Title

Identifying differentially expressed genes from microarray experiments via statistic synthesis

Authors

Yang, Yee Hwa
Xiao, Yuanyuan
Segal, Mark R

Publication Date

2004-07-02

Identifying differentially expressed genes from microarray experiments via statistic synthesis

Yee Hwa Yang[1†] Yuanyuan Xiao[2†] Mark R. Segal[3*]

July 2, 2004

Departments of [1]Medicine, [2]Biopharmaceutical Sciences, [3]Epidemiology and Biostatistics,
[1,3]Center for Bioinformatics and Molecular Biostatistics,
University of California, San Francisco, CA 94143, USA.

[†]These two authors contributed equally to this work.
[*]To whom correspondence should be addressed.

Abstract

Motivation: A common objective of microarray experiments is the detection of differential gene expression between samples obtained under different conditions. The task of identifying differentially expressed genes consists of two aspects: ranking and selection. Numerous statistics have been proposed to rank genes in order of evidence for differential expression. However, no one statistic is universally optimal and there is seldom any basis or guidance that can direct toward a particular statistic of choice.

Results: Our new approach, which addresses both ranking and selection of differentially expressed genes, integrates differing statistics via a distance synthesis scheme. Using a set of (Affymetrix) spike-in data sets, in which differentially expressed genes are known, we demonstrate that our method compares favorably with the best individual statistics, while achieving robustness properties lacked by the individual statistics. We further evaluate performance on one other microarray study.

Availability: The approach is implemented in an R package called `DEDS`, which is available for download from `http://www.biostat.ucsf.edu/jean/DEDS.htm`.

Contact: mark@biostat.ucsf.edu

1 Introduction

Microarrays have become increasingly common in biological and medical research. They enable the simultaneous study of thousands of genes and afford unprecedented ability to provide gene expression information on a whole genome level. There are several types of microarray technology including spotted arrays (DeRisi et al. (1996)) and high-density oligonucleotide chips (Lockhart et al. (1996)). The reader is referred to Schena (2000) and Bowtell and Sambrook (2003) for a more detailed introduction to the biology and technology underlying microarrays.

Microarray experiments generate large and complex multivariate data sets which present numerous data analytic challenges. These include, firstly, adjusting for the many sources of variability arising from probe and target preparation, array fabrication, and imaging; secondly, devising methods that are geared to the novel data structures – thousands of inter-related variables (genes), small sample sizes (arrays), and little or no replication; and lastly, where needed, accommodating multiple testing concerns. These complexities call for robust and integrated analysis tools to enhance reliability and

efficiency.

A very common data analytic goal is the identification of important genes amongst the many for which expression measures have been obtained. Importance typically corresponds to association with a response of interest. When we are contrasting expression between different groups or conditions (i.e. the response is polytomous), such important genes are said to be "differentially expressed" (DE). The task of identifying differentially expressed genes can be divided into two aspects: ranking and selection. Ranking requires specification of a statistic or measure which captures evidence for differential expression (DE) on a per gene basis. Selection requires specification of a procedure (e.g. stipulation of a critical-value) for arbitrating what constitutes "significant" DE. Ranking is fundamental and, arguably, is easier than selection. The primary importance of ranking arises from the fact that only a limited number of genes can be biologically validated from follow-up experiments, these being necessary to affirm DE in view of the present maturity of microarray measurements. The goals of this paper are to develop and illustrate a novel approach to ranking and selection that integrates differing measures of DE.

The paper is organized as follows. Section 2 presents a brief overview of some recently proposed methods for identifying DE genes in multiple-array experiments. Our proposed approach, *Differential Expression via Distance Synthesis* (DEDS), is presented in Section 3. We apply the method to two different data sets and present results in Section 4. Finally, Section 5 discusses our findings, extensions and open questions.

2 Existing Methods for Detecting DE Genes

The importance of developing new data analytic techniques to effectively identify differentially expressed genes is illustrated by the appreciable effort and literature dedicated to this area. We will overview several customized approaches to both aspects of DE detection with an emphasis on ranking.

2.1 Ranking Genes

For simplicity, and without loss of generality, we focus on a dichotomous response; i.e., deal with two-group comparisons. We designate the groups as treatment (T) and control (C). One may

consider the question of ranking genes in terms of DE in a discriminant analysis framework, i.e. consider DE genes as "features" or "variables" that best separate groups T and C (Ghosh (2003)). However, such approaches are focused on class prediction, and the rankings of genes so obtained are strongly influenced by between gene dependencies and feature selection strategies, so that individual gene DE is at best a by-product. For this reason, we limit our discussion to more direct approaches which assess differences between distinct groups on a single-gene basis, and we focus on those methods that we subsequently apply and describe the classes of statistics that they represent.

For two-channel competitive hybridization experiments, we assume that the comparisons of log-ratios are all indirect; that is we have n_T arrays in which samples from group T are hybridized against a reference sample R, giving n_T log-ratios $M_{T_i} = \log_2(T_i/R);\ i = 1, \ldots n_T$, and similarly we get n_C log-ratios $M_{C_j} = \log_2(C_j/R);\ j = 1, \ldots n_C$ from group C. For Affymetrix oligonucleotide array experiments, we have n_T chips with gene expression measures from group T and n_C chips with gene expression measures from group C.

Fold Changes and t-Statistics

The simplest gene ranking method is based on the fold change (FC) (i.e., ratio) in expression means between the two groups. Thus, for each gene, we compute the difference between (log) means $FC = \overline{M}_T - \overline{M}_C$, where $\overline{M}_T = 1/n_T \sum_{i=1}^{n_T} M_{T_i}$ and similarly for $\overline{M}_C$. While use of FC is conceptually appealing, ranking genes based on fold change alone implicitly assigns equal variance to every gene.

In contrast, the t-statistic defined as

$$t = \frac{\overline{M}_T - \overline{M}_C}{s_p \sqrt{1/n_T + 1/n_C}},$$

where s_p is the pooled standard deviation, takes into account differing gene-specific variation across arrays. Use of t-statistics is also widespread in assessing DE (Dudoit et al. (2002)). However, as indicated next, some care is needed in accommodating variation.

