

# Putative Null Distributions Corresponding to Tests of Differential Expression in the Golden Spike Dataset Are Intensity Dependent.

Daniel P. Gaile<sup>abc1</sup>, Jeffrey C Miecznikowski<sup>abc</sup>, Sung Eun Choe, Marc Halfon

---

<sup>a</sup>Department of Biostatistics, University at Buffalo, Buffalo, NY 14214-3000, USA

<sup>b</sup>Department of Biostatistics, Roswell Park Cancer Institute, New York 14263

<sup>c</sup>These authors contributed equally to this work.

Short title:

**Intensity Dependent Null Distributions**

---

<sup>1</sup>Corresponding Author. Department of Biostatistics, School of Public Health and Health Professions, 249 Farber Hall, University at Buffalo, 3435 Main Street, Buffalo NY 14214-3000, USA. Tel:+1(716) 829 2754. e-mail:dpgaile@buffalo.edu

## Abstract

We provide a re-analysis of the Golden Spike dataset (Choe et al., 2005), a first generation “spike-in” control microarray dataset. A recent publication (Dabney and Storey, 2006) reports that the p-values associated with tests of differential expression for null probesets (i.e., probesets designed to be fold change 1 across the two arms of the experiment) are not uniformly distributed. We provide evidence that the pre-processing algorithms considered in Choe et al. (2005); Dabney and Storey (2006) fail to provide expression values that are adequately centered or scaled. Furthermore, we demonstrate that the distributions of the p-values, test statistics, and probabilities associated with the relative locations and variabilities of the expression values vary with signal intensity. We provide diagnostic plots and a simple logistic regression based test statistic to detect these intensity related defects in the processed data. These diagnostics should prove useful in the analysis of the next generation of “spike-in” control experiments and should be modified for use in experimental settings where the invariant subset of genes is truly unknown.

## 1 Introduction

Normalization of microarray data is essential for removing systematic variation and biases that are present due to the nature of the assay. In experiments where the goal is to determine differential expression scientists have developed a variety of tests and algorithms to identify differentially expressed genes. One such experiment was the “Golden Spike” experiments by Choe et al. (2005). In the experiment six Affymetrix chips were divided into two groups: a control group (C) and a spike group (S). The S sample contains the same cRNAs as the C sample, except for ten selected groups of approximately 130 cRNAs per group that are present at a defined increased concentration compared to the C sample. This results in 3860 cRNAs, where 1309 cRNAs are spiked in with differing concentrations between the S and C samples. The rest (2551) are present at identical relative concentration between the two sets of microarrays. This type of experiment models the general paradigm of experiments meant to detect differential expression. Recently, however, the validity of inference based upon the Golden Spike experiment has been questioned (Dabney and Storey, 2006).

A key component to the Golden Spike dataset is knowledge of the null p-values for tests of differential expression, that is, information of the genes that are present in a 1:1 ratio on the S chips and the C chips provides knowledge of which tests for differential expression are truly null. Figure 1 provides a schematic of the two stage procedure used to obtain the sets of null p-values referenced in Choe et al. (2005) and Dabney and Storey (2006). The raw Golden Spike dataset consists of data generated by the scanning device used to measure the relative spot fluorescence values across each microarray chip. For oligonucleotide (Affymetrix) experiments such as the Golden Spike, the nature of the design demands heavy statistical intervention.

In microarray experiments, the end-stage analysis usually consists of simple two-sample test statistics such as the t-statistic or the Wilcoxon Rank Sum test statistic to test for differential expression. However, it is important to note that these statistics generally operate upon data matrices which have been subjected to potentially significant amounts of pre-processing. With this technology, there are several steps required in order to process the data in order to achieve a single value representing the intensity for a given probe. It is worthwhile to consider the Affymetrix data acquisition in two stages. A Stage I analysis includes image processing where each spot is deemed to consist of a collection of pixels. From the collection of pixels at a spot an overall signal value is determined by taking a summary measure (often a median) of the pixel set at each hybridization location on the chip. In the Affymetrix data design there are 15 probe pairs spotted for each gene or SNP. Each probe pair contains two 25-mer DNA oligonucleotide probes; the perfect match (PM) probe matches perfectly to the target RNA, and the mismatch (MM) probe which is identical to its PM partner probe except for a single homomeric mismatch at the central base-pair position. The MM probe serves to estimate the nonspecific signal. In this stage, the PM and MM signals are combined into one score representing the expression signal for a specific probe. The major software packages for Stage I analysis include Bioconductor’s “Affy” package, dChip and MAS 5.0 executables (Choe et al., 2005). Each software package varies in how the image processing is performed and how the PM and MM values are combined. After obtaining a signal for each probe, the next step in the Stage I analysis is to “normalize” the data accounting for between chip effects, spatial effects, intensity effects, a possible grid effect, and any nonlinear intensity/variation effects. Popular normalizing methods include lowess and loess smoothers to remove systematic sources of noise (Bolstad et al., 2003; Schadt et al., 2001).

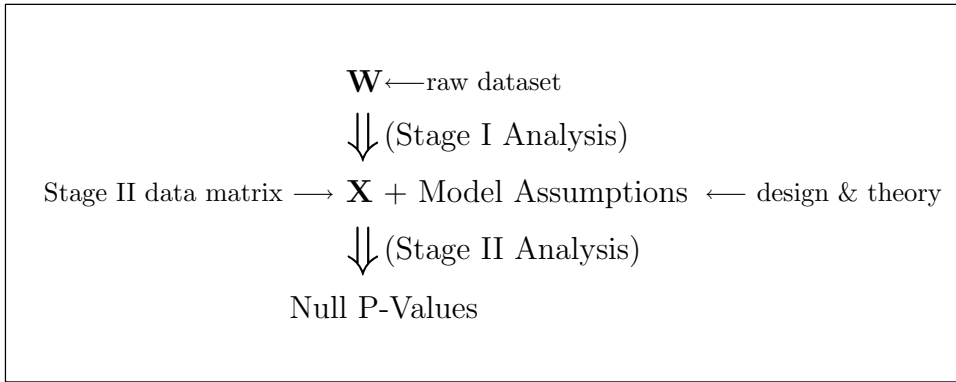


Figure 1: A two-stage procedure was used to obtain the sets of null p-values referenced in Choe et al. (2005) and Dabney and Storey (2006). The first stage of the procedure involves the application of algorithms designed to correct and normalize the raw data matrix,  $\mathbf{W}$ . The second stage of the procedure involves the evaluation of a test statistic for differential expression using information from the Stage II data matrix,  $\mathbf{X}$ . Choe et al. (2005) considered 150+ unique combinations of algorithms and input parameter values for the Stage I analysis and proposed a subset of the 10 best. Dabney and Storey (2006) determined that the distributions of the null p-values for the Choe et al. (2005) 10 best Stage I analyses were non-uniform for the most common choices for the Stage II analysis (e.g., t-test, permutation t-test, and Wilcoxon Rank Sum test). Dabney and Storey (2006) concluded that this non-uniformity implies that the technical replicates generated by the experiment do not constitute adequate approximations of biological replicates and hence, the Stage II model assumptions associated with these tests are not met. We provide evidence that the non-uniformity can be attributed, at least in part, to the failure of the Stage I analyses to correct for systematic biases in the raw data matrix,  $\mathbf{W}$ .

