BMC Bioinformatics

Research article

Open Access

Leveraging two-way probe-level block design for identifying differential gene expression with high-density oligonucleotide arrays

Leah Barrera[1,2], Chris Benner[1,2], Yong-Chuan Tao[3], Elizabeth Winzeler[1,4] and Yingyao Zhou*[1]

Address: [1]Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, California 92121, USA, [2]Bioinformatics Graduate Program, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, [3]Novartis Institutes for Biomedical Research, 100 Technology Square, Cambridge, MA 02139, USA and [4]The Scripps Research Institute, La Jolla, California 92037, USA

Email: Leah Barrera - lbarrera@bioinf.ucsd.edu; Chris Benner - cbenner@bioinf.ucsd.edu; Yong-Chuan Tao - yong-chuan.tao@pharma.novartis.com; Elizabeth Winzeler - winzeler@gnf.org; Yingyao Zhou* - zhou@gnf.org

* Corresponding author

Published: 20 April 2004

BMC Bioinformatics 2004, **5**:42

Received: 23 December 2003

Accepted: 20 April 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/42

Abstract

Background: To identify differentially expressed genes across experimental conditions in oligonucleotide microarray experiments, existing statistical methods commonly use a summary of probe-level expression data for each probe set and compare replicates of these values across conditions using a form of the *t*-test or rank sum test. Here we propose the use of a statistical method that takes advantage of the built-in redundancy architecture of high-density oligonucleotide arrays.

Results: We employ parametric and nonparametric variants of two-way analysis of variance (ANOVA) on probe-level data to account for probe-level variation, and use the false-discovery rate (FDR) to account for simultaneous testing on thousands of genes (multiple testing problem). Using publicly available data sets, we systematically compared the performance of parametric two-way ANOVA and the nonparametric Mack-Skillings test to the *t*-test and Wilcoxon rank-sum test for detecting differentially expressed genes at varying levels of fold change, concentration, and sample size. Using receiver operating characteristic (ROC) curve comparisons, we observed that two-way methods with FDR control on sample sizes with 2–3 replicates exhibits the same high sensitivity and specificity as a *t*-test with FDR control on sample sizes with 6–9 replicates in detecting at least two-fold change.

Conclusions: Our results suggest that the two-way ANOVA methods using probe-level data are substantially more powerful tests for detecting differential gene expression than corresponding methods for probe-set level data.

Background

The use of DNA microarrays for monitoring the expression levels of thousands of genes simultaneously has generated a stream of methodological and computational challenges. In particular, the reliable identification of differentially expressed genes across different tissues, time points or treatment conditions is the most common and central task in the majority of such experiments [1]. This task has been cast as a multiple hypothesis-testing problem of the simultaneous test for each gene *j* of the null hypothesis of no change in expression level between two or more experimental conditions. Tackling this problem generally involves the following key steps: (1) computing a test statistic for each gene *j*, T_j and determining the significance of each test statistic using parametric assumptions or by appropriate estimation of a null distribution, and (2) employing an appropriate multiple testing procedure to determine which hypotheses to reject while controlling an appropriate error rate [2,3].

A slew of statistical models has been developed to overcome the limitations of the classical *t*-test, rank-sum methods, and other one-way ANOVA methods currently applied to detecting differential gene expression [4]. Under non-normal situations the classical parametric *t*-test is too conservative, and like the Wilcoxon test, with its lack of distributional assumptions, suffers from low power [5]. Non-parametric variants of the *t*-test include the use of permuted data sets to estimate the null distribution of *t*-statistics for each gene [6]. With a small number of replicates, the former method suffers from coarse resolution, resulting in too few or too many genes called differentially expressed depending on the significance threshold. A mixture-modeling approach to calculate the distribution of *t*-statistic type scores has been proposed to overcome that limitation [6]. This approach is similar in spirit to the significance analysis of microarrays (SAM), an increasingly popular method which also uses a *t*-statistic type score [6]. SAM uses permutations of repeated measurements and then pools estimated null statistics for each gene to compute an overall error rate defined as the false discovery rate (FDR) for genes identified as differentially expressed [2,7].

The false discovery rate (FDR) is the rate at which features called significant are truly null. Here, it is the expected proportion of genes erroneously identified as differentially expressed. The control of the FDR as a multiple testing procedure was proposed by Benjamini and Hochberg as a more powerful alternative to controlling the family-wise error rate (FWER) when considering multiple null hypotheses simultaneously [8]. Control of the FDR implies control of the FWER when all the null hypotheses are true [9]. Bonferroni type procedures which control FWER are considered too stringent because they control the probability of making any Type I error among the hypotheses under consideration, thus rejecting too few hypotheses when identifying differentially expressed genes [3]. On the other hand, control of the FDR has been increasingly favored for high-throughput screenings such as microarray experiments, striking a balance between FWER control and the per-comparison-error-rate (PCER) control which often yields too many false positives.

The persisting high cost of microarrays, in particular of commercial high-density oligonucleotide arrays (HDAs) such as the Affymetrix GeneChip, and the scarcity of samples in many experiments, continue to severely limit the number of replicates used per condition, and thus restrict the potential gain in statistical power of the statistical methods described above with increasing sample size. In addition, the statistical methods described above are generally applied to experiments using both cDNA microarrays and HDAs. The differences in design between the two microarray platforms have warranted different algorithms for aspects of array analysis such as gene expression level calculation, image analysis, and normalization [10].

In this light, instead of developing a new statistical method that can be generally applied to experiments using both cDNA microarrays and HDAs as those described above, we can leverage the unique design of HDAs for better differential gene expression identification. On HDAs supplied by Affymetrix, 11–20 25-base oligonucleotide probes that are exact complements to different fragments of the same gene target form a probe set. Unlike cDNA microarrays, where a single intensity ratio is collected for each gene, 11–20 probe-level measurements per probe set are collected simultaneously for any single array hybridization. However, these redundant measurements are typically summarized as one value in the form of an average difference (AD) or model-based expression index (MBEI) for the purpose of statistical analysis [11]. Using probe-level measurements in identifying differentially expressed genes and blocking on the probe in an analysis of variance (ANOVA), combined with FDR adjustment for the multiple testing problem, are the key differences between our proposed approach and previously described related methods.

Although carrying out statistics at the probe-level immediately increases the sample size by at least an order of magnitude, it is warranted due to the large and systematic differences that are known to exist among probes that survey the same gene [11]. Due to these probe-specific biases, variation induced by probes is larger than that induced by array replicates [12]. The use of the probe as a blocking factor in testing for differential gene expression in a two-way ANOVA on probe-level data is thus expected to be more sensitive than previously described methods.

Chu et al. also took an ANOVA approach at the probe level, however the experimental design of their study was different from ours, which led to a more complicated model than what we propose here [13]. Chu et al. compared their method to SAM on the same data set, but identified a very different set of differentially expressed genes (Table 3 in [13]). As pointed out by other researchers, this method cannot be easily benchmarked only based on data sets of unknown positives and negatives [14]. Lemon et al. recently proposed a probe-level Logit-t method that was shown to be superior to other popular probe set methods [14]. Independent from these two studies, we reached the same conclusion that using probe-level data could significantly improve the quality of resultant gene list. In addition, we demonstrate the use of a rank-based Mack-Skillings test, which does not depend on any distribution models required by the two parametric studies mentioned above. Furthermore, by using an FDR-based criterion, our method not only ranks genes but also suggests statistically rigorous thresholds for gene selection.

