1:80–83.
16. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*. 2009; 25(6):765–771. [PubMed: 19176553]
17. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*. 2006; 7:359. [PubMed: 16872483]
18. Parrish RS, Spencer HJ 3rd, Xu P. Distribution modeling and simulation of gene expression data. *Comput Stat Data Anal*. 2009; 53:1650–1660.
19. Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ, Tsai CA. Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics*. 2007; 8:412. [PubMed: 17961233]
20. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971; 76:378–382.
21. Gentleman R, Carey V, Huber W, Hahne F. **genefilter**: methods for filtering genes from microarray experiments. R package version 1.38.0. 2012
22. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*. 2004; 5:R80. URL <http://www.bioconductor.org>. [PubMed: 15461798]
23. Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*. 2004; 103(7):2771–2778. [PubMed: 14684422]
24. Li X. **ALL**: A data package. R package version 1.4.12. 2009
25. Merk S. **colonCA**: exprSet for Alon et al. (1999) colon cancer data. R package version 1.4.9.
26. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA*. 1999; 96(12):6745–6750. [PubMed: 10359783]
27. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999; 286(5439):531–537. [PubMed: 10521349]
28. Golub T, Carey V. **golubEsets**: exprSets for golub leukemia data. R package version 1.4.11. 2012
29. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for dna microarrays. *Bioinformatics*. 2001; 17(6):520–525. [PubMed: 11395428]
30. Hastie T, Tibshirani R, Narasimhan B, Chu G. **impute**: Imputation for microarray data. R package version 1.32.0. 2013
31. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2012. ISBN 3-900051-07-0 URL <http://www.R-project.org>

32. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*. 2003; 19(2):185–193. [PubMed: 12538238]
33. Gautier L, Cope L, Bolstad BM, Irizarry RA. **affy**—analysis of affymetrix genechip data at the probe level. *Bioinformatics*. 2004; 20(3):307–315. [PubMed: 14960456]
34. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci USA*. 2004; 101(3):811–816. [PubMed: 14711987]
35. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. 2002; 1(2):203–209. [PubMed: 12086878]
36. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci USA*. 2001; 98(24):13784–13789. [PubMed: 11707590]
37. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA*. 2001; 98(24):13790–13795. [PubMed: 11707567]
38. Haslett JN, Sanoudou D, Kho AT, Bennett RR, Greenberg SA, Kohane IS, Beggs AH, Kunkel LM. Gene expression comparison of biopsies from duchenne muscular dystrophy (DMD) and normal skeletal muscle. *Proc Natl Acad Sci USA*. 2002; 99(23):15000–15005. [PubMed: 12415109]
39. Pescatori M, Broccolini A, Minetti C, Bertini E, Bruno C, D'Amico A, Bernardini C, Mirabella M, Silvestri G, Giglio V, Modoni A, Pedemonte M, Tasca G, Galluzzi G, Mercuri E, Tonali PA, Ricci E. Gene expression profiling in the early phases of DMD: a constant molecular signature characterizes DMD muscle from early postnatal life throughout disease progression. *FASEB J*. 2007; 21(4):1210–1226. [PubMed: 17264171]
40. Shi L, Jones WD, Jensen RV, Harris SC, Perkins RG, Goodsaid FM, Guo L, Croner LJ, Boysen C, Fang H, Qian F, Amur S, Bao W, Barbacioru CC, Bertholet V, Cao XM, Chu TM, Collins PJ, Fan XH, Frueh FW, Fuscoe JC, Guo X, Han J, Herman D, Hong H, Kawasaki ES, Li QZ, Luo Y, Ma Y, Mei N, Peterson RL, Puri RK, Shippy R, Su Z, Sun YA, Sun H, Thorn B, Turpaz Y, Wang C, Wang SJ, Warrington JA, Willey JC, Wu J, Xie Q, Zhang L, Zhong S, Wolfinger RD, Tong W. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics*. 2008; 9(Suppl 9):S10. [PubMed: 18793455]
41. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*. 1995; 57:289–300.
42. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Schrf U, Thierry-Mieg J, Wang C, Wilson M, Wolber PK, Zhang L, Amur S, Bao W, Barbacioru CC, Lucas AB, Bertholet V, Boysen C, Bromley B, Brown D, Brunner A, Canales R, Cao XM, Cebula TA, Chen JJ, Cheng J, Chu TM, Chudin E, Corson J, Corton JC, Croner LJ, Davies C, Davison TS, Delenstarr G, Deng X, Dorris D, Eklund AC, Fan XH, Fang H, Fulmer-Smentek S, Fuscoe JC, Gallagher K, Ge W, Guo L, Guo X, Hager J, Haje PK, Han J, Han T, Harbottle HC, Harris SC, Hatchwell E, Hauser CA, Hester S, Hong H, Hurban P, Jackson SA, Ji H, Knight CR, Kuo WP, LeClerc JE, Levy S, Li QZ, Liu C, Liu Y, Lombardi MJ, Ma Y, Magnuson SR, Maqodi B, McDaniel T, Mei N, Myklebost O, Ning B, Novoradovskaya N, Orr MS, Osborn TW, Papallo A, Patterson TA, Perkins RG, Peters EH, Peterson R, et al. The microarray quality control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 2006; 24(9):1151–1161. [PubMed: 16964229]
43. Parrish RS, Spencer HJ 3rd. Effect of normalization on significance testing for oligonucleotide microarrays. *J Biopharm Stat*. 2004; 14(3):575–589. [PubMed: 15468753]