At this point in the Stage I analysis we have a matrix of dimension  $p$  by  $m$  where the  $p$  rows refer to the different probes, and the  $m$  columns refer to the different chips. The general procedure in normalizing this data is to use loess smoothers on the data set. One of the motivations for the Choe et al. (2005) experiment was to examine the numerous and varied normalization methods that currently exist for this data. Most of the normalization methods consider the data as a function of the matrix column. The goal of any of these normalization schemes is to reduce the systematic variation that exists in each chip. By considering each column of the data matrix as a separate chip, in each column we can scale and center the values, via loess smoothers so that each column has roughly the same “center” and “scale.” This general approach (as pointed out by Bolstad et al. (2003)) does not deal well with nonlinear relationships between arrays. Another method from Bolstad et al. (2003) is to transform the data via quantile regression so that the distribution of probe intensities is the roughly the same across arrays. At this stage, the normalization should result in a dataset where the systematic variation is reduced in order to get a clearer glimpse of the biological variation, or “interesting” variation that is present in these experiments.

Ultimately, the Stage I analysis results in an  $X$  matrix of dimension  $p$  by  $m$  for each experiment where the  $p$  rows correspond to each (smoothed) probe value and the  $m$  columns correspond to the sample. In the Golden Spike datasets, numerous options in the Stage I analysis were examined, resulting in 152 different  $X$  matrices with each matrix corresponding to a

different set of parameters in a Stage I analysis. From these 152 datasets, 10 “best” datasets were chosen that represented the best combination of processing in terms of detecting approximately 95 percent of true differentially expressed genes (DEGs) with changes greater than twofold, but less than 30 percent with changes below 1.7 fold before exceeding a 10 percent false-discovery rate. At this point, each data matrix  $X$  represents the input for Stage II analysis. The goal of Stage II analysis is to answer the researcher’s questions of the experiment. Usually in the microarray setting this consists of a ranked list of genes determined to be differentially expressed between two groups such as treatment versus control. The methods of Stage II generally take in to account facets of the experimental design and allow the user to control for things like the false discovery rate (FDR) within a given two sample test environment. Often the validity of Stage II analyses depends upon the assumption that the Stage I analysis has provided a Stage II data matrix such that the two sample test statistic null p-values are uniformly distributed.

Dabney and Storey (2006) provide a re-analysis of the Golden Spike dataset in which they consider the most common choices for the Stage II analysis (e.g., t-test, permutation t-test, and Wilcoxon Rank Sum test) and demonstrate that the null p-values for the Choe et al. (2005) 10 best datasets were non-uniform in all cases. Dabney and Storey (2006) note that statistical methods to control the FDR require the assumption that the true null p-values are uniformly distributed and hence the Golden Spike data can not be utilized to assess the performance of such methods. Furthermore, Dabney and Storey (2006) conclude that the non-uniform distributions of p-values are the direct consequence of an experimental design which requires that technical replicates adequately approximate biological replicates. The authors provide simulation results which demonstrate that technical replicates analyzed as biological replicates can provide non-uniform null p-value distributions but fail to provide any evidence that the parameter values that evoke this behavior are consistent with the set of conditions under which the Golden Spike experiment was conducted. Presumably the reader is left to infer that because the null p-values are non-uniform and because technical replicates analyzed as biological replicates can provide non-uniform null distributions, then the technical replicates generated by the Golden Spike experiment do not adequately approximate biological replicates.

We have replicated and extended the analyses of Dabney and Storey (2006) and we agree with the assessment that the null p-values are indeed non-uniform. We also agree with the conclusion that, given current pre-processing (i.e., Stage I) technologies, the Golden Spike datasets should not serve as reference datasets to evaluate FDR controlling methodologies. However, we disagree with the assessment that the non-uniform p-values are merely the byproduct of testing for differential expression under the assumption that chip data are approximate to biological replicates when, in fact, they are not. Whereas Dabney and Storey (2006) attribute the non-uniform p-values to violations of the Stage II model assumptions, we provide evidence that the non-uniformity can be attributed to the failure of the Stage I analyses to correct for systematic biases in the raw data matrix. Specifically, we demonstrate that the 10 best Stage I analyses considered in Choe et al. (2005) and Dabney and Storey (2006) provide Stage II data matrices in which the columns are neither adequately centered nor adequately scaled and that observed deviations in centering and scaling are intensity dependent.

Our re-analysis of the Golden Spike data provides evidence that even if the columns of the Stage II data matrix are adequately centered, that fundamental issues pertaining to relative scale can still invalidate the null distributions of commonly used two sample test statistics. We demonstrate the utility of plotting the observed test statistics and p-values as a function of signal intensity and we propose simple diagnostic plots to assess whether or not the relative center and scales of the underlying distributions for the control and spike-in expression values vary as a function of signal intensity. We advocate that the diagnostic plots presented in this

manuscript be applied to the next generation of spiked-in datasets and speculate that they may also be useful for detecting egregious intensity dependent effects in experiments where the null genes are unknown.

The remainder of the manuscript is organized as follows. In Section 2 we provide our analysis of the Golden Spike datasets. In Section 2.1 we present the model of Dabney and Storey (2006) and demonstrate that it fails to provide a reasonable approximation to the true state of nature. In Section 2.2 we demonstrate that the inclusion of empty probesets in the invariant set of “null genes” used in in the 10 best Stage I analyses of Choe et al. (2005) had a deleterious effect on the Stage II data matrices. In Section 2.3 we demonstrate that other common two-sample test statistics fail to provide null p-value distributions which are uniform. In Section 3 we provide a set of diagnostic plots and an approximate diagnostic test and apply them to the Golden Spike datasets.

## 2 Re-analysis of the Golden Spike Dataset

Our analysis of the Golden Spike Dataset reveals that the null two sample t-test p-value distributions are non-uniform across the 152 combinations of Stage I analyses. The fact that all distributions were non-uniform implies that this problem can not be attributed to the procedure utilized to identify the ten best datasets. The two sample test was conducted using the equal variance t-test so that the analysis would be consistent with the one presented in Dabney and Storey (2006). Other test statistics (i.e., Wilcoxon Rank Sum, permutation t-test, and Welch’s t-test) were considered and yielded similar results, a finding which is also consistent with those reported in Dabney and Storey (2006). Figure 2 contains sample quantile plots for the 152 sets of null p-values corresponding to the 152 datasets described in Choe et al. (2005). The black curves in Figure 2(a) correspond to the ten best datasets (i.e., datasets labeled 9a-e and 10a-e). The grey curves in Figure 2(a) correspond to the remaining 142 datasets and demonstrate that non-uniform null p-values were observed in datasets other than the ten best.