In this study, we compared both the sensitivity and specificity of parametric two-way ANOVA and the nonparametric Mack-Skillings test on probe-level data against the commonly-used *t*-test and Wilcoxon test on probe-set level data. For all tests, we employed FDR-controlling procedures described above to account for the multiple testing problem. Two public data sets are used for benchmarking purposes: the Lemon set, where thousands of genes are expected to be differentially expressed and the Affymetrix Latin-square data set where only 14 spiked genes out of over 9000 genes on the array are expected to show real change [15-17]. We systematically tested the effects of key factors such as expression level (concentration) of the RNA transcripts, number of replicates, amount of change to be detected, in addition to the statistical methods. In almost all cases, the proposed probe-level methods outperformed previous methods based on probe-set level calculations. We also found that the two-way methods are most sensitive to transcript concentration between 4 pM and 128 pM and fold change greater than two. By comparing receiver operating characteristic (ROC) curves, we demonstrated that by taking advantage of the HDA design, the two-way methods applied on only 2–3 replicates can exhibit the same high sensitivity and specificity as a SAM-like *t*-test with FDR-control using 6–9 replicates for detecting at least two-fold change. Therefore, by taking advantage of the HDA design, the present limitations of one-way ANOVA-type methods can be overcome. Matlab scripts for our methods are available on http://carrier.gnf.org/publications/ProbeStatistics.

Results

We compared the performance of the commonly used one-way ANOVA methods described above, against the two-way ANOVA methods using two publicly available microarray data sets. The first set of microarray experiments involves groups of human fibroblast cells in three conditions – serum starved, serum stimulated, and a 50:50 mixture of starved/stimulated – with six replicate Affymetrix HuGeneFL arrays in each group [16]. For this set, a total of 7011 probe sets were examined per array after the preprocessing steps. The second data set is the Affymetrix Latin Square Data for Expression Algorithm Assessment [17]. In 11 experiments (denote these as experiments A-K), 14 groups of human gene transcripts in 14 different known concentrations were spiked into a background RNA mixture and hybridized to 3 replicate microarrays. In two additional experiments (denote these as experiments L and M), the same Latin Square design is followed but 12 instead of 3 replicates were used per condition. In the following study, we tracked only 12 of 14 genes due to errors in the original data set for two of the probe sets. Transcript concentrations for each spiked gene ranged from 0 to 1024 pM over the various experiments [15]. For this data, the Affymetrix HG_U95A array is used and a total of 9024 probe sets in each array were analyzed as described in the following sections after the preprocessing step.

Sensitivity

We first assessed the relative sensitivity of the statistical tests by comparing the number of genes identified as differentially expressed when controlling the FDR using either an LSU procedure or a resampling-based approach. We compared the serum starved and serum stimulated data sets between which, the expression levels of a large number of genes were expected to vary significantly [16]. We randomly sampled three replicates per condition to make the results comparable to later analyses for which only three replicates are available. The process was repeated 100 times and results were averaged.

We found that the parametric two-way ANOVA combined with an LSU FDR-controlling procedure identified the largest number of genes at varying levels of q. The nonparametric two-way method, the Mack-Skillings test, combined with the LSU-procedure also identified a significantly greater number of genes compared with the *t*-test and the Wilcoxon test (Fig. 1). The Wilcoxon-test lacked the sensitivity to identify any genes differentially expressed at a reasonable FDR, while the *t*-test performed in the intermediate range.

The use of a resampling-based FDR-controlling procedure decreased the number of genes called differentially expressed by nearly half when using the two-way ANOVA and Mack-Skillings tests while not significantly altering the number called when using the *t*-test. Despite this decrease, the two-way ANOVA methods remained more

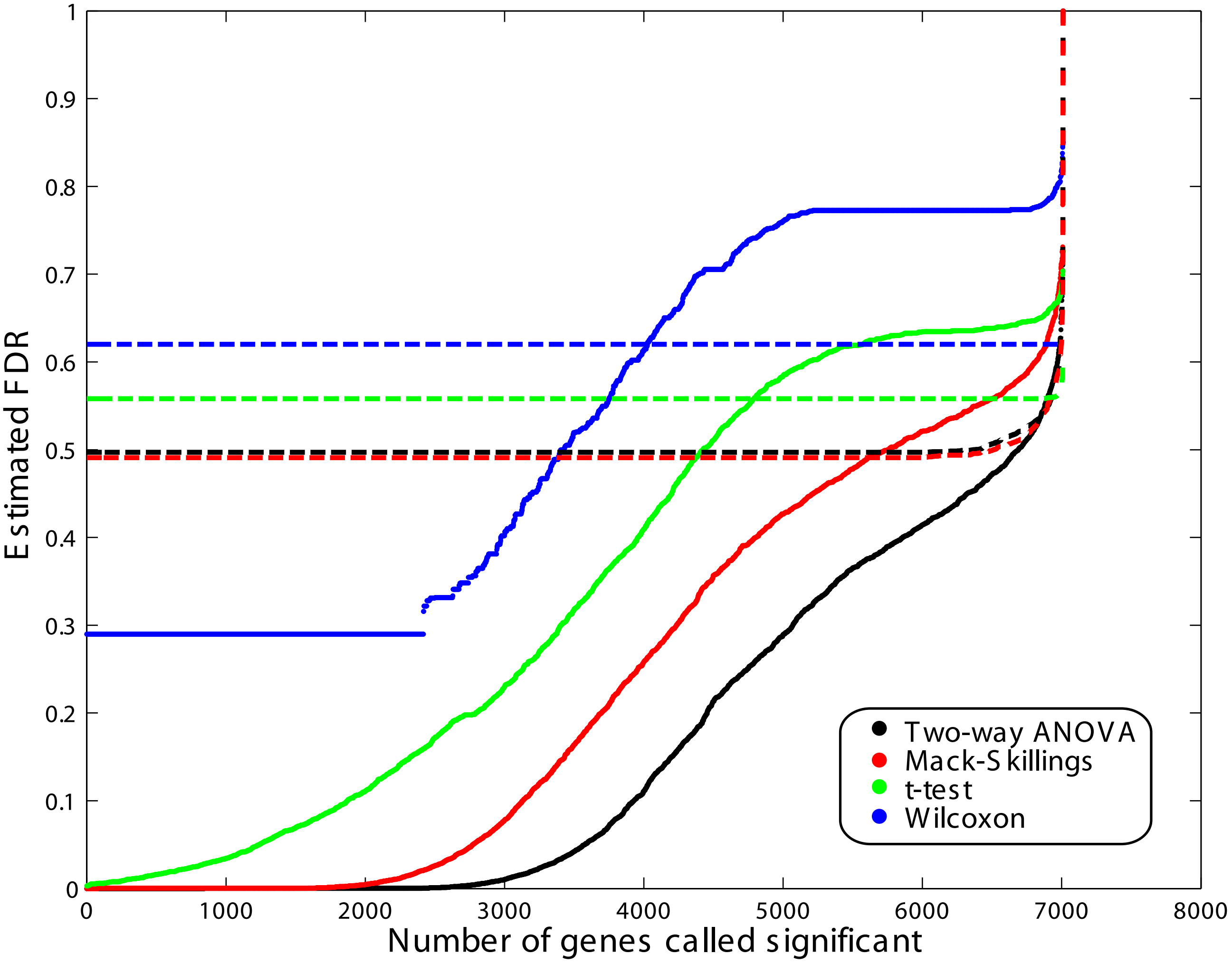

Figure 1
Number of probe sets called significant versus LSU-adjusted FDR for all methods. Dashed lines indicate the control versus control comparisons.

sensitive at typical levels of control such as q = 0.05 (Fig. 2). This decrease can be attributed to the extreme sensitivity of the two-way methods and the large number of genes with differential gene expression limiting the success of resampling-based methods for estimating a good null distribution.