Penalized Statistics

The primary shortcoming of using t-statistics for ranking genes lies in the unstable variance estimates that arise when sample size is small. Such small sample sizes are common occurrences in micorarray experiments due to high costs and/or resource (RNA) limitations. Relatedly, with tens

of thousands of t-statistics (corresponding to tens of thousands of genes) there is frequently a number of large t-statistics driven by very small denominator standard deviations (s_p's), even though their numerators, measuring expression differences $\bar{M}_T - \bar{M}_C$, may also be small. To overcome these shortcomings a variety of approaches have been proposed to provide a more reliable variance estimate; they can be categorized into two groups. The first group consists of variance stabilizing functions, and the second group contains error fudge factors and Bayesian methods. The former seeks to decouple the mean-variance dependency by modeling the variance of the expression of a gene as a function of the mean expression of the gene (Rocke and Durbin (2001); Huber et al. (2002); Jain et al. (2003)).The latter regularizes the t-statistics by inflating their denominators:

$$t_a = \frac{\overline{M}_T - \overline{M}_C}{a + s_p\sqrt{1/n_T + 1/n_C}}.$$

There are differing ways of motivating such penalization and, accordingly, of estimating the penalty parameter a. What unifies these approaches is the (sometimes implicit) desire to utilize *between* gene information rather than relying solely on individual (within) gene information as afforded by t-statistics. What distinguishes them is the formalism invoked in applying information sharing. At one end are somewhat *ad hoc* methods that avoid modeling between-gene relationships such as SAM (Significance Analysis for Microarrays; Tusher et al. (2001)). At the other are Bayesian approaches (Newton et al. (2001)) with empirical Bayes methods along the way (B-statistics, Lönnstedt and Speed (2001)). Smyth (2004) extends and resets the hierarchical Bayesian model of Lönnstedt and Speed (2001) in the context of general linear models, and uses the moderated (penalized) t- (or F) statistics for inference about contrasts of biological interest.

Mixture Models

These Bayes and empirical Bayes approaches are arrived at via mixture models: it is postulated that we have a mixture of non-DE and DE genes with the latter group sometimes refined into up- and down-regulated components. Often mixing of these components is done at the level of one-dimensional, gene-specific summary statistics (Efron et al. (2001); Lee et al. (2000); Pan et al. (2002); Allison et al. (2002); Ghosh (2004)). Newton et al. (2004) argue that efficiency and sensitivity gains may be realized by effecting mixing on the gene expression (mean) values themselves, rather than on derived summaries. They develop so-called semi-parametric hierarchical mixture models (SHMMs) that have a number of features. Firstly, they are flexible (non-parametric) where

data are rich (lots of genes) and parametric where data paucity mandates assumption making (few replicates per gene). Secondly, the provision of per gene posterior probabilities of DE allows a straightforward approach to calibration in terms of false discovery rates (see below). Arguably, the main limitation surrounding the SHMM approach is the adequacy of the parametric model. Recognizing this, Newton et al. (2004) provide graphical diagnostics (stratified $Q-Q$ plots) for assessment of their chosen (gamma) model with their accompanying software.

Linear (Mixed) Models

In an effort to explicitly accommodate factors systematically impacting microarray gene expression values, a number of linear model/ANOVA based approaches have been advanced. Kerr et al. (2000) use fixed effect ANOVA models for logged intensities (single channel measurements) including terms for dye, array, treatment and gene main effects, as well as select interactions between these factors. By assuming a common variance across genes, a pooled (over genes) analysis is enabled, with large attendant increases in degrees of freedom for error variance estimation. However, there will likely be appreciable erosion of these degrees of freedom in order to accommodate interaction terms needed to "recover" the adjustments afforded by use of log ratios in two channel settings. To overcome the (strong) common variance assumption Jin et al. (2001) and Wolfinger et al. (2001) adopt mixed model ANOVA formulations, performing gene-by-gene analyses treating factors such as dye and array as random effects.

2.2 Ascribing Significance

Subsequent to gene ranking comes selection – the task of declaring which genes are significantly differentially expressed. Informal approaches include graphical examination of the ranking statistics via $Q-Q$ plots, whereas more formal approaches involve testing suitably constructed (joint) null hypotheses of equal expression (non-DE). Such joint formulations are mandated by the multiple testing concerns that follow from evaluating thousands of genes. Two main approaches to this problem have emerged. One seeks to control family-wise type I error rates (see Dudoit et al. (2002), Ge and Dudoit (2002)), using step-down Bonferroni correction (Westfall and Young (1993)). The other (Efron et al. (2000),Tusher et al. (2001), Storey (2002)) extends the false discovery rate (FDR) ideas of Benjamini and Hochberg (1995). Detailed discussion and comparisons of competing approaches to multiple testing correction are deferred to Dudoit et al. (2003) and Storey et al. (2002).

3 Differential Expression via Distance Synthesis

It is immediately apparent from Section 2.1 that there are numerous statistics available for ranking genes as a prelude to arbitrating DE. Furthermore, selecting the best statistic or ordering statistics in terms of merit is problematic due to the complexity of the variability structure present in all microarray data. Performance characteristics (e.g. efficiency, robustness) of two-sample statistics depend on data attributes such as underlying distribution(s) and nature and extent of outliers and/or contamination. To date, there has been no delineation of such attributes for microarray data, making pre-selection of the theoretically "best" statistic problematic. Likewise, no comparisons across a sufficiently wide range of benchmark data sets have been undertaken. In general, there has been a lack of microarray experiments specially designed for assessing DE gene identification. Such experiments often require careful inclusion of spike-in controls. In addition, there is a lack of a large scale independent validation using quantitative approaches, such as Northern Blotting or RT-PCR. Hence, investigators must make rather arbitrary choices when deciding which ranking statistic to use. It is in part to eliminate some of this arbitrariness but primarily to borrow strength across related measures that we propose synthesizing DE ranking schemes.

A somewhat related approach is that of Pareto Fronts (PF, Fleury et al. (2002)). Briefly, by regarding the set of measures as defining a multivariate point cloud (points corresponding to a gene's vector of measures) and employing a standard (coordinate-wise) partial order, a set of "non-dominated" (Pareto-optimal) genes is identified. In our context these could include genes ranked (for DE) very highly by one measure but very lowly by another.

We now sketch our approach to detecting differential expression via distance synthesis (DEDS). We also begin with the multivariate point cloud with a point corresponding to a gene's vector of DE measures. Rather than employing partial order, we exploit the fact that all measures are attempting to capture DE and synthesize (reduce to a scalar) as follows. Firstly, we define an "extreme point". Without loss of generality, assume that large values of all measures indicate DE. Then the extreme point is a vector of coordinates, each of which is the overall maximum of both observed and permuted values of that measure. Note that the extreme point may not correspond to any observed point. Next, the distance from all points to the extreme is computed. Intuitively, those points closest to the extreme are most likely to correspond to DE genes, but, we need to calibrate "close". This is done by generating null referent distributions analogously

to the methods Tibshirani et al. (2001) devised for calibrating gap statistics. Figure 1 provides a graphical illustration of the motivation behind DEDS. Specifically, panel (b) represents a common occurrence where the concordance between two DE measures (M1 and M2 in the plot) is relatively low (compared to panel (a)). We hope by measuring the distance of spots to the "extreme point" in the direction of DE and therefore taking both measures into consideration, the rankings of the genes that show discord between measures (blue spots in Figure 1(b)) will be lowered to a certain extent. The steps for calculating DEDS and selecting DE genes are detailed in Box 1.