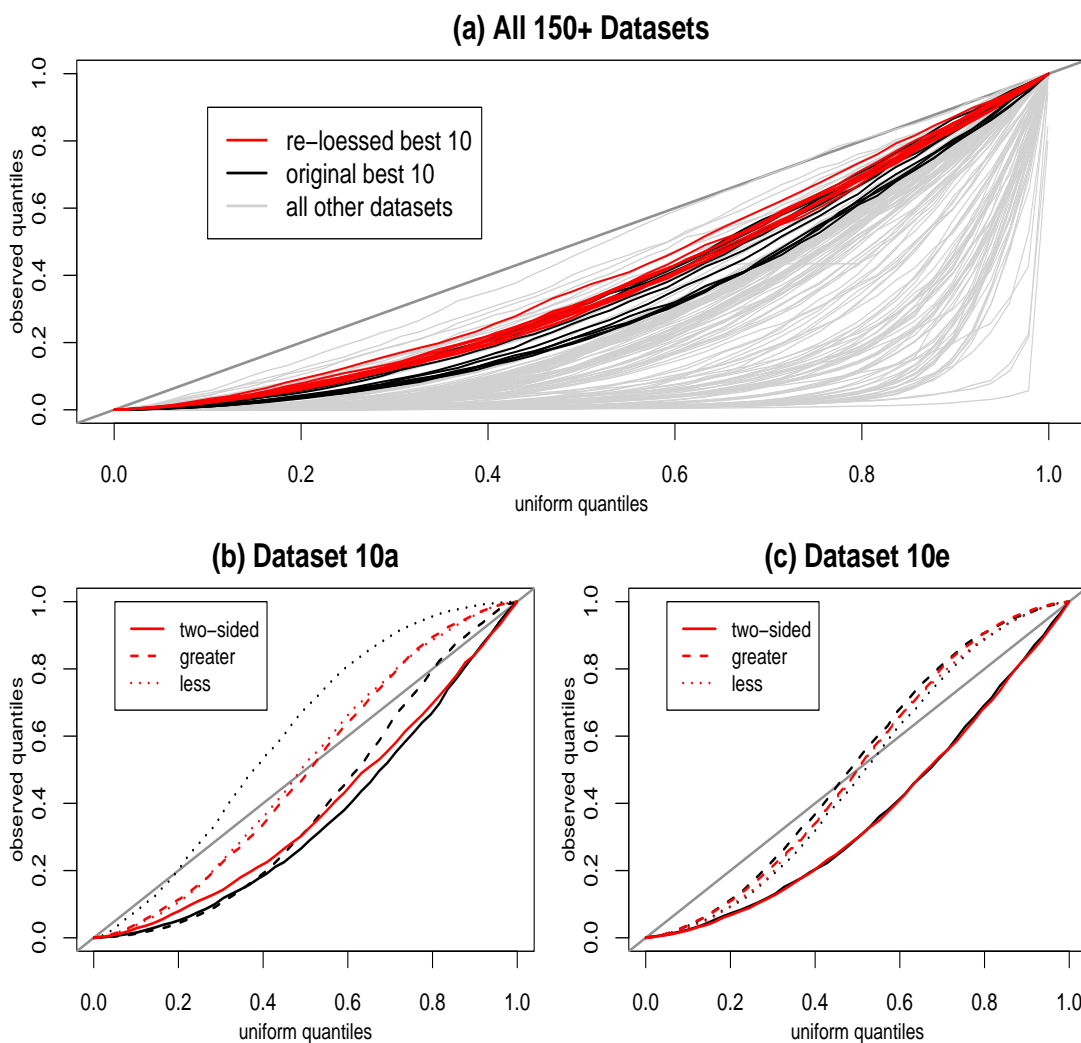


Figure 2: Sample quantile plots for various sets of null p-values. The x-axes correspond to the expected quantiles for a uniform distribution and the y-axes correspond to the observed (sample) quantiles. (a) Sample quantile plots for the t-test p-values associated with the 152 preprocessing combinations described by Choe et al. (2005). Black lines correspond to the 10 best datasets and are consistent with the curves presented in Figure 1 of Dabney and Storey (2006). The red lines correspond to re-loessed datasets that were obtained using the same combinations of preprocessing steps as the original 10 sets with the exception that the invariant subsets consisted only of the “present null” (present with fold change=1) probesets (versus both the “present null” and “empty null” probesets used in Choe et al. (2005)). The distribution of the p-values thus depends upon the choice of the invariant subset. (b) Sample quantile curves for dataset 10a. Solid lines correspond to the two-sided p-values and the dashed and dotted lines correspond to the p-values associated with the one sided tests. The model presented in Dabney and Storey (2006) does not account for the discrepancy in the one-sided p-values observed for this dataset, which is not manifest in the re-loessed data (red lines). Similar results are seen with datasets 10b, c, d and 9a, b, c, d. (c) As in (b) but showing sample quantile curves for dataset 10e; dataset 9e is similar. The p-value discrepancies are much less pronounced for these two datasets. This figure appears with permission in the response to Dabney and Storey (2006).

## 2.1 Observed p-value Distributions Inconsistent with Model of Dabney and Storey (2006)

Dabney and Storey (2006) attributed the non uniform distribution of p-values to the fact that the Golden Spike experimental design requires technical replicates to masquerade as biological replicates. In their response to Dabney and Storey (2006), the authors of Choe et al. (2005) acknowledged that the three spike-in and three control chips were technical replicates but they argued that the differences in the relative concentrations of the fold change one genes within the master spike-in sample (i.e., prior to splitting into three samples) compared to those in the master control sample should have had a negligible impact on the observed expression values.

Dabney and Storey (2006) proposed the the following model for  $i$  genes,  $i = 1, 2, \dots, m$ ,  $j$  treatments,  $j = C, S$  and  $k$  technical replicates,  $k = 1, 2, 3$  and the Stage II expression data matrix  $X$ :

$$X_{ijk} = \mu_{ij} + \epsilon_{ij} + \phi_{ijk} \quad (1)$$

where

$$\begin{bmatrix} \mu_{iC} \\ \mu_{iS} \end{bmatrix} = \begin{bmatrix} \text{mean of gene } i \text{ for the control set} \\ \text{mean of gene } i \text{ for the spike-in set} \end{bmatrix}$$

$$\begin{bmatrix} \epsilon_{iC} \\ \epsilon_{iS} \end{bmatrix} \sim N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_i^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

and  $\phi_{ijk} \sim N(0, \tau_i^2)$ . In the model stated in Equation (1), straightforward calculations show that for gene  $i$  we have the following distribution:

$$\begin{bmatrix} x_{iC1} \\ x_{iC2} \\ x_{iC3} \\ x_{iS1} \\ x_{iS2} \\ x_{iS3} \end{bmatrix} \sim N_6 \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left[ \begin{array}{ccc|ccc} \sigma_i^2 + \tau_i^2 & \sigma_i^2 & \sigma_i^2 & \sigma_i^2 \rho & \sigma_i^2 \rho & \sigma_i^2 \rho \\ \sigma_i^2 & \sigma_i^2 + \tau_i^2 & \sigma_i^2 & \sigma_i^2 \rho & \sigma_i^2 \rho & \sigma_i^2 \rho \\ \sigma_i^2 & \sigma_i^2 & \sigma_i^2 + \tau_i^2 & \sigma_i^2 \rho & \sigma_i^2 \rho & \sigma_i^2 \rho \\ \hline \sigma_i^2 \rho & \sigma_i^2 \rho & \sigma_i^2 \rho & \sigma_i^2 + \tau_i^2 & \sigma_i^2 & \sigma_i^2 \\ \sigma_i^2 \rho & \sigma_i^2 \rho & \sigma_i^2 \rho & \sigma_i^2 & \sigma_i^2 + \tau_i^2 & \sigma_i^2 \\ \sigma_i^2 \rho & \sigma_i^2 \rho & \sigma_i^2 \rho & \sigma_i^2 & \sigma_i^2 & \sigma_i^2 + \tau_i^2 \end{array} \right] \right)$$