To assess whether we are detecting biologically meaningful change, we also applied the statistical tests on random pairs of combinations of data values within the same treatment condition, i.e. comparing serum starved samples or serum stimulated samples among themselves, respectively. The dashed lines in Fig. 1 and Fig. 2 show that as expected all methods do not identify any genes as differentially expressed within a reasonable level of FDR control. The result ensures that the extra sensitivity of the proposed two-way methods was not gained at the expense of sacrificing robustness. The specificity of the methods will be further studied later.

To study the necessity of explicitly modeling the probe-treatment interaction in our ANOVA model, F-tests were applied to all genes. 5151 out of 7009 genes (73%) tested had P-values less than 0.05, even after a Bonferroni adjustment. A possible explanation is that the interaction term captures the changes in probe cross-hybridization properties caused by the large differences in the mRNA content between the two sample groups.

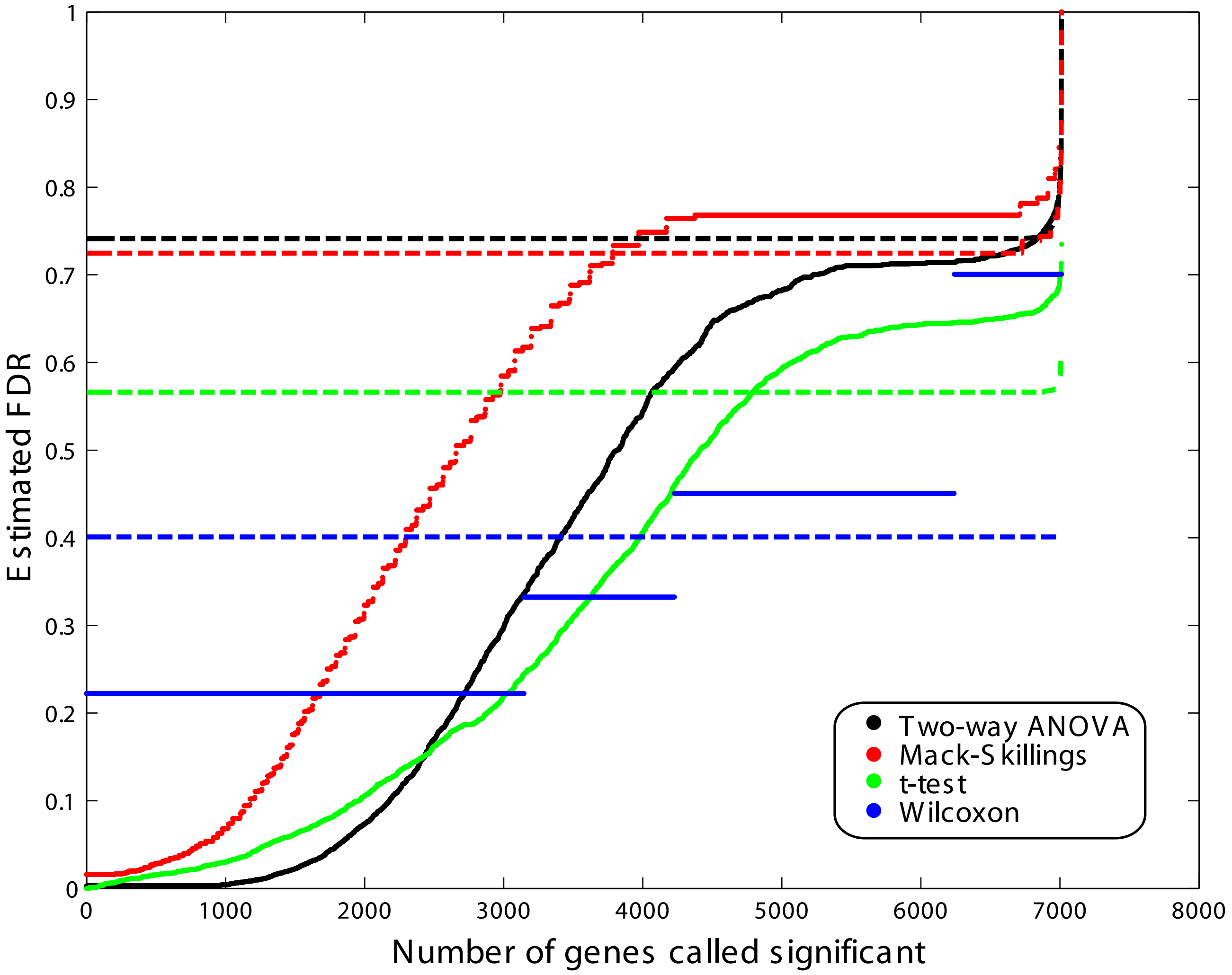

Figure 2
Number of probe sets called significant versus resampling-based FDR for all methods. Dashed lines indicate the control versus control comparisons.

Hoffmann *et al.* studied the consistency between various one-way methods and found the differentially expressed genes identified by SAM is approximately a subset of genes identified by nonparametric methods, which is in turn a subset of *t*-test results [18]. To check the agreement among the various methods described above, we also compared the proportion of genes that can be detected simultaneously in the Lemon data set when comparing combinations of statistical methods and FDR controlling procedures. When controlling the LSU FDR at level q = 0.05, 3576 genes are called differentially expressed using parametric two-way ANOVA; 2773, Mack-Skillings test; 1238, *t*-test; and none, Wilcoxon-test. In Table 1, we show the proportion of identical genes called significant by pairs of methods. Percentages are relative to the methods defined in the column headings. These results demonstrate general agreement in calls of differential gene expression among the different methods, with slightly higher correspondence between the two-way methods.

A representative run of the probe-level Logit-t method [14] on the Lemon data set identified 1032 genes as differentially expressed when controlling the LSU-FDR at level q = 0.05 as above – less than half the number identified by the two-way methods. The Logit-t method

Table 1: Percentage of identical genes called significant by pairs of procedures

	Mack-Skillings; LSU	*t*-test; LSU	Wilcoxon; LSU
two-way;LSU	96	93	--
Mack-Skillings; LSU		88	--
t-test; LSU			--

Percentages are relative to the method defined in the column heading. -- No genes passed the criteria of FDR control at q = 0.05.

demonstrated a level of sensitivity similar to *t*-test (Fig. 8) [see Additional File 1].

Effect of concentration and fold change

Figs. 1 and 2 highlighted the greater sensitivity of two-way methods using the Lemon data set in which the expression of a large number of genes were expected to change between the starved and stimulated conditions. However, the identities of these true positives and the magnitudes of relative and absolute change are unknown. Using the set of 11 experiments with 3 replicates each from the Affymetrix Latin Square Data Set; we examined the effects of known concentration and fold-change on the sensitivity of the tests coupled with the LSU FDR-controlling procedure. We did 55 pairwise comparisons of the 11 experiments giving a wide range of fold change and maximum spike-in concentration combinations (Fig. 3). As expected, increasing fold change combined with increasing maximum spike-in concentration allows for better detection using all methods.

With as little as three replicates, we see in Fig. 3 that the parametric two-way ANOVA and the Mack-Skillings test are very sensitive to two-fold changes when testing within a maximum concentration range of 4 to 128 pM (Fig. 3a). With a four-fold change, the two-way methods are able to successfully detect nearly all spiked gene transcripts in all pairs of experiments at FDR level q = 0.05 with the exception of one changing from 0.25 to 1 pM (Fig. 3b). Only with an eight-fold change and maximum concentration between 32 and 128 pM do we begin to detect the spiked genes when using the *t*-test and controlling the FDR at q = 0.05 (Fig. 3c). These differences may explain the higher sensitivity of the two way methods shown in Figs. 1 and 2.