*** Place Figure 1 about here ***

Box 1: DEDS – Permutation Based Algorithm

Calculating DEDS

1. Apply $j = 1, \ldots, J$ appropriate (DE measuring) statistics t_j to each of $i = 1, \ldots, N$ genes in the target data set yielding values t_{ij}. Without loss of generality, assume larger values indicate increased DE. Let the (observed) coordinate-wise extreme point be $E_0 = (\max_i(t_{i1}), \ldots, \max_i(t_{iJ}))$.

2. Locate the overall (observed, permutation) extreme point E:

   (a) Obtain $b = 1, \ldots, B$ permuted data sets by randomly assigning n_T arrays to class "T" and n_C arrays to class "C". For each permuted data set recalculate the J DE statistics for each of the N genes yielding t^b_{ij} and store the results (see below). Obtain the corresponding coordinate-wise maximum as above: $E_b = (\max_i(t^b_{i1}), \ldots, \max_i(t^b_{iJ}))$.

   (b) Obtain the coordinate-wise permutation extreme point E_p by maximizing over the B permutations: $E_p = (\max_b(E_{b1}), \ldots, \max_b(E_{bJ}))$.

   (c) Obtain E as the overall maximum: $E = \max(E_p, E_0)$.

3. Calculate a distance d from each gene to E. For example, one choice for a scaled distance is

$$d_i = \frac{(t_{i1} - E_1)^2}{MAD(t_1)^2} + \frac{(t_{i2} - E_2)^2}{MAD(t_2)^2} + \ldots + \frac{(t_{iJ} - E_J)^2}{MAD(t_J)^2},$$

   where MAD is the median absolute deviation from the median. Order the distances: $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(N)}$.

Assessing DE Significance

1. Using the stored statistics for each permuted data set b, analogously compute distances for each gene. Order the distances: $d^b_{(1)} \leq d^b_{(2)} \leq \ldots \leq d^b_{(N)}$.

2. For the i^{th} gene as ordered by the original d, define "falsely called genes" for each of the B sets of permutations as those genes, whose $d^b_{(j)}$ satistfy: $d^b_{(j)} \leq d_{(i)}, j = 1, \ldots, N$. Compute the median number of falsely called genes among the B sets of permutations.

3. The q value that controls FDR for the i^{th} ordered gene is computed as the median of the number of falsely called genes divided by the number of genes called significant (i).

4 Illustrative Examples

We evaluate the performance of our proposed method, DEDS, on two diverse data sets, each featuring a number of sub-studies. Both two-channel spotted arrays and one-channel Affymetrix arrays are represented, as are situations with minimal and considerable DE anticipated. Furthermore, the inclusion of a spike-in experiment permits assessment in the rare setting where a gold standard (a set of known DE genes) is available.

4.1 Study 1: Affymetrix Spike-in Experiment

4.1.1 Data Description

The spike-in experiment represents a portion of the data used by Affymetrix to develop their MAS 5.0 preprocessing algorithm. Here we utilize both MAS 5.0 and RMA (Irizarry et al. (2003)) probe level summaries in order to showcase a robustness property of DEDS. The data features 14 human genes spiked-in at a series of 14 known concentrations ($0, 2^{-2}, 2^{-1}, \ldots, 2^{10}$ pM) according to a Latin square design including 12612 null genes. Each "row" of the Latin square (given spike-in gene at a given concentration) was replicated (typically 3 times, two rows 12 times, 59 arrays in total). Further details are available at `http://www.affymetrix.com/analysis/download_center2.affx`. We analyze this data set as 91 ($= \binom{14}{2}$) pairs of two-sample comparisons corresponding to all pairwise comparisons of differing concentrations.

4.1.2 Results

The use of spike-ins allows computation of receiver operating characteristic (ROC) curves since we know which genes are DE and which are null. For each two-sample comparison, we compute five different DE measures: FC, t statistic, a moderated t statistic that coincides with the B statistic (Lönnstedt and Speed (2001)), SAM with the standard deviation penalty a taken as the median (over all genes) within gene standard deviation, and our synthesized measure (DEDS) based on the first four statistics.

Thus, we obtain 91 ROC curves for each of the 5 measures. ROC curves are created by plotting the true positive rates versus false positive rates. The summary ROC curves as presented in Figure 2 are just averages of the 91 individual curves. The two panels of Figure 2 correspond

to differing approaches for probe level summarization: panel (a) uses RMA while (b) uses MAS 5.0. Contrasting panels, the ensemble of ROC curves affirms the findings of the RMA developers (Irizarry et al. (2003)) that RMA summaries improve accuracy without sacrificing precision, relative to MAS 5.0. However, more important from the present standpoint, is that even though for each approach to probe level summarization there is one measure that performs relatively poorly (t statistics for RMA, FC for MAS 5.0), in both settings DEDS performed well. Thus, by utilizing several DE measures in the analysis of the spike-in experiment, DEDS is able to not only perform competitively with the best, but is also not adversely affected by the worst. And, of course, in practice we do not know a priori which measures will be suitable, as the spike-in data exemplifies.

*** Place Figure 2 about here ***

Furthermore, two array groups among the 14 array groups contains 12 replicates. This subset of data were used to evaluate the ability of DEDS in determining the correct cutoffs for declaration of differential expression. We apply the permutation procedure described in Box 1 to simulate the DEDS reference distribution and thereby estimate the number of differentially expressed genes. Controlling the FDR at 0.01, 16 genes are found to be differentially expressed, 11 out of which are among the 14 true DE genes. The top 20 DE genes have 13 of the 14 genes selected. The gene that is not detected is the most "difficult" one, since it is spiked in at the two lowest concentrations (0 pM, 0.25 pM). This represents an extremely low log-ratio / log-intensity for microarray based detection. No other DE statistic was able to detect this gene.

4.2 Study 2: Spt Splicing (SPT) Experiment

4.2.1 Data Description

The SPT experiment consists of 22 two-channel spotted oligonucleotide splicing arrays (Clark et al. (2002)). The primary goal of this experiment is to investigate the roles of eukaryotic chromatin elongation factors, Spt4-Spt5, in splicing (Hartzog et al. (1998)). We have analyzed a *spt4* null mutation *spt4*Δ, and three partial loss-of-function *spt5* mutations, *spt5-194*, *spt5-242* and *spt5-4*. In addition, we include analysis of *ceg1-250*, a temperature-sensitive mutation that causes rapid inactivation of the capping enzyme at non-permissive temperature (Fresco and Buratowski (1996)). Two independent mRNA samples are prepared for each mutant and a pair of dye-swap experiments

are performed for each mRNA sample. Figure 3 shows a graphical representation of the actual experiment which includes 4 hybridizations between each mutant (mut) and wild-type (wt) and an additional two self-self hybridizations of the wild-type.