44. Smyth GK, Michaud J, Scott HS. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*. 2005; 21(9):2067–2075. [PubMed: 15657102]

Summary

This paper presents an extensive empirical evaluation of the consistency and accuracy of five different methods to detect differentially expressed (DE) genes based on microarray data. Our study compares five different methods for determining differential expression, including the traditional *t*-test, the significance analysis of microarrays (SAM), the empirical Bayes *t*-test (eBayes), *t*-tests relative to a threshold (TREAT), and assumption adequacy averaging (AAA). SAM, eBayes, and TREAT can all be considered 'moderated' versions of the *t*-test, because they use a robust or moderated estimate of the sample variance of gene expression measurements. Our Monte Carlo study evaluated nine different data sets, four different subset sizes, and four different list lengths. Consistency was evaluated both within a single study and between different studies, using nine different data sets, four different subset sizes, and four different list lengths. In addition, we evaluated the power of each method based on a simulation approach which mimics the multivariate distribution of the original microarray data and does not rely on strong distributional assumptions. We found that TREAT had the highest overall consistency in each scenario, in agreement with prior studies which found gene rankings based on fold-changes produced the most consistent gene lists. For smaller sample sizes (6 or less per group), the moderated versions of the *t*-test (SAM, eBayes, and TREAT) were superior in terms of both power and consistency relative to the *t*-test and AAA, which combines the *t*-test with a non-parametric approach. The differences in consistency were most pronounced for comparisons between two different data sets for the same disease. However, the AAA method had the highest power for detecting small effect sizes in larger samples (25 or more per group). In contrast TREAT, due to its intentional design of eliminating genes with low log2 ratios from the DEG list, had the lowest power for detecting small effect sizes. Thus, while moderated versions of the *t*-test can generally be recommended for smaller sample sizes, for larger sample sizes selection of a method to detect DEGs may involve a compromise between consistency and power.

Figure 1.

nPOG and nPOGR scores for the five methods for determining DEG lists. Scores are based on comparing DEG lists for subsets of the data with DEG lists for the full data set, where each row represents a different data set, each column a consistency score, and the x-axis represents different subset sizes (increasing from left (Subset A) to right (Subset D)). Points represent mean values while vertical bars indicate the 5th and 95th percentiles from the simulation runs (1000 total simulations). In each case the list length for DEGs was 100, for complete results see Supplementary File B.