The decrease in sensitivity at higher concentrations over the lower fold changes for all methods prompted an investigation of the associated expression values used. A log-log plot of the average difference values against the known concentrations of the spiked transcripts (Fig. 4) suggests a nonlinear relationship between measured intensities and the spiked concentrations at the higher concentrations supporting the lack of sensitivity of all tests in that range. The nonlinear relationship at the low concentration range is mainly caused by hybridization noise. Despite low signal-noise ratio in this range, the two-way methods can consistently detect two-fold changes for a gene concentration as low as 4 pM whereas the SAM-like one-way methods were generally unsuccessful for the range of concentrations at this fold change for the data set (Fig. 3).

Specificity and Sample Size Effect

The clearly higher sensitivity of the two-way methods in discriminating a wider range of fold change at various transcript concentrations with as little as three replicates (Fig. 3) prompted the simultaneous evaluation of sensitivity and specificity using receiver-operator characteristic (ROC) curves. We compared the ability of the various methods to discern a known two-fold change over the range of concentrations in experiments L and M of the Latin Square Data set using only three replicates. We obtained results for sample size n = 3 by computing the adjusted FDR from the average p-value for each probe set over 100 comparisons of random pairs of samples of size n taken from each condition. For these data, Fig. 5 shows that the two-way ANOVA methods combined with the LSU FDR-controlling procedure clearly outperform the one-way statistical tests and does not trade off sensitivity for specificity. For the parametric two-way ANOVA, the ROC curve indicates a 91% sensitivity with a 99.84% specificity. In other words, we expect to find 11/12 spiked genes with only 14 false positives in this data set. The Mack-Skillings test follows with 75% sensitivity at the same specificity range, whereas the *t*-test and the Wilcoxon test clearly lack power under those conditions. These results suggest that with the same number of replicates, the improved sensitivity of two-way methods (Figs. 1, 2, 3) is due to the accurate detection of lower fold changes at a wider range of concentrations.

After observing the relative lack of power of the one-way methods compared with the two-way methods with only three replicates per sample, we systematically assessed the extent of the sample size effect on the relative power of these tests in detecting the same two-fold change. Using the same pair of Latin Square Data experiments with 12 replicates each, we compared the performance of the four

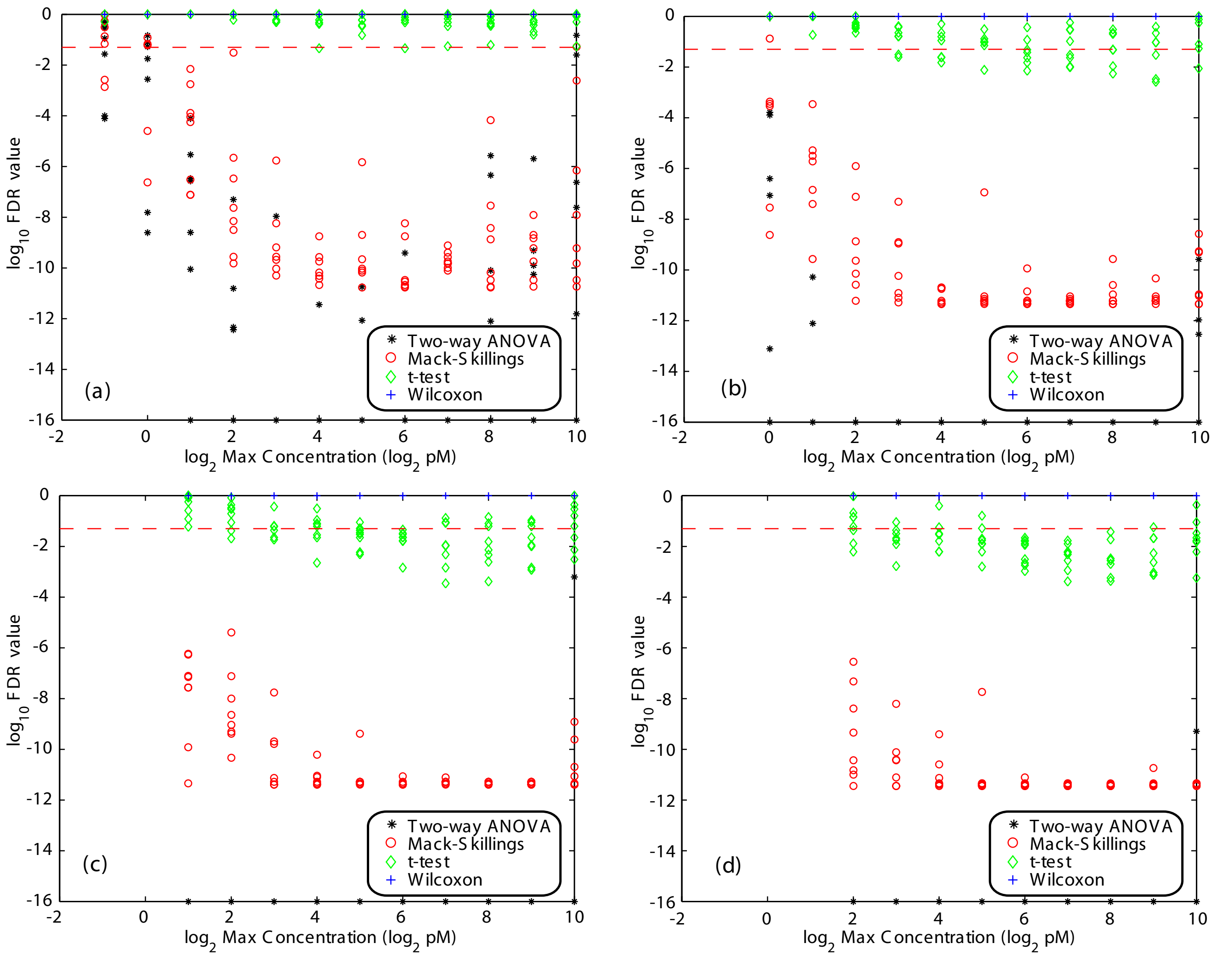

Figure 3

log-log plots of FDR versus the maximum spike-in concentration at varying levels of fold change (FC). The dashed line in each plot is the log FDR value corresponding to *q* = 0.05. Plots for higher fold changes are available at our web site. (a) FC = 2. (b) FC = 4. (c) FC = 8. (d) FC = 16. Due to the precision of the Matlab routines used for this study, log FDR values below -16 were cut off at -16.

methods using additional sample sizes of *n* = 2, 6, 9, 12. The adjusted FDR values for *n* = 2, 6, 9 were computed in a similar way as for *n* = 3. As shown by the ROC curves in Fig. 6(a), the two-way methods exhibit relatively high sensitivity and specificity when applied to data from as little as two replicate experiments. Visual comparison suggests that as many as 9 replicates may be needed using the *t*-test to attain the same high power exhibited by the two-way ANOVA using as little as three replicates (Fig. 5, Fig. 6). This is not a surprising result because the two-way ANOVA methods take advantage of probe information which effectively increase the sample size by an order of magnitude.