To distinguish between spliced and unspliced transcripts for intron-containing yeast genes, oligonucleotide probes on these arrays were designed to detect splice junctions (SJ), introns (Int) and second exons (Ex) of spliced genes. After normalization, array measurements are summarized into two indices that capture splicing alterations (Clark et al. (2002)). The intron accumulation (IA) index assesses the change (relative to wt) of the intron probe signal in the mutants, normalized by the change in the corresponding exon. This is defined as

$$\text{IA index} = \log \frac{\text{mut}_{Int}/\text{wt}_{Int}}{\text{mut}_{Ex}/\text{wt}_{Ex}}.$$

Similarly, the splice junction (SJ) index, defined as

$$\text{SJ index} = \log \frac{\text{mut}_{SJ}/\text{wt}_{SJ}}{\text{mut}_{Ex}/\text{wt}_{Ex}},$$

measures a similarly normalized gain or loss of the splice junction probe signal in the mutants. A total of 254 SJ indices and 263 IA indices are formed. Details of this experiment can be found in Xiao et al. (2004).

*** Place Figure 3 about here ***

4.2.2 Models

To effectively analyze the splicing data, we apply five competing statistical models for evaluating DE and the synthesized measure DEDS. The nested experimental design of the splice mutant study illustrated by Figure 4 motivates the use of four different mixed ANOVA models described next. Furthermore, the limited amount of replication coupled with the small number of genes makes the semi-parametric mixture model (Newton et al. (2004)) appealing for conceptual reasons. Thus, we apply five individual statistics separately and identically to the SJ and IA indices and subsequently investigate their fusion via DEDS.

Distinguished by including wild-type self-hybridizations or not, DE assessment can be pursued by either *o*ne-sample (4 SJ or IA mutant indices per gene corresponding to the 4 replicate hybridiza-

tions) or *t*wo-sample (4 SJ or IA mutant indices *vs.* 2 SJ or IA wild-type indices) comparisons. A more detailed discussion of the difference between the direct (one-sample) and indirect (two-sample) comparisons can be found at Speed and Yang (2002). In addition to the above two approaches, we also consider another two approaches distinguished by allowing gene-specific variance heterogeneity or not. This latter case imposes the assumption that all genes exhibit a similar degree of variability and so can be jointly analyzed using a common estimate of error variance. This pooling dramatically increases error degrees of freedom (df). The former approach, on the other hand, does not impose the common variance assumption, allowing different variances for different genes. The resulting model is then fitted gene by gene. So, we have a 2×2 factorial of approaches, indicated by Models I - IV in Table 1).

*** Place Figure 4 about here ***

In addition to these four models, we also employ Newton et al. (2004) semi-parametric hierarchical mixture model (SHMM) (Model V in Table 1; see Section 2). Further details concerning the fitting of the five different models and the comparisons are provided in Xiao et al. (2004).

4.2.3 Results

Figure 5 displays a scatter plot matrix of $-\log_{10}(p)$, where p either corresponds to the Model I through IV unadjusted p-value for tests of DE or to the Model V posterior probability for non-DE for mutants *ceg1-250* (panel (a)) and *spt4*Δ (panel (b)). Note that by relating Model I through IV results to Model V results we may seemingly be perpetuating the "severe pedagogical problem of misinterpreting p-values as posterior probabilities" (Berger et al. (1997)). However, this is not the case. At no stage do we make probabilistic statements in terms of these quantities. Rather, they simply constitute a quantification of DE. The large number of DE genes anticipated for mutant *ceg1-250* is evident in Figure 5(a) (note the scales) and we observe high correlations between the five models. By contrast, minimal DE is anticipated for *spt4*Δ and results from the different models show this along with much lower agreement (panel (b)). The fact that Model V conforms more closely to the homoscedastic models (I and III) than to the heteroscedastic models (II and IV) is not surprising, since the SHMM utilizes information sharing between genes which is absent for the gene specific heteroscedastic models.

*** Place Figure 5 about here ***

Here is a situation where there is no clear advantages of one model over the others. Therefore, rather than trying to arbitrate between models and pick a single model on which to base DE rankings and declarations, or informally distilling sets of genes that are DE under two or more models, we employ DEDS as a robust means for synthesizing results and compare its performance with individual models.

As the sample sizes are too small (4 *vs* 2 for the two-sample statistics) to employ an effective permutation scheme, we elect to use p-values for the calculation of DEDS distances. The observed distances are then calibrated against expected values under the referent null distribution, which is simulated by drawing from marginal uniform $\mathcal{U}(0,1)$ variates with correlation structure conforming to the observed data; see Tibshirani et al. (2001). The algorithm is further motivated analogously to the mixture model approaches described in Section 2.1 but on the p-value scale, with non-DE corresponding to uniformity. Figure 6(a) and (c) show histograms of the p-values from Model I applied on *ceg1-250* SJ indices and p-values from the Affymetrix spike-in experiment (see Section 4.1). The dashed lines are the frequency we would expect when all genes were null. As can be seen, there are many DE genes in (a) but minimal DE genes in (c). Further details of the algorithm are provided in Supplementary Data A.

*** Place Figure 6 about here ***

Critical to ascribing DE is appropriate specification of cutoffs. Table 1 compares the numbers of genes achieving significance under FDRs of 0.01 and 0.05 for the five models and DEDS for mutant *ceg1-250*. Immediately striking are the differences in numbers of genes declared DE by the different models. While this can be attributed in part to differing operating characteristics, it serves to showcase how sensitive results are to statistic choice.