If we consider the linear combination:

$$W_i = \bar{X}_{iS} - \bar{X}_{iC} \quad (2)$$

where  $\bar{X}_{iS}$ , and  $\bar{X}_{iC}$  represent the sample mean for probe  $i$  under condition  $S$  and  $C$ , respectively, then it follows that

$$W_i \sim N(\mu_S - \mu_C, 2(1 - \rho)\sigma_i^2 + \frac{2}{3}\tau_i^2). \quad (3)$$

The standard two-sample t-statistic is given by

$$T = \frac{\bar{X}_{iS} - \bar{X}_{iC}}{\sqrt{s_S^2/3 + s_C^2/3}}. \quad (4)$$

where  $s_S$  and  $s_C$  represent the sample standard deviation of probe  $i$  under condition  $S$  and  $C$  respectively. It follows from (3) that a t-test statistic calculated with respect to random variables governed by model (1) constitutes an observation from a distribution which is heavier in the tails than a  $t_4$  distribution provided  $2(1 - \rho)\sigma_i^2 > 0$ . This follows from the fact that square

of the denominator of the test statistic is an unbiased estimator of  $\frac{2}{3}\tau_i^2$  and hence, a negatively biased estimator of the variance of  $W_i$ . Hence, evaluating t-test statistics such as (4) against a  $t_4$  distribution will provide p-values which are negatively biased. This is the crux of the Dabney and Storey (2006) critique of the Golden Spike experimental design.

Unfortunately the experimental design does not provide enough data to fit model (1) and directly estimate the relative magnitudes of  $\sigma_i^2$  and  $\tau_i^2$ . However, it is still possible to determine that model (1) does not adequately explain all aspects of the observed p-value distributions for all Stage II datasets. There is an underlying symmetry to this putative model mis-specification because if the t-test statistic underestimates the actual variance, then the distributions of the one-sided p-values should be parsimonious with the distribution of the two-sided p-values. In actuality, the one-sided p-values for eight of the ten best datasets proved to be inconsistent with the two sided p-values. Figure 2 (b) contains the sample quantile curves for dataset 10a where solid lines correspond to the two-sided p-values and the dashed and dotted lines correspond to the p-values associated with the one sided tests. The distributions of the one-sided p-values are not in agreement. Surprisingly, the set of p-values associated with the “less than” alternative appearing to contain a disproportionate number of large p-values and an insufficient number of small p-values. Datasets 9a-d and 10b-d provided results similar to those observed for dataset 10a. Figure 2 (c) contains the sample quantile curves for dataset 10e in which the two sets of one-sided p-values appear to share the same underlying distribution.

Most importantly, the model (1) does not adequately explain the most intriguing aspect of the observed p-value distributions for all Stage II datasets; that the distributions are not invariant with respect to the overall signal intensity. Figures 3(a) and (c) contain curves which estimate the underlying population quartiles for the p-value distributions as a function of signal intensity for datasets 10a and 10e. The observed p-values were modeled as a function of a 4<sup>th</sup> order polynomial for rank-it intensity,  $\frac{\text{rank of value}}{\text{total \# of values} + 1}$ . The curves were fit using quantile regression (Koenker, 2006; Koenker and D’Orey, 1987) where the black lines correspond to the fits for  $\tau = 0.5$  (solid) and  $\tau = 0.25, 0.75$  (dashed). Solid and dashed grey lines indicate the theoretical medians and quartiles, respectively. Inspection of Figure 3(a) reveals that the p-values for dataset 10a are negatively biased across all intensities but that the magnitude of the bias is intensity variant. Inspection of Figure 3(b) reveals that the p-values for dataset 10e are also negatively biased across all intensities but that the magnitude of the bias does not vary with intensity to the extent which was observed for dataset 10a.

Figures 4(a) and (c) contain curves which estimate the underlying population quartiles for the t-test distributions as a function of signal intensity for datasets 10a and 10e. As in the previous figure, the observed t-tests were modeled as a function of a 4<sup>th</sup> order polynomial for rank-it intensity. The curves were fit using the quantile regression and are coded as in Figures 3(a) and (c). For dataset 10a, the test statistics corresponding to the null genes with overall signal intensities falling below the median all appear to be positively biased and exhibit a greater degree of variation than is compatible with the null  $t_4$  distribution. This observation is consistent with the previous observation that the p-values associated with the “less than” alternative contain an excessive number of large p-values. For dataset 10a, the test statistics corresponding the null genes with overall signal intensities falling in the lower 15-20% appear to be negatively biased while test statistics corresponding the null genes with overall signal intensities falling in the upper 15% appear to be positively biased. The results in Figures 2(c) and 4(c) suggest that the effect that these biases have upon the relative distributions of the one-sided p-values appears to wash out across all signal intensities even though the distributions are different for genes with overall signal intensities at the extremes.