We show the effect of sample size on the resulting FDRs given by each test on some representative probe sets representing spiked genes in Fig. 7. Similar plots for all spiked probe sets are available as Supplementary Material http://carrier.gnf.org/publications/ProbeStatistics. Note the expected decrease in FDR with increasing sample size using all methods, and the slight difference between Fig. 7(a) and Fig. 7(b) due to the effect of the maximum and

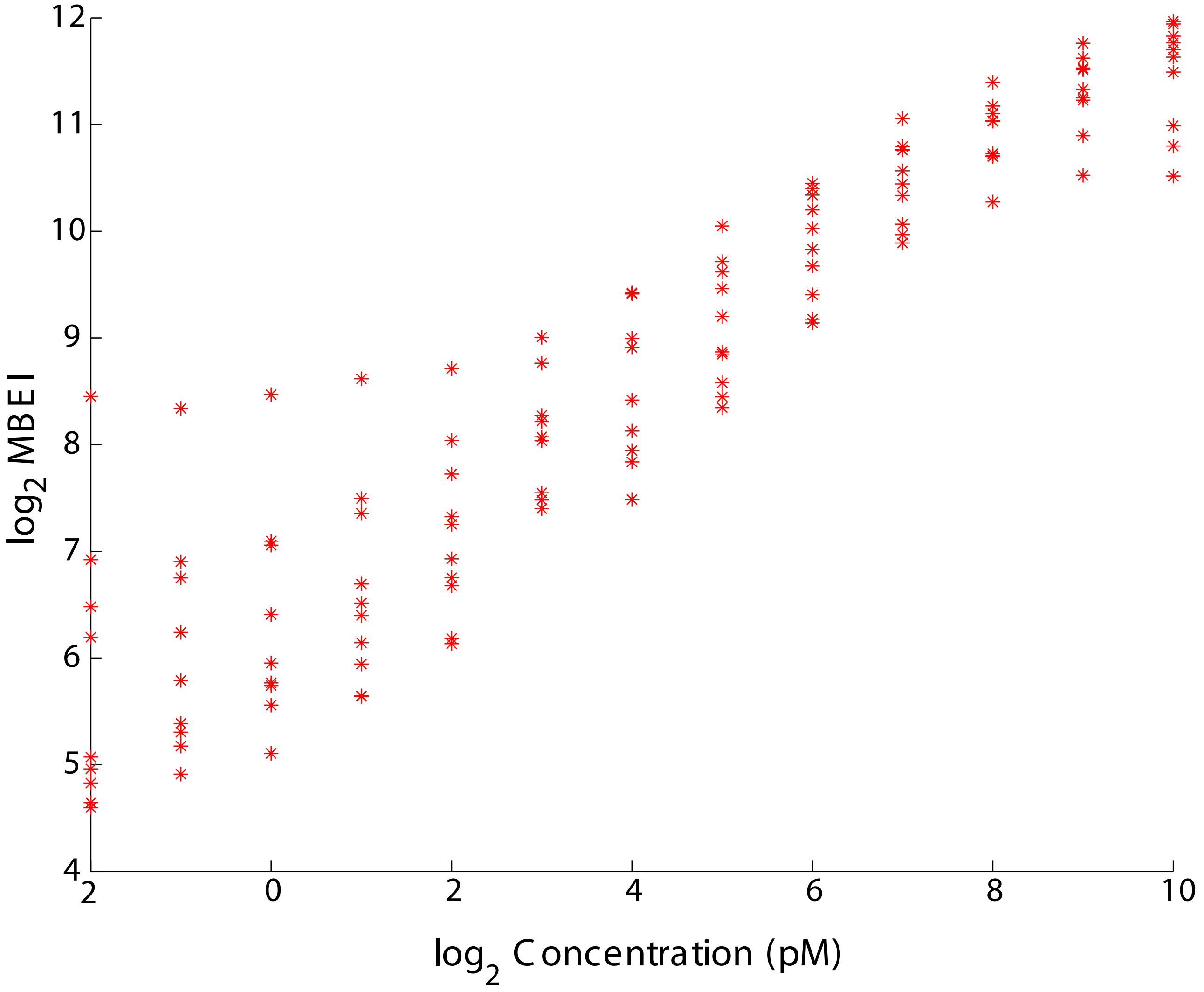

Figure 4
Comparison of the estimated expression values against the known spiked transcript concentrations.

minimum concentration on the absolute magnitude of the two-fold change to be detected.

Discussion

In the previous sections, we compared the application of two-way ANOVA methods on probe-level data to standard statistical methods on probe-set level data in identifying differential gene expression in microarray experiments. We aimed to show the importance of leveraging HDA design in the choice of statistical test and not discarding information by working with a probe-set summary or average of probe-level data.

Using two-way ANOVA methods, we systematically accounted for probe-specific biases in hybridization or measurement efficiency, and thus achieved higher sensitivity and specificity compared with the *t*-test in the range of conditions investigated with varying levels of sample size, fold change, and maximum spike-in concentration. In the Lemon serum-starved and serum-stimulated data set, the two-way methods coupled with LSU-FDR control identified more than twice as many genes as differentially expressed compared with the *t*-test. With the Latin Square data set, we confirmed the specificity of the two-way methods by analyzing the ROC curves and observed that with as few as three replicates, the two-way ANOVA has a 91% sensitivity with a 99.84% specificity.