Here, evaluating differences in DE determinations by the different models is problematic since we have no gold standard and, unlike the spike-in study, we do not know which genes are truly DE. However, as DEDS synthesizes over individual statistics, we believe its rankings of genes to be more "robust" than single measures. To examine this, we have defined and compared the characteristics of the top DE genes for *ceg1-250* SJ indices by DEDS and the five individual DE models. As mentioned, *ceg1-250* is known to profoundly affect splicing and accordingly we treat the top 133 genes from each model as 'significantly" DE. While this number is somewhat arbitrary, the flavor

of the results below is not overly sensitive to this specification. We then classify genes into three major groups. Group I consists of DE genes by DEDS; group II contains non-DE genes by all six models. Group III is comprised of genes that are non-DE by DEDS but DE by one or more single measures and genes in this groups are further separated into different classes as illustrated in Figure 7(a). To aid comparisons between genes of different groups, we display three-dimensional scatter plots between different models. Plotted on all axes are $-log_{10}p$ of the corresponding models. Group I genes, represented by black spots, illustrate good concordance among DE models; whereas group III genes, represented as colored numbers, lie mainly off diagonal, indicating that such genes are ranked higher in one measure than the others. Thus, by ascribing high rankings to genes that exhibit agreement on DE among different measures and low rankings to genes that demonstrate discord among related measures, DEDS arguably provides a more robust gene ranking.

*** Place Figure 7 about here ***

5 Discussion

In this paper we have reviewed various statistical methods for the identification of differentially expressed genes in replicated microarray experiments. Additionally, we have advanced a novel method (DEDS) for this purpose. The DEDS algorithm synthesizes statistics or methods that estimate the same quantity of interest. The underlying principle behind DEDS is that genes that are highly ranked by different measures are more likely to be truly differentially expressed than genes that rank highly on a single measure.

Consider three widely used measures as an example: FC, t and SAM. The major limitations surrounding FC and t are the "equal denominator" and "small denominator" problems respectively. Concerns surrounding SAM include criteria for, and accuracy of, estimates for the penalty parameter a. Another problem is that with tens of thousands of statistics calculated, there is frequently a set of genes possessing large statistics for one measure only, these arising by chance and/or because of shortcomings associated with the measures. Such genes are likely to be "false positives". The advantage offered by DEDS over single measures is that, by combining over measures, such false positives are ranked lowly and become "true negatives". Therefore, the set of DE genes obtained via DEDS tends to be robust against limitations associated with individual measures.

The intuition behind DEDS simply draws on the concept of intersection, i.e. it attempts to select genes that are ranked highly on all measures. However, there is a clear distinction between DEDS and a simple intersection of results from individual measures, which treats the measures as independent. To further illustrate such differences, we use an additional data set of 6320 genes for which, due to the nature of the comparison groups, we anticipate substantial DE. The data derives from a study of cardiomyopathy in transgenic mice as influenced by overexpression of a G protein-coupled receptor, *Ro1*; details are provided in (Redfern et al. (2000)), while details on fitting of the various DE measures is available in Supplementary Data B. Figure 8 displays a volcano plot between two measures: FC and t statistics for the *Ro1* data. The categorization of genes here is similar to Figure 7. Black points are the top 905 genes selected by DEDS, and the gray ones are those failing to be among the extremes (top 905) of any measures. The horizontal and vertical reference lines are the cutoff values estimated using t statistic ($|t| \geq 2.98$) and FC ($|FC| \geq 0.66$) respectively. Classes 1 (red), 2 (cyan) and 3 (blue) are genes that are extreme in single measures, FC, t or SAM, only and their occurrences are possibly due to limitations associated with each measure. A simple intersection treats all measures as independent; thus, it will select only the genes from areas A and B. On the other hand, DEDS extends the intersection idea by considering all DE measures simultaneously, implicitly taking their correlation into account, thereby giving rise to a less stringent criteria and possibly a lower false negative rate.

*** Place Figure 8 about here ***

Natural questions that arise in using DEDS are choices of component DE measures and the distance metric. For the former, we have found that synthesizing t, FC and a penalized statistic, such as SAM, gives good performance for all data sets we have analyzed using DEDS. To investigate whether including highly correlated measures increases variation and so erode the efficiency of DEDS, we applied DEDS on the Affymetrix spike-in RMA data by synthesizing 2, 3 or all 4 measures from t, FC, SAM and moderated t. We have observed stable performance of DEDS as the obtained ROC curves for all combinations were largely overlapping. With regard to choice of distance metric, we recommend using a distance scaled according to variation of the component statistics so that DEDS results are not dominated by a single measure. In addition, even though we have demonstrated good performance of DEDS in the assessment of DE, it should be noted that sufficient replication is still the key in obtaining power and stability in analysis, and DEDS is not a panacea for poorly replicated experiments.

Finally, the DEDS method proposed in this paper is implemented in a `R` package (Ihaka and Gentleman (1996)) `DEDS` (Differential expressed via distance synthesis), which may be downloaded from `http://www.biostat.ucsf.edu/jean/DEDS.htm`.

Acknowledgements

We are grateful to Professor Grant Hartzog and Todd Burkin at the University of California, Santa Cruz, who have provided us the microarray data of the SPT experiment and have participated in many valuable discussions related to the analysis of the SPT data. We also thank the anonymous reviewers whose comments improved the quality of the paper. Y.X. wishes to acknowledge Professor C. Anthony Hunt at the University of California, San Francisco for funding support.

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, 39(1):1–20.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300.

Berger, J. O., Boukai, B., and Wang, Y. (1997). Unified frequentist and bayesian testing of a precise hypothesis. *Statistical Science*, 12(3):133–148.

Bowtell, D. and Sambrook, J., editors (2003). *DNA Microarrays: A Molecular Cloning Manual*. Cold Spring Harbor Press.

Clark, T. A., Sugnet, C. W., and Ares, M. (2002). Genomewide analysis of mrna processing in yeast using splicing-specific microarrays. *Science*, 296:907–910.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14:457–460.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139.

Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2000). Microarrays and their use in a comparative experiment. Technical report, Department of Statistics, Stanford University.

Efron, E., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160.

Fleury, G., Hero, A., Yoshida, S., Carter, T., Barlow, C., and Swaroop, A. (2002). Pareto analysis for gene filtering in microarray experiments. European Signal Processing Conference. `http://www.eecs.umich.edu/~hero/Preprints/eusipcogene.pdf`.

Fresco, L. D. and Buratowski, S. (1996). Conditional mutants of the yeast mrna capping enzyme show that the cap enhances, but is not required for, mrna splicing. *RNA*, 2:584–596.

Ge, Y. and Dudoit, S. (2002). Multiple testing procedures. `R` package, `http://lib.stat.cmu.edu/R/CRAN/`.

Ghosh, D. (2003). Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, 59(4):992–1000.

Ghosh, D. (2004). Mixture models for assessing differential expression in complex tissues using microarray data. *Bioinformatics*. Epub ahead of print, `http://bioinformatics.oupjournals.org/cgi/reprint/bth139v1`.

Hartzog, G. A., Wada, T., Handa, H., and Winston, F. (1998). Evidence that spt4, spt5, and spt6 control transcription elongation by rna polymerase ii in *Saccharomyces cerevisiae*. *Genes & Development*, 12:357–369.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differetial expression. *Bioinformatics*, 1(1):1–9.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31:e15.