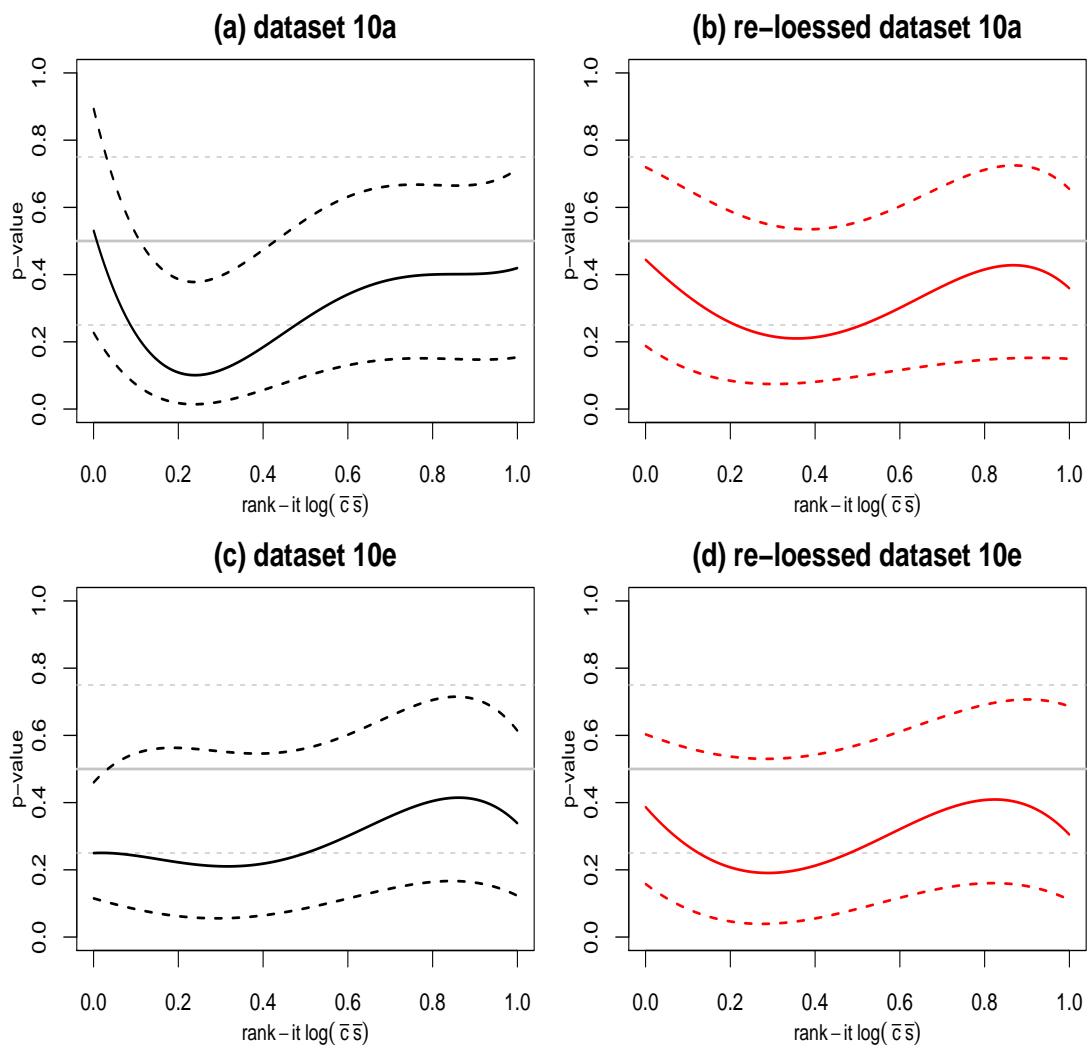


Figure 3: Estimates of the null p-value quartiles vary as a function of signal intensity for datasets 10a (a, b) and 10e (c, d); although less so for the re-loessed data. The x-axes correspond to the rank-it (i.e.,  $\frac{\text{rank of value}}{\text{total \# of values} + 1}$ ) of the log of the product of the expression means. The y-axes correspond to the observed two-sided p-values. Solid and dashed grey lines indicate the theoretical medians and quartiles, respectively. The null p-values were modeled as a function of a 4<sup>th</sup> order polynomial for rank-it intensity. Black and red lines correspond to the quantile regression fits for  $\tau = 0.5$  (solid) and  $\tau = 0.25, 0.75$  (dashed). Portions of this figure appear with permission in the response to Dabney and Storey (2006).

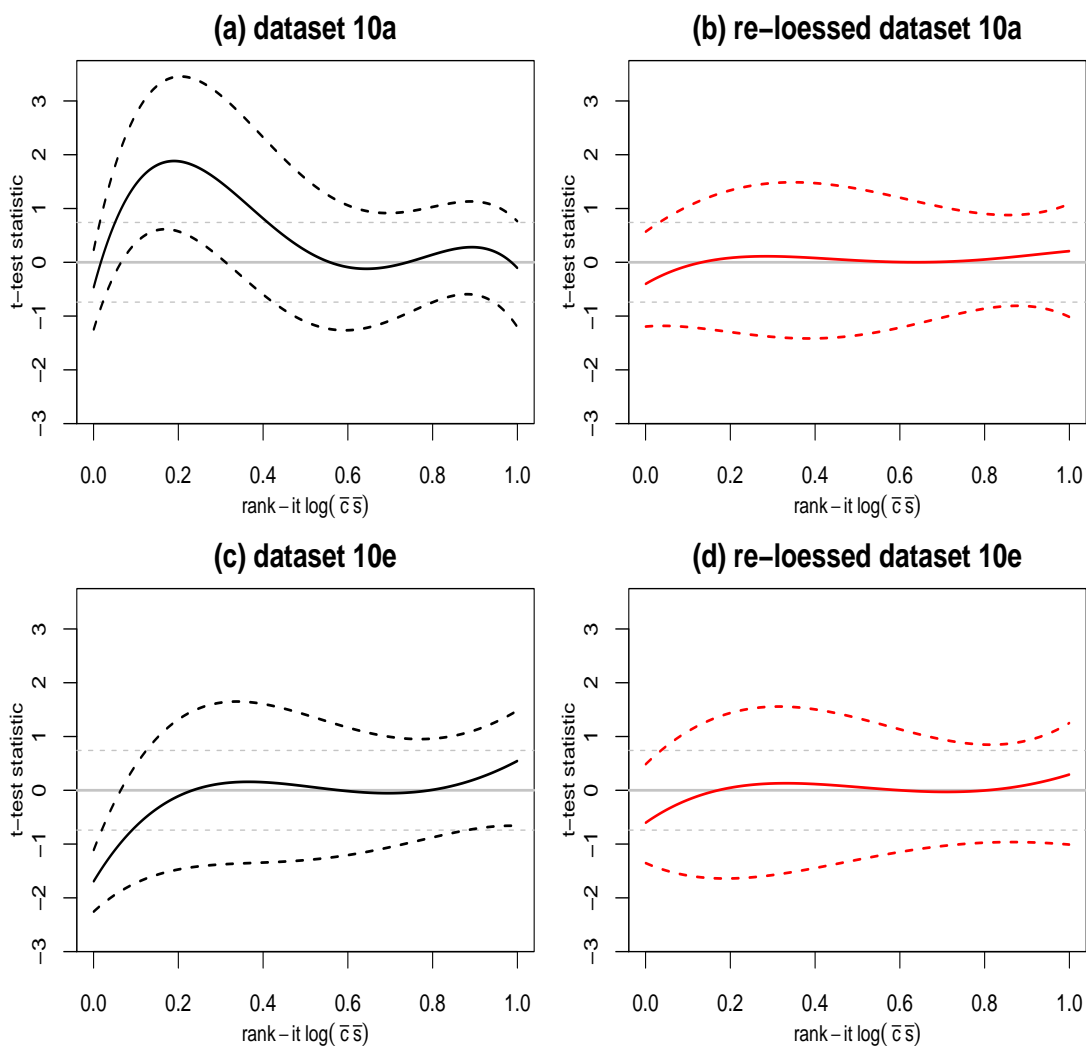


Figure 4: Estimates of the null t-test statistic quartiles vary as a function of signal intensity for datasets 10a (a, b) and 10e (c, d); although less so for the re-loessed data. The x-axes correspond to the rank-it (i.e.,  $\frac{\text{rank of value}}{\text{total \# of values} + 1}$ ) of the log of the product of the expression means. The y-axes correspond to the observed two-sided t-test statistics. Solid and dashed grey lines indicate the theoretical medians and quartiles, respectively. The null t-test statistics were modeled as a function of a 4<sup>th</sup> order polynomial for rank-it intensity. Black and red lines correspond to the quantile regression fits for  $\tau = 0.5$  (solid) and  $\tau = 0.25, 0.75$  (dashed). The overwhelming positive deviation of the null distribution in (a) is consistent with the discrepancy between the one-sided p-values observed in Figure 2(b). Portions of this figure appear with permission in the response to Dabney and Storey (2006).

## 2.2 Re-Loessing Golden Spike Dataset Improves Null Distributions

Each of the ten best Choe et al. (2005) Stage II data matrices were obtained using Stage I steps that included correcting the observed intensity with a loess curve that was fit to values from an invariant set of genes. This invariant set included present null (i.e., present with a putative fold change of one) as well as empty null (i.e., not present in either sample) probesets. The inclusion of the empty null probesets appears to have had a deleterious effect on the distributions of the null p-values. We have calculated a new set of ten best datasets in which the invariant set contains only the present null probesets. These calculations were performed at our request by the authors of Choe et al. (2005) using analysis scripts which were identical to those used for the original analyses except for passages of the code relating to the identification of the invariant set. The red curves in Figures 2(a)-(c) correspond to the sample quantile curves for the re-loessed datasets. The “re-loessed” datasets are still significantly non-uniform, although noticeably less so than the original datasets. Inclusion of the empty nulls in the original invariant sets appears to have contributed to the observed biases in the underlying t-distributions as inspection of Figures 4(b) and (d) indicates that this bias appears to be mitigated in the re-loessed datasets.