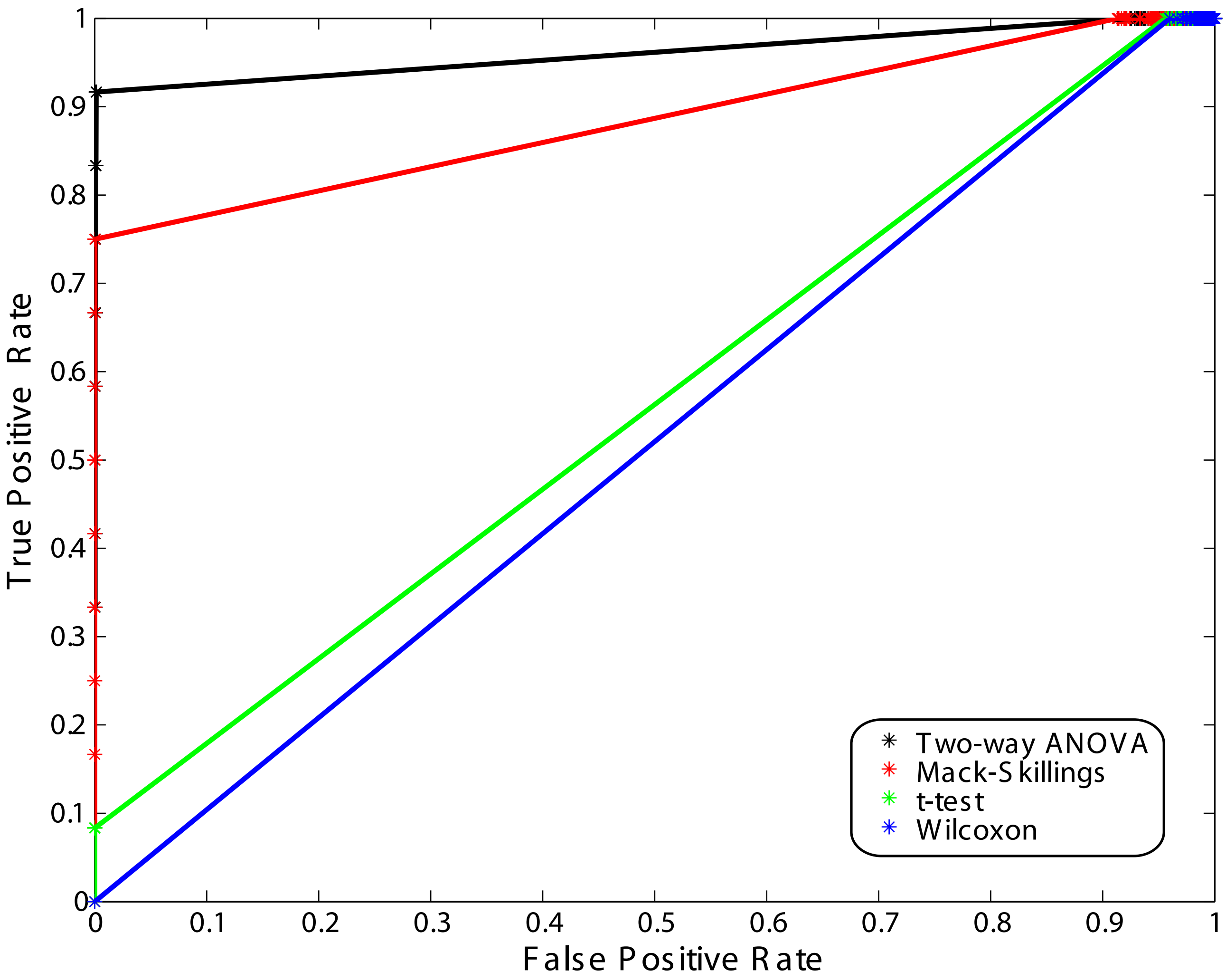

Figure 5
ROC curve comparing the power of each method when sample size, *n* = 3.

Parametric methods are commonly criticized for their lack of sensitivity and specificity when detecting differential gene expression. However, we discovered that the use of an LSU FDR-controlling procedure with the parametric two-way ANOVA method yielded the most promising results in terms of higher sensitivity and specificity for detecting differentially expressed genes. The outstanding performance of the parametric two-way ANOVA with the LSU FDR-controlling procedures relative to the other combinations of nonparametric tests and the resampling-based FDR in our study suggests that in the case of gene expression analysis with HDAs, there is a substantial gain in power by working with probe-level data, and that proper treatment of this data by appropriate normalization procedures and the application of appropriate transformations (logarithm, square root) can allow us to maintain assumptions critical to the method chosen.

Even without parametric assumptions, the advantage of treating the probe as a blocking factor was clearly demonstrated by the results using the Mack-Skillings test. Thus, if more conservative estimates from a two-way ANOVA analysis are desired, we can choose to use the results from the nonparametric Mack-Skillings test and still have a substantial gain in power over the *t*-test. The same Affymetrix Latin Square data set has been recently studied by Lemon et al. using a probe-level Logit-t method and a low false positive rate of 0.03% was achieved at the sensitivity of 87% [14]. The parametric two-way ANOVA achieved essentially the same performance and the nonparametric

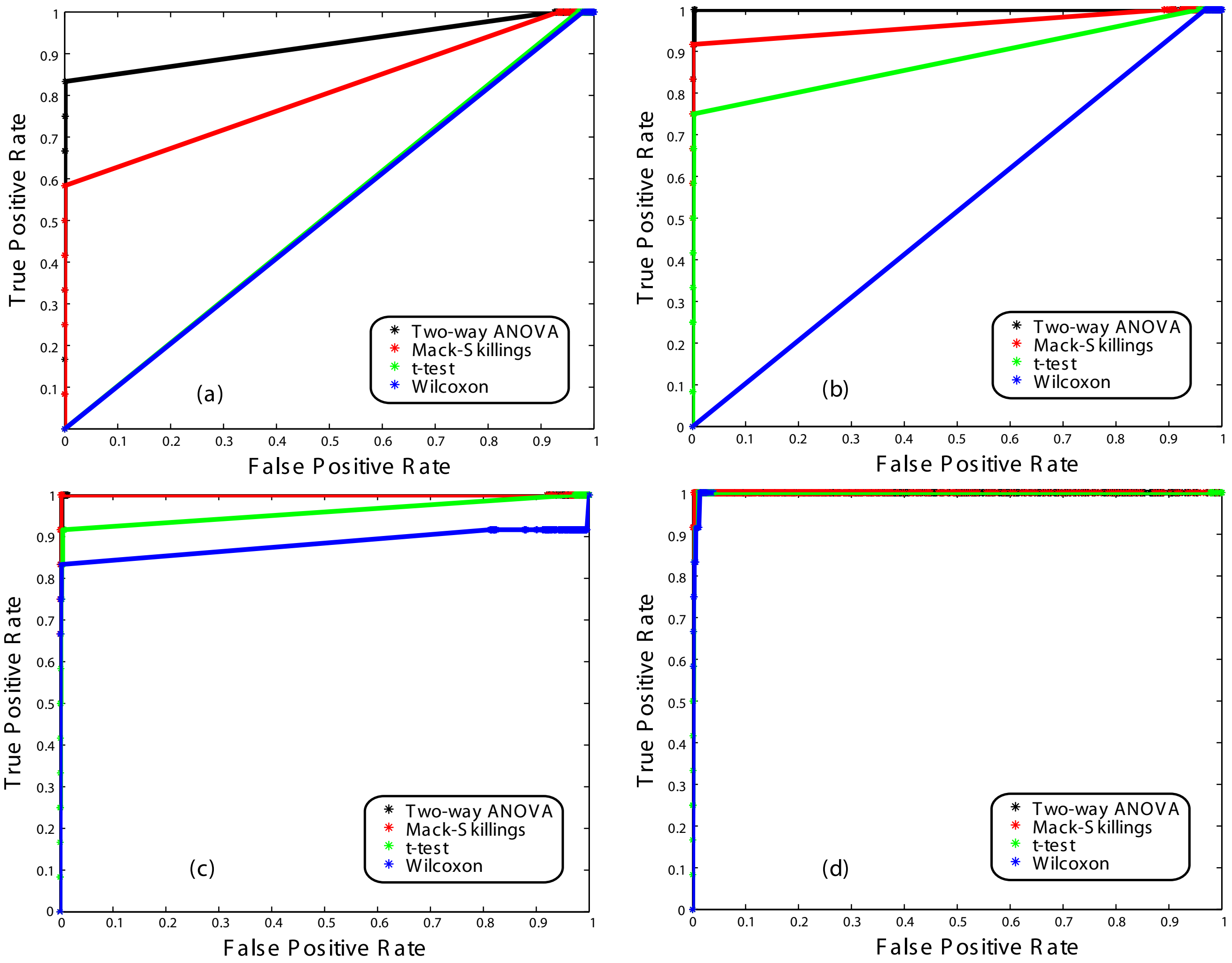

Figure 6
ROC curves highlighting the effect of sample size on the power of each method. (a) *n* = 2. (b) *n* = 6. (c) *n* = 9. (d) *n* = 12.

Mack-Skillings method showed an even better false positive rate of 0.01% with the same sensitivity http://carrier.gnf.org/publications/ProbeStatistics.

The power of a statistical test is a function of its sensitivity and this further depends on (1) the magnitude of the real difference to be measured, (2) the noise level or standard deviation of sample measurements, (3) the significance level at which the tests are done, and (4) the sample sizes [19]. Limitations inherent to the technology platform and suboptimal data preprocessing procedures can reduce the magnitudes of the real differences being measured. As we observed, there is a nonlinear relationship between expression values and the actual spike-in concentrations at the lower and higher ends of the concentration spectrum for the Latin Square data set due to detection and measurement saturation issues [20]. With the use of probe-level data in a two-way ANOVA, we take advantage of informative probe-level differences between treatments and eliminate noise due to probe efficiency differences. In this way, the two-way ANOVA methods are better able to discern treatment differences. Control of the third factor depends on the number of false leads that one is willing to incur and this in turn varies with the goals of the experiments. Finally, the control of the fourth factor is limited by resource constraints, and in microarray experiments, this continues to be a key issue due to the costs of microarrays (two to three per condition in most labs) and availability of samples to be analyzed.