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M., and Lee, J. K. (2003). Local-pooled-error test for indentifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, 19(15):1945–1951.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G., and Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. *Nature Genetics*, 29:389–395.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837.

Lee, M.-L. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci.*, 97:9834–9839.

Lockhart, D. J., Dong, H. L., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., and Horton, H. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680.

Lönnstedt, I. and Speed, T. P. (2001). Replicated microarray data. *Statistica Sinica*, 12(1):31–46.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8:37–52.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method gene expression profiles. *Biostatistics*, 5:155–176.

Pan, W., Lin, J., and Le, C. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? a mixture model approach. *Genome Biology*, 3:research0022.1–0022.10.

Redfern, C. H., Degtyarev, M. Y., Kwa, A. T., Salomonis, N., Cotte, N., Nanevicz, T., Fidelman, N., Desai, K., Vranizan, K., Lee, E. K., Coward, P., Shah, N., Warrington, J. A., Fishman, G. I., Bernstein, D., Baker, A. J., and Conklin, B. R. (2000). Conditional expression of a g_i-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy. *Proc. Natl. Acad. Sci.*, 97:4826–4831.

Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8(6):557–570.

Schena, M., editor (2000). *Microarray Biochip Technology*. Eaton.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 3.

Speed, T. P. and Yang, Y. H. (2002). Direct and indirect hybridizations for cDNA microarray experiments. *Sankhya: The Indian Journal of Statistics, Series A*, 64:706–720.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64:479–498.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2002). A unified estimation approach to false discovery rates. Technical Report 623, Department of Statistics, University of California, Berkeley.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 63:411–423.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to teh ionizing radiation response. *Proc. Natl. Acad. Sci.*, 98:5116.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625–638.

Xiao, Y., Yang, Y. H., Burckin, T., Shiue, L., Hartzog, G. A., and Segal, M. R. (2004). Analysis of a splice array experiment. In prepartion.

Table 1: A summary of the five competing DE models and the estimated number of DE genes from each model.

Model No	Model Description	FDR 0.01 *ceg1-250*	FDR 0.05 *ceg1-250*	FDR 0.05 *spt5-194*
I	Mixed ANOVA: one-sample/homoscedastic errors	114	149	6
II	Mixed ANOVA: one-sample/heteroscedastic errors	2	20	1
III	Mixed ANOVA: two-sample/homoscedastic errors	20	46	3
IV	Mixed ANOVA: two-sample/heteroscedastic errors	0	2	0
V	Semi-parametric hierarchical mixture model	89	142	33
DEDS Synthesis		**133**	**159**	**12**

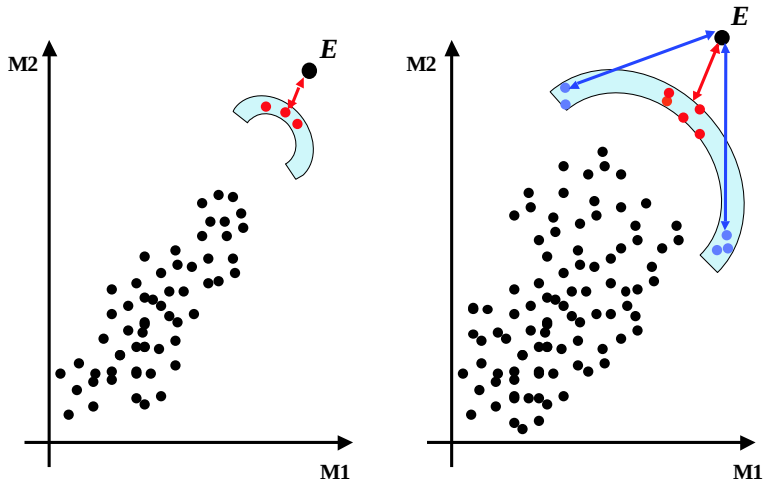
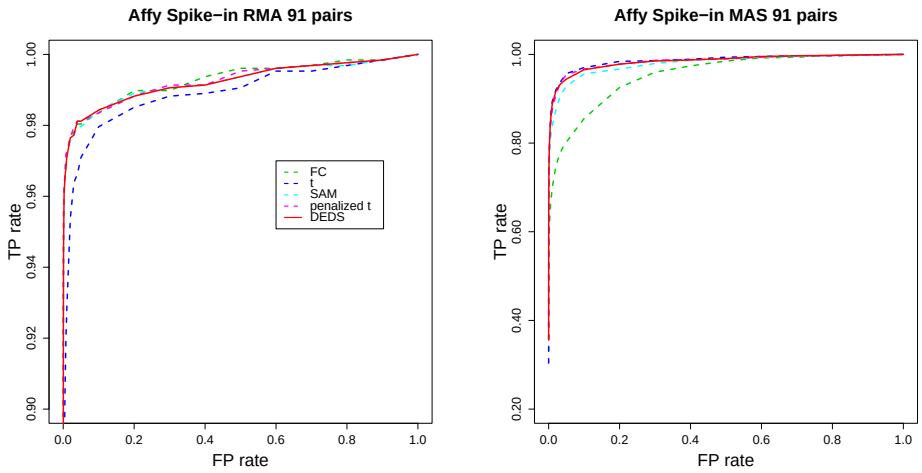


Figure 1: A graphical representation of the motivation behind DEDS. Plotted are the multivariate point cloud with a point corresponding to a gene's vector of two related DE measures, M1 and M2. E is the "extreme origin" in the direction of DE. Two scenarios are presented: when M1 and M2 show strong (panel (a)) or weak (panel (b)) concordance. Red spots are genes that are ranked highly by both measures and blue spots are genes that are ranked highly by one measure but lowly but the other.



(a) (b)

Figure 2: Comparisons of different DE measures using ROC curve for Affymetrix spike-in experiment. (a) Use RMA probe summary. (b) Use MAS 5.0 probe summary.

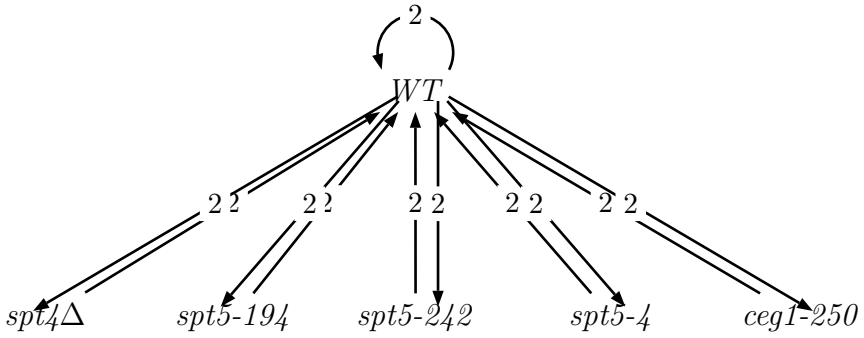


Figure 3: Study design of the SPT experiment. In this representation, *vertices* correspond to target mRNA samples and *edges* to hybridizations between two samples. By convention, we place the green-labeled sample at the tail and the red-labeled sample at the head of the arrow.