## 2.3 Other Common Two Sample Tests Failed to Provide Uniform Null Distributions

Although removal of the empty nulls from the invariant set provides data that is better centered than the original ten best, the results depicted in Figures 3(b) and (d) indicate that the p-value distributions are still non-uniform and intensity dependent. We re-analyzed the re-loessed ten best datasets using three other common two sample test procedures and found that none were robust to the problems which remain in the underlying Stage II data matrices. Figure 5 contains the results of this analysis for re-loessed dataset 10a.

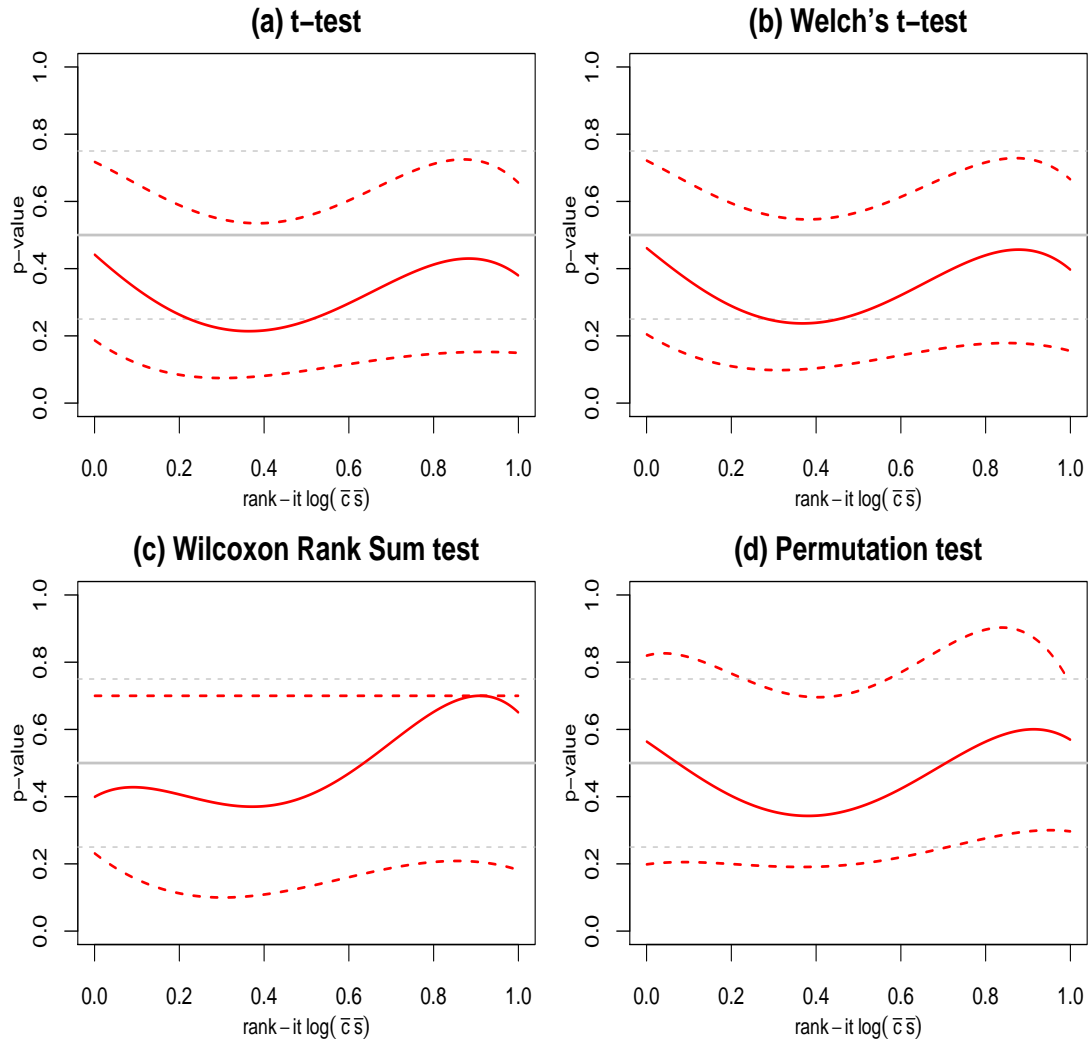


Figure 5: Estimates of the null p-value quartiles vary as a function of signal intensity for four common two sample test procedures applied to the re-loessed dataset 10a. The x-axes correspond to the rank-it (i.e.,  $\frac{\text{rank of value}}{\text{total \# of values} + 1}$ ) of the log of the product of the expression means. The y-axes correspond to the observed two-sided p-values. Solid and dashed grey lines indicate the theoretical medians and quartiles, respectively. The null p-values were modeled as a function of a  $4^{\text{th}}$  order polynomial for rank-it intensity. Red lines correspond to the quantile regression fits for  $\tau = 0.5$  (solid) and  $\tau = 0.25, 0.75$  (dashed). (a) P-values for the two sample t-test conducted under the assumption of equal variances. (b) P-values for the two sample t-test conducted under the assumption of unequal variances (a.k.a. Welch's test). Relaxing the assumption of equal variances provides for only a slight improvement in the distribution of the null p-values (c) P-values for the Wilcoxon Rank Sum test. (d) P-values for the permutation test (Hothorn, 2001; Hothorn and Hornik, 2006). Permutation based approaches are not robust to the systematic errors manifest in the Stage II data matrix.

### 3 Distribution Free Diagnostic Plots

A distribution free analysis of the ten best datasets (original and re-loessed) reveals that removal of the empty nulls from the invariant set provides for Stage II datasets which are adequately centered but inadequately scaled. We (loosely) refer to the analysis as distribution free because it does not include distributional assumptions associated with a test statistic. The adequacy of the centering and scaling of the data is, of course, relative. The re-loessed data appears to be adequately centered in that the probability that randomly selecting a null probeset such that the average expression value for the control samples is larger than that for the spike-in samples is approximately one half regardless of the overall signal intensity. The re-loessed data appears to be inadequately scaled in that the probability that randomly selecting a null probeset such that the variation in the expression values for the control samples is larger than that for the spike-in samples is less than one half. Further this variation is dependent on the overall signal intensity.

Figure 6 contains a panel of distribution free diagnostic plots to assess the adequacy of the centering and scaling of spike-in experiment (i.e., where true null fold changes are known) Stage II data. To evaluate relative centering, we propose modeling the probability that, for a randomly sampled probeset, the control samples will have a median expression value greater than the matched spike-in samples using the logit of a 4<sup>th</sup> order polynomial for rank-it intensity. To evaluate relative scaling, we propose modeling the probability that, for a randomly sampled probeset, the control samples will have a median absolute deviation (MAD) greater than the matched spike-in samples using the logit of a function of a 4<sup>th</sup> order polynomial for rank-it intensity. Given that there were only three replicates in the Golden Spike dataset we used the average of the two absolute deviations from the median value in place of the more common formulation of the MAD (which would have provided only the minimum of the two non-zero absolute deviations). The curves presented in Figure 6 correspond to the logistic regression fitted values. Note that the curves corresponding to the relative centering of the expression values (solid lines) are consistent with the biases observed in the t-statistics (depicted in Figure 4).