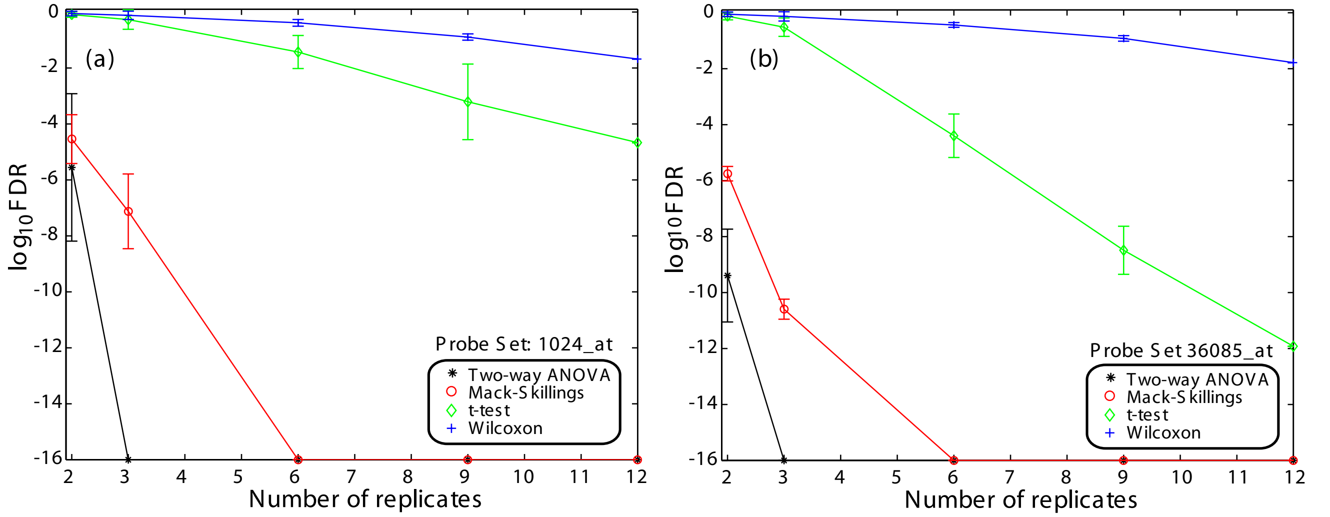

Figure 7
Effect of sample size on the log FDR value for representative probe sets. (a) 1024_at (b) 36085_at. Due to the precision of the Matlab routines used for this study, log FDR values below -16 were cut off at -16.

It is well-known that increasing sample size increases sensitivity for all statistical tests, and given enough samples, one can discern biologically meaningful changes well below the differences currently measured. That we can detect differentially expressed genes using two-way ANOVA methods with only two or three replicates and get comparable results with the use of the *t*-test on at least 6–9 replicates is evidence of the higher power of these methods on probe-level data all other factors being equal.

In this study, we have shown that coupled with an easily implemented linear step-up (LSU) FDR-controlling procedure, parametric and nonparametric two-way ANOVA methods using probe-level data are substantially more powerful tests than standard methods applied to probe-set level data for detecting differential gene expression. Their advantage in power is especially pronounced when working with samples with as few as two or three replicates – the most common sample sizes for microarray experiments [1]. Although we only examined two sets of conditions in our data sets, the two-way ANOVA is a general design which easily handles other array experiment setups with two or more levels of treatments or time series points. As a well-known and extensively used statistical method in many fields, the two-way ANOVA has inspired a body of literature for dealing with many special cases, such as unequal group sizes due to missing data from replicates, which frequently occur in microarray experiments [21,22]. Clearly, the ease of implementation of the two-way ANOVA-type methods coupled with LSU-FDR control, and the results shown herein, strongly suggest its use and further development for identifying differentially expressed genes.

Methods

For the following study, we use methods focusing on the two-sample case. We briefly describe the four well-known statistical tests and the two forms of FDR control employed in our study. Since the parametric statistical tests require the key assumption of equal group variances, logarithms of probe-level intensities and summarized expression values were taken to provide a better approximation [23].

Statistical tests

Parametric t-test

The *t*-statistic and its variants are powerful measures for detecting differential expression because they permit selection of genes with maximal difference in mean level of expression between two groups and minimal variation of expression within each group [4]. Here we employ the classical *t*-test which is a statistically equivalent test to the parametric one-way ANOVA in the two-sample case [22]. As done in Reiner *et al.*, we obtain the *p*-values directly using the *t*-distribution with appropriate degrees of freedom depending on the sample sizes [9].

Wilcoxon rank sum test

For each gene, the distribution-free rank sum test transforms the sorted gene expression values across experiments into ranks and then tests the null hypothesis of equality of the means of the ranked values between experimental conditions [6]. For small sample sizes, exact *p*-values can be obtained from pre-calculated statistical tables. A normal approximation of standardized test statistics is typically used to obtain *p*-values for larger sample sizes. In this case, it was used for samples of size 9 and greater.

Two-way ANOVA

The use of ANOVA in testing for the equality of group means relies on the computation of the ratio of the mean square variation among group means to the mean square variation within groups. A large ratio indicates a significant difference between group means. The one-way ANOVA model, a generalization of the *t*-test reliably detects differences between group means only when other factors, which can cause large variation within groups, are controlled.

In the case of HDAs, probe-level intensities are a source of large and systematic variation. Thus, instead of using the summarized expression indices for each probe set for hypothesis testing and ignoring individual probe effects, we use intensity values for each probe in a probe set and control for probe-specific biases by considering probe type as a blocking factor in a two-way ANOVA. For each probe set, replicate measurements of log-transformed probe-level intensities for each probe are segregated into blocks across the treatment conditions. Two types of hypothesis tests can be performed in this case: (1) the test of the equality of probe or block means to assess the significance of explicitly modeling probe-level effects, and (2) the test of equality of treatment means having accounted for variation caused by individual probes. The ANOVA model is:

$$Y_{ijk} = \mu + P_i + T_j + PT_{ij} + \varepsilon_{k(ij)},$$

where Y_{ijk} is the logarithm of the probe-level intensity measurement, μ is the overall mean, P_i is the effect of the probe i, T_j is the effect of treatment j, PT_{ij} is the effect of the interaction between the probe i and treatment j, and $\varepsilon_{k(ij)}$ is the error. The probe-treatment interaction term is necessary based on our results on the Lemon data set (see Results for details).

In the first test we can measure the ratio of the mean square variation among blocks to the mean square variation within groups, where each group is a treatment/block combination. The significance of these probe-level differences have been documented and were again confirmed by the extremely low *p*-values associated with block effects in our study [11]. However, it is not of particular interest that measured intensities for probe A differ significantly from those of probe B in a probe set when testing for differential gene expression [12]. Here we only measure the amount of such fluctuations and remove it from the estimate of within group variability. In the second test, the test of interest for identification of differential gene expression, we measure the ratio of the mean square variation among treatments to the mean square variation within treatment/block groups. The *p*-values corresponding to the ratios for the second test are determined using an *F*-distribution whose numerator has degrees of freedom equal to k-1 where k is the number of treatments, and whose denominator has $pk(r\text{-}1)$ degrees of freedom, where p is the number of probes in the probe set and r is the number of replicates. In this study, we maintain the assumption of equal group sizes because there are corresponding probes for each probe set across experimental samples profiled using the same array type, and in the data sets used, there are equal numbers of replicates [22,23]

Mack-Skillings test

This distribution-free alternative to the classic two-way ANOVA model above transforms the probe-level intensities into ranks for each probe across the samples (replicates and conditions). It is a generalization of the nonparametric Friedman test when there are replicates. This test of no change across experimental conditions uses the Mack-Skillings statistic to measure the squared deviation of the sum of the ranks across the probes in a probe set for each treatment condition, from the expected sum based on no treatment differences. As with the Wilcoxon test, the exact *p*-values for small sample sizes can be found in statistical tables or computed numerically. Large-sample approximation allows the estimation of *p*-values using a chi-square distribution with k-1 degrees of freedom, where k is the number of experimental conditions [21].