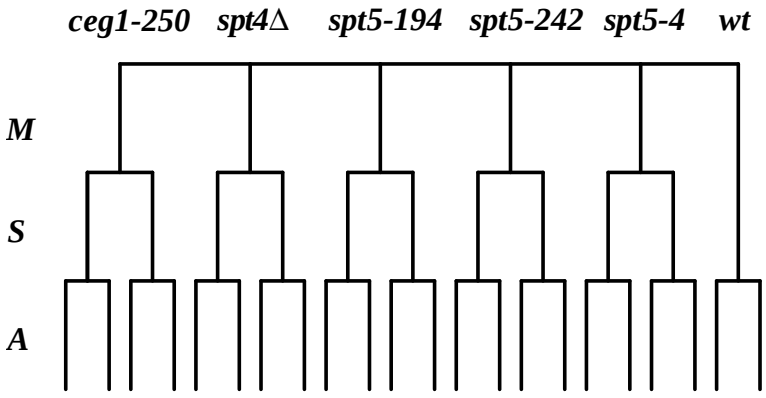
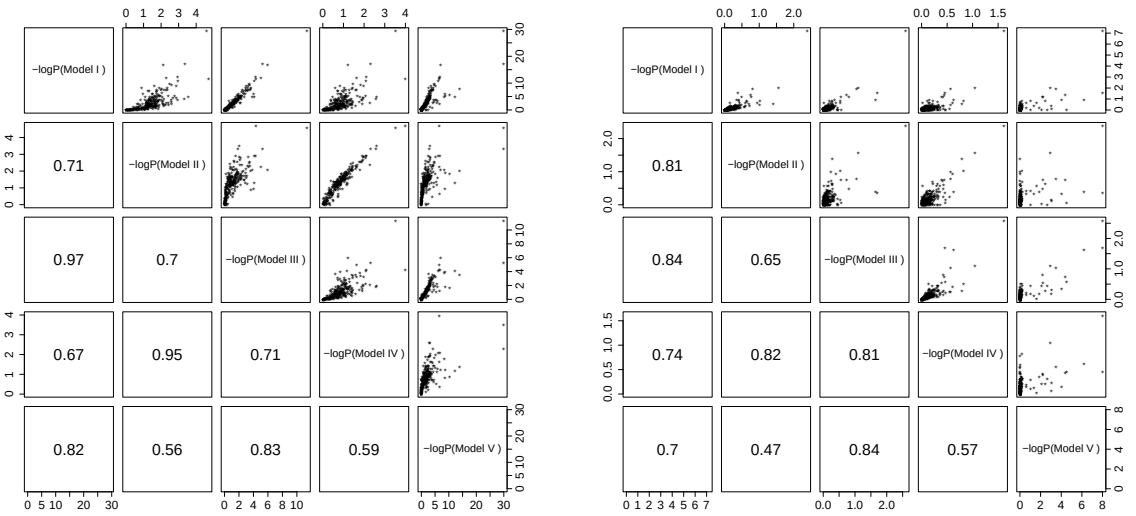


Figure 4: Nested study design of the SPT experiment. Each mutant (M) has two samples (S) and each sample is hybridized to two arrays (A). Therefore, A is nested in S and S is in turn nested in M.



(a) (b)

Figure 5: Pairwise scatter plots for the five models from a subset of the SPT experiment, (a) mutant *ceg1-250* and (b) mutant *spt4*Δ. The correlation coefficients of $-log_{10}p$ between corresponding models are shown in the lower-triangle grids.

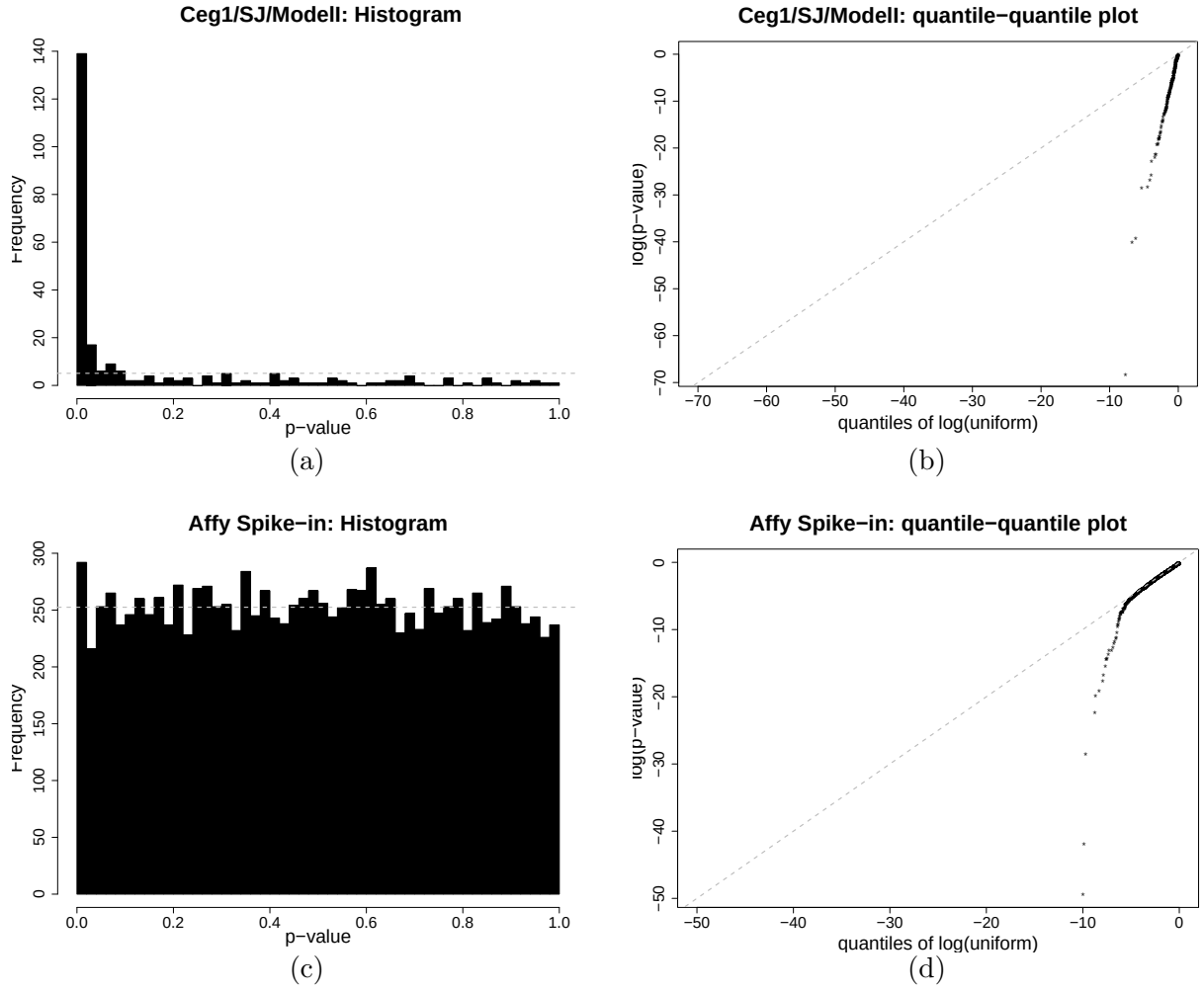
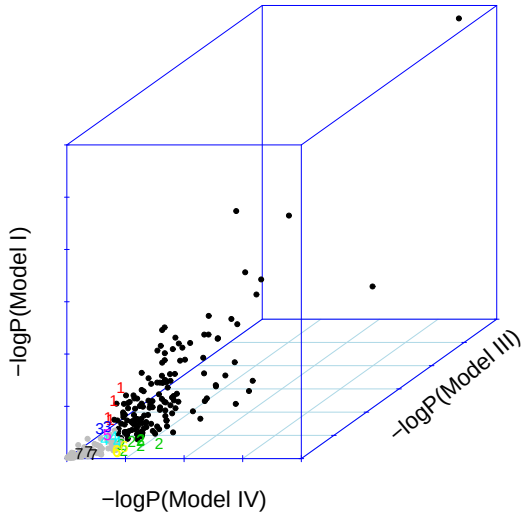