Table 1 contains results which indicate that the removal of the empty nulls from the invariant set provides for Stage II datasets which are adequately centered but are still inadequately scaled. For each of the 20 datasets considered, logistic models were fit as described above (i.e., the appropriate probability was modeled as the logit of a 4<sup>th</sup> order polynomial for rank-it intensity) and were tested against a null model that the appropriate probability was constant with respect to rank-it intensity. The deviances and asymptotic p-values (in bold) are reported. Given the possibility for cross hybridization of probesets, the assumption that the observed expression values are independent is dubious, although less tenuous than in a non-controlled experiment. The p-values, albeit approximate, indicate that the relationship between relative centering and intensity is highly significant in the original datasets and insignificant at a marginal level of 0.05 for all re-loessed datasets save 10e. The tabulated results indicate that the relationship between relative variability and intensity is highly significant for all datasets. However, the deviance values are significantly improved for several re-loessed datasets, most notably 10a-d.

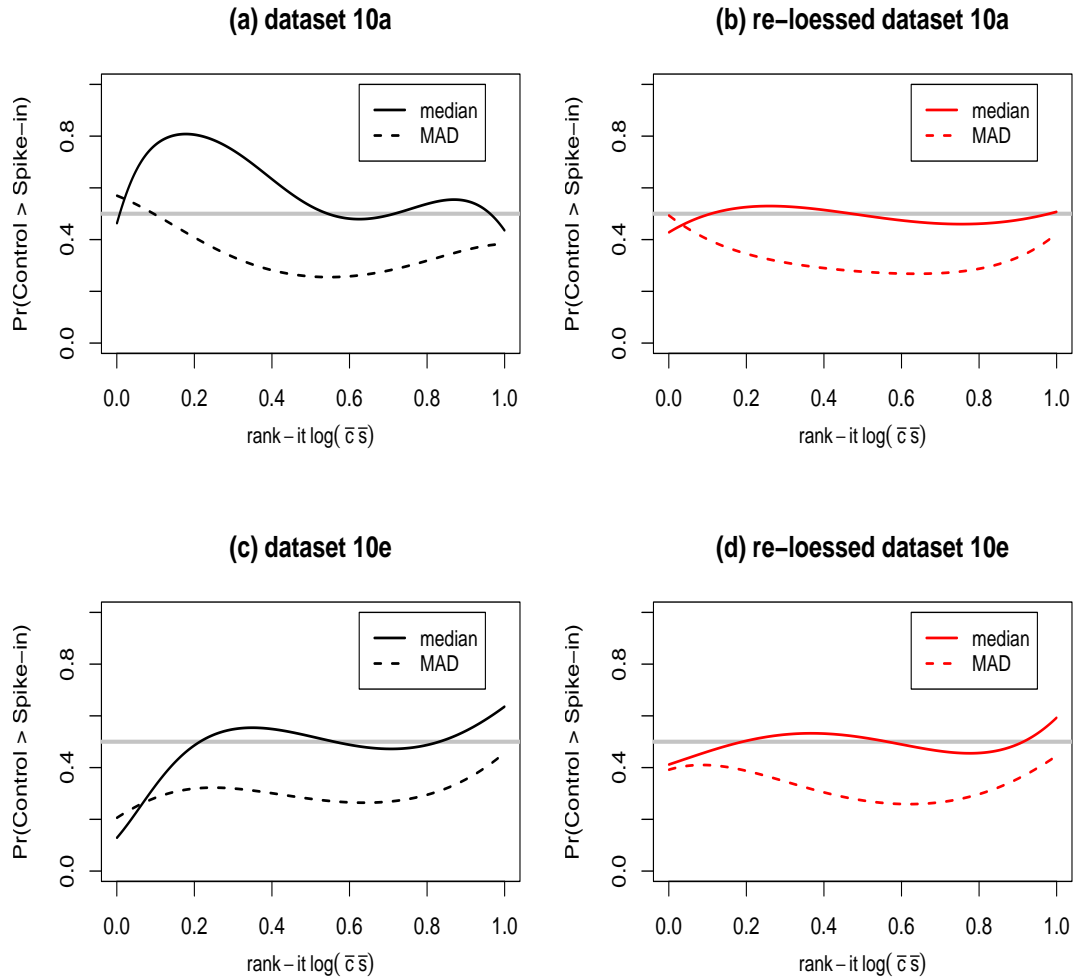


Figure 6: A diagnostic plot to assess, as a function of signal intensity, whether or not the underlying distributions for the control and spike-in expression values share the same center and scale. The x-axes correspond to the rank-it (i.e.,  $\frac{\text{rank of value}}{\text{total \# of values} + 1}$ ) of the log of the product of the expression means. The y-axes correspond to the probability that the control samples will have a value greater than the spike-in sample; values for the median and the MAD (median absolute deviation) were considered. The horizontal solid gray line corresponds to a probability of  $\frac{1}{2}$ . The probability that, for a randomly sampled probeset, the control samples will have a value greater than the matched spike-in samples was modeled as the logit of a 4<sup>th</sup> order polynomial for rank-it intensity. Solid (dashed) lines correspond to the logistic regression fits for the median (MAD). Diagnostic plots (a) and (c) indicate that, prior to re-loessing, the control and spike-in expression values were not equivalently centered and scaled for all signal intensities. Diagnostic plots (b) and (d) indicate that re-loessing the data provided control and spike-in expression values which were equivalently centered and but not equivalently scaled. Although loess correcting using only the set of true invariants can provide Stage II data which is adequately re-centered, issues pertaining to relative scale may remain and can invalidate the null distributions of commonly used two sample test statistics.

Table 1: Results of logistic regression for intensity dependence. The probability that the control samples will have a value greater than the matched spike-in samples was modeled as the logit of a function of a 4<sup>th</sup> order polynomial for rank-it intensity. Values for the median and the MAD (median absolute deviation) were considered. The deviances and p-values (in bold) for the comparison of the polynomial model to a constant null model are provided and are consistent with the results presented in Figure 6. Re-loessing the data using only the fold change 1 all but eliminates the relationships between intensity and relative centering of the two sample populations. However, the relationships between intensity and relative variability of the expression values remain, although they are greatly diminished.

dataset	Median		MAD	
	original	re-loess	original	re-loess
9a	184 ( <b>8.23e-39</b> )	4.37 ( <b>0.358</b> )	81.5 ( <b>8.28e-17</b> )	62.6 ( <b>8.26e-13</b> )
9b	246 ( <b>5.02e-52</b> )	3.2 ( <b>0.525</b> )	48.1 ( <b>9.06e-10</b> )	49.9 ( <b>3.79e-10</b> )
9c	225 ( <b>1.56e-47</b> )	3.41 ( <b>0.492</b> )	83.4 ( <b>3.26e-17</b> )	62.3 ( <b>9.6e-13</b> )
9d	271 ( <b>1.85e-57</b> )	4.02 ( <b>0.403</b> )	71.7 ( <b>9.93e-15</b> )	59 ( <b>4.73e-12</b> )
9e	104 ( <b>1.28e-21</b> )	7.61 ( <b>0.107</b> )	24.2 ( <b>7.38e-05</b> )	45.3 ( <b>3.37e-09</b> )
10a	151 ( <b>1e-31</b> )	6.61 ( <b>0.158</b> )	82.3 ( <b>5.69e-17</b> )	35.6 ( <b>3.47e-07</b> )
10b	190 ( <b>4.86e-40</b> )	3.19 ( <b>0.527</b> )	102 ( <b>4.63e-21</b> )	32.5 ( <b>1.54e-06</b> )
10c	214 ( <b>4.52e-45</b> )	8.12 ( <b>0.0874</b> )	124 ( <b>6.76e-26</b> )	47.7 ( <b>1.11e-09</b> )
10d	238 ( <b>2.1e-50</b> )	4.62 ( <b>0.329</b> )	157 ( <b>6.06e-33</b> )	39.4 ( <b>5.63e-08</b> )
10e	105 ( <b>8.49e-22</b> )	12 ( <b>0.0171</b> )	21.9 ( <b>0.000208</b> )	36.4 ( <b>2.43e-07</b> )