FDR control

Linear step-up (LSU) procedure

The linear-step up (LSU) procedure originally described by Benjamini and Hochberg (1995) controls the FDR rate at level q by rejecting all hypotheses $H_{(i)}$, $i = 1,...,k$ where $k = \max\{i : P_{(i)} = \frac{i}{m}q\}$ and the $P_{(i)}\}$ are the ordered *p*-values. Here, we compute the multiplicity adjusted *p*-values:

$$P^{LSU}_{(j)} = min_{j \le i}\left\{P_{(i)}\frac{m}{i}\right\},$$

and thus associate an FDR for each hypothesis test [9].

Resampling-based procedure

Resampling-based methods seek to gain more power by utilizing the empirical dependency structure of the data to construct more powerful FDR-controlling procedures [3,9]. Here we generate an $m \times n$ matrix of resample-based p-values $[p_{ik}]$ for m probe sets using n permutations of treatment labels (n = 100 in this study). We naively estimate a resampling-based FDR for each probe set by ordering the observed p-values P_j and starting with the largest p-value $P_{(m)}$ we compute:

$$P_{(j)}^{FDR} = min\left\{\frac{V}{R}, P_{(l)}^{FDR}, 1\right\}, l > j$$

$$\text{where } V = \frac{1}{n}\sum_{k=1}^{n} \#\left\{p_{ik} \le P_{(j)}\right\}$$

$$R = \#\left\{P_{(i)} \le P_{(j)}\right\}$$

V is the average number of assumed null p-values from all permutations as extreme as the observed value under consideration, whereas R is the number of observed p-values as extreme as the same value under consideration. The ratio of these values gives an estimate of the FDR associated with the rejection of the hypothesis under consideration.

The statistical tests described above were performed using Matlab. Built-in Matlab functions were used to compute the test statistics and associated p-values, and FDR adjustments were implemented as described above.

Data preprocessing

Microarray intensity normalization and gene expression calculations were performed using dChip [11]. Probe values were first normalized and their background intensities subtracted. Probe set expression values were computed using the PM-only model for expression using standard outlier detection. An additional normalization step was used to adjust the probe set expression values of each array to a median expression level of 200. Aside from the previously published advantages for using only PM probes intensity calculations using only PM probes tend to result in higher values and few if any negative values, alleviating complications when log transforming the data [11,10]. In addition to the preprocessing using dChip, we also filtered the probe sets so that at least one sample group has an average expression level of 20. This is done in order to prevent comparing expression levels of genes that are either insignificantly expressed in both treatment conditions or are expressed at the noise level.

Abbreviations

ANOVA: Analysis of variance

FDR: false discovery rate

HDA: high-density oligonucleotide array

ROC: receiver operating characteristic

LSU: linear step-up procedure

Authors' contributions

LB conducted most of the analyses presented here and drafted the manuscript. CB conducted most of the preliminary analyses for testing the ideas. TC conceived of the idea of using two-way tests. EW provided initial data sets for testing. YZ conceived and coordinated the study. All authors read and approved the final manuscript.

Additional Files

Additional file 1

File name Figure 8

File type PDF

Description of the file: Number of probe sets identified by Logit-t method in the Lemon data set.

Number of probe sets called significant versus LSU-adjusted FDR in the Lemon data set computed with t-test, Wilcoxon test, Logit-t method, parametric two-way ANOVA and nonparametric Mack-Skillings method. Dashed lines indicate the control versus control comparisons.

References

1. Pan W: **A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.** *Bioinformatics* 2002, **18:**546-554.
2. Storey JD, Tibshirani R: **SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays.** In *To appear in The Analysis of Gene Expression Data: Methods and Software. Springer, New York* 2003.
3. Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Stat Sinica* 2002, **12:**111-139.
4. Troyanskaya O, Garber ME, Brown PO, Botstein D, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data.** *Bioinformatics* 2002, **18:**1454-1461.
5. Pan W, Lin J, Le CT: **How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach.** *Genome Biol* 2002, **3:**research0022.1-0022.10.
6. Pan W, Lin J, Le C: **A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data.** *To appear in Functional & Integrative Genomics 2003. (Also Report 2003–004, Division of Biostatistics, University of Minnesota)* 2003.
7. Tusher V, Tibshirani R, Chu G: **Significance analysis of microarrays applied to transcriptional responses to ionizing radiation.** *Proc Natl Acad Sci USA* 2001, **98:**5116-5121.
8. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Stat Soc B* 1995, **57:**289-300.
9. Reiner A, Yekutieli D, Benjamini Y: **Identifying differentially expressed genes using false discovery rate controlling procedures.** *Bioinformatics* 2003, **19:**368-375.

10. Zhou Y, Abagyan R: **Algorithms for high-density oligonucleotide array.** *Curr Opin Drug Discov Devel* 2003, **6:**339-345.
11. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection,.** *Proc Natl Acad Sci* 2001, **98:**31-36.
12. Yang Y, Hoh J, Broger C, Neeb M, Edington J, Lindpainter K, Ott J: **Statistical Methods for Analyzing Microarray Feature Data with Replications.** *J Comput Biol* 2003, **10:**157-169.
13. Chu TM, Weir B, Wolfinger R: **A systematic statistical linear modeling approach to oligonucleotide array experiments.** *Math Biosci* 2002, **176:**35-51.
14. Lemon WJ, Liyanarachchi S, You M: **A high performance test of differential gene expression for oligonucleotide arrays.** *Genome Biology* 2003, **4:**R67.
15. *Affymetrix Latin Square Data for Expression Algorithm Assessment: Affymetrix Inc., Santa Clara, CA, USA* [http://www.affymetrix.com/analysis/download_center2.affx].
16. Lemon WJ, Palatini JJT, Krahe R, Wright FW: **Theoretical and experimental comparisons of gene expression indices for oligonucleotide arrays.** *Bioinformatics* 2002, **18:**1470-1476.
17. Liu W, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho M, Baid J, Smeekens SP: **Analysis of high-density expression microarrays with signed-rank call algorithms.** *Bioinformatics* 2002, **18:**1593-1599.
18. Hoffmann R, Seidl T, Dugas M: **Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray analysis.** *Genome Biol* 2002, **3:**research0033.1-0033.11.
19. Rice JA: **Mathematical Statistics and Data Analysis.** 2nd edition. Duxbury Press, Belmont, CA; 1995.
20. Rajagopalan D: **Comparison of statistical methods for oligonucleotide arrays.** *Bioinformatics* 2003, **19:**1469-1476.
21. Hollander M, Wolfe D: **Nonparametric Statistical Methods.** *Wiley, New York* 21999.
22. Sokal RF: **Biometry.** *Freeman, New York* 31995.
23. Jobson JD: **Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design.** *Springer-Verlag, New York* 1991.