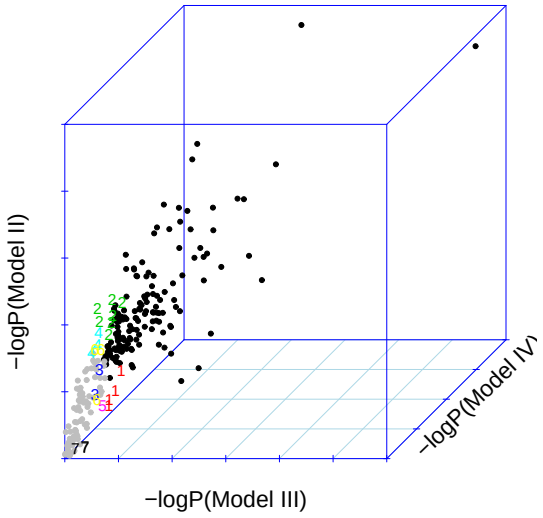
Figure 6: (a) Histograms and (b) $Q-Q$ plot of p-values from the mutant *ceg1-250* in the SPT experiment. (c) Histogram and (d)$Q-Q$ plot of p-values from the Affymetrix spike-in experiment where all but 14 genes remain unchanged. When no gene is expected to be differentially expression, the distribution of p-values is uniform, shown as dashed line in all panels.

Class	Model I	Model II	Model III	Model IV	Model V	No. genes
1	1	0	1	0	1	9
2	0	1	0	1	0	9
3	1	0	0	0	0	3
4	0	1	0	0	0	3
5	0	0	1	0	0	1
6	0	0	0	1	0	5
7	0	0	0	0	1	3

(a)



(b)



(c)

Figure 7: Relationship among differing DE models for the mutant *ceg1-250* (SJ indicies) in the SPT experiment. Black spots are DE genes by DEDS and gray spots are non-DE genes by all measures including DEDS. Colored numbers represent non-DE genes by DEDS but are DE genes by one or more of the five models. (a) Scatter plot of $-log_{10}p$ values from DE models I, III and IV; (b) Scatter plot of $-log_{10}p$ values from DE models II, III and IV; (c) Coding for the colored numbers in (a) and (b): 1 and 0 denote DE and non-DE respectively.

Class	FC	t	SAM	No. genes
1	1	0	0	145
2	0	1	0	258
3	0	0	1	9
4	1	1	0	0
5	1	0	1	1
6	0	1	1	34

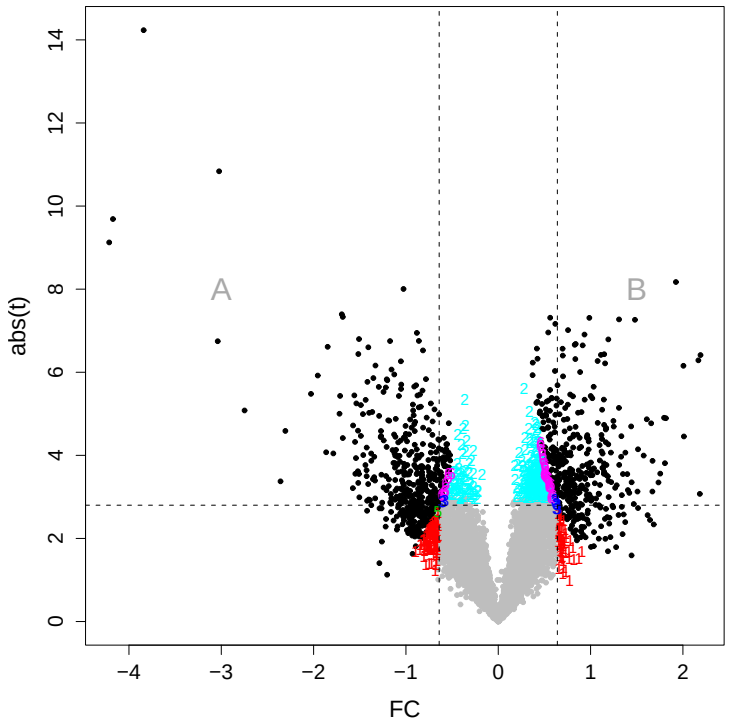


Figure 8: Volcano plot between FC and $|t|$ for the Ro1 experiment (C versus T). Black spots are DE genes by DEDS and gray spots are non-DE genes by all measures including DEDS. Colored numbers represent non-DE genes by DEDS but are DE genes by one or more of the three DE measures. Coding for the colored numbers is provided in the table; 1 and 0 denote DE and non-DE respectively.