Figure 2.

Bar chart showing the distribution of ranks for each method across all combinations of data set, subset size, and list length for consistency of DEG lists. The bottom panel displays consistency scores based on comparisons between the full data set versus a subset of the data, while the top panel is based on comparisons between two DEG lists from two different data sets of the same disease. Bars are color-coded by rank from highest rank (1 = light orange) to lowest rank (5 = dark orange), and height represents the percentage of that rank for the corresponding method. Methods are ordered from left to right according to overall average ranking from highest to lowest. Separate panels are given for nPOG (left) and nPOGR (right) scores.

Figure 3.

nPOG and nPOGR scores for the five methods for determining DEG lists. Scores are based on comparing DEG lists from subsets of two different data sets concerning the same disease, where each row represents a different data set, each column a consistency score, and the x-axis represents different subset sizes (increasing from left (Subset A) to right (Subset D)). Points represent mean values while vertical bars indicate the 5th and 95th percentiles from the simulation runs (1000 total simulations).

In each case the list length for DEGs was 100, for complete results see Supplementary File B. Results are based on using the smaller data set for each disease (first data set listed in each case in Table 3) as the reference data set.

Single Dataset Scores, List Length = 100

t-test ○ AAA △ eBayes + Treat × SAM ◇

Figure 4.

Overall power for the each of the five methods for detecting true DEGs, based on the simulation study. Separate panels are given for each data set.

Table 1

Characteristics of each of the five methods evaluated for determining DEGs in microarray data.

Method	Description
<i>t</i> -test	Ratio of difference in means divided by the standard error of the difference. Robust to violations of normality. May be sensitive to chance fluctuations in the estimate of variability.
SAM [11]	Moderated version of the <i>t</i> -test, which includes a positive constant in the denominator of the <i>t</i> -statistic as a stabilizing factor
eBayes [12, 44]	Moderated version of the <i>t</i> -test similar to SAM, but based on an empirical Bayes approach which averages between the per-gene sample variance and a global (pooled) estimate of the variance
TREAT [16]	Extension to the eBayes method which tests whether differences in gene expression are above a given threshold. Essentially eliminates genes with low log ₂ ratios from the DEG list.
AAA [14]	Averages between a parametric (<i>t</i> -test) and a non-parametric (Wilcoxon rank-sum) test for differential expression.

Table 2

List of data sets used for evaluating the consistency of DE detection methods within the same data set. Data sizes are samples×probes in each case.

Data Set	Raw Size	Used Size	Group (number)	Platform
CHI [23]	128×12,625	79×4,399	BCR/ABL (37), Neg (42)	Affy ¹
ALO [26]	62×2,000	62×2,000	Tumor (40), Normal (22)	Affy
GOL [27]	72×7,129	72×2,400	ALL (47), AML (25)	Affy

¹ Affy = Affymetrix.

List of data sets used for evaluating consisting of DE detection methods between different studies of the same disease. Data sizes are samples×probes in each case.

Table 3

Disease	Data Set	Raw Size	Used Size	Group (n)	Platform
PC ¹	LAP [34]	103×43,008	86×2,072	T ⁵ (54), N ⁶ (32)	cDNA
	SIN [35]	102×12,625	102×2,072	T (52), N(50)	Affy ⁴
LC ²	GAR [36]	18×24,192	18×1,893	T (13), N (5)	cDNA
	BHA [37]	38×12,625	38×1,893	T (21), N (17)	Affy
DMD ³	HAS [38]	24×12,625	24×3,158	DMD (12), N(12)	Affy
	PES [39]	36×22,283	36×3,158	DMD (22), N (14)	Affy

¹PC = prostate cancer,
²LC = lung cancer,
³DMD = Duchenne muscular dystrophy,
⁴Affy = Affymetrix,
⁵T = Tumor,
⁶N = Normal.