A set of diagnostic plots were created to assess whether the differences in relative centering and variability could be attributed to a one or two rogue samples. Figure 7 includes an example panel of the diagnostic plots for sample datasets 10a and 10e. These plots constitute a variation on the theme of the plots presented in Figure 6. The probability that a given sample will have an expression value greater than the median of the expression values for the balance of samples was modeled as the logit of a 4<sup>th</sup> order polynomial for rank-it intensity. Inspection of Figures 6(a) and (c) reveal that the within subpopulation (i.e., control and spike-in) logistic model fits are remarkably consistent for dataset 10a and are less so for dataset 10e. None of the plots support the hypothesis that a minority of the samples (i.e., one or two samples) are wildly inconsistent with the majority.

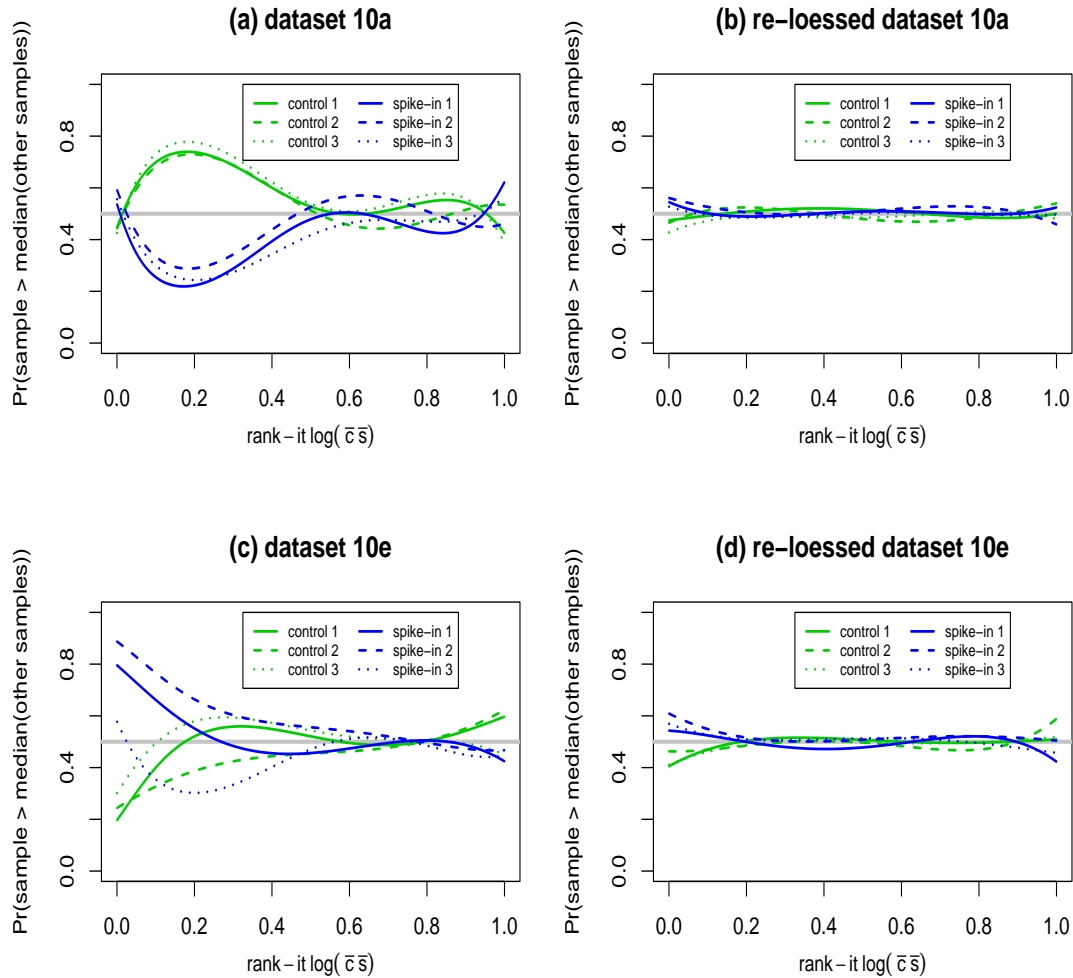


Figure 7: A diagnostic plot to assess, as a function of signal intensity, whether or not the underlying distributions for the expression values from each chip/sample share the same center. The x-axes correspond to the rank-it (i.e.,  $\frac{\text{rank of value}}{\text{total \# of values} + 1}$ ) of the log of the product of the expression means. The y-axes correspond to the probability that the observed expression value for a given sample will exceed the median of the expression values for the samples not under direct examination. The horizontal solid gray line corresponds to a probability of  $\frac{1}{2}$ . The probability that the sample under consideration will have an expression value greater than the median of the expression values for the samples not under direct examination, was modeled as the logit of a 4<sup>th</sup> order polynomial for rank-it intensity. Colored lines correspond to the logistic regression fit values. The within subpopulation logistic model fits are remarkably consistent for dataset 10a and are less so for dataset 10e. Plots (a) and (b) suggest that problems with relative centering can not be attributed to one or two “outlying” samples. Rather, these plots support the hypothesis that the Stage I pre-processing algorithms could not adequately adjust for differences in the underlying population distributions of the expression values for the empty probesets.

## 4 Discussion

The Golden Spike dataset was generated to address a dearth of controlled spiked-in array datasets. The original analysis of the data was presented in Choe et al. (2005) and concluded, among other things, that common methods to control the false discovery rate had failed to control at the nominal level. Dabney and Storey (2006) determined that the failure of the FDR algorithms was not methodological, rather the distributions of the null p-values corresponding to the most common choices for the Stage II analysis (e.g., t-test, permutation t-test, and Wilcoxon Rank Sum test) were non-uniform for the datasets which were considered. Dabney and Storey (2006) concluded that the Stage II model assumptions (e.g., that the denominator of t-test is appropriate estimator of the underlying variation) associated with these tests are not met, as the non-uniform p-value distributions imply that the technical replicates generated by the experiment do not constitute adequate approximations of biological replicates. We provide evidence that the non-uniform distributions arise, at least in part, from the failure of the Stage I analyses to correct for systematic biases in the raw data matrix, and that the p-value distributions are intensity dependent. We demonstrate that removing the empty probesets from the invariant set used in the analyses of Choe et al. (2005) can provide Stage II data which is adequately re-centered, a result which is dependent upon artificial knowledge of the true invariant set. Unfortunately, even under these ideal conditions, issues pertaining to higher moments (e.g., relative scale) remain and these issues appear to invalidate the null distributions of commonly used two sample test statistics.

Our analysis constitutes proof of principle that the distributions of the p-values, tests statistics, and probabilities associated with the relative locations and variabilities of the expression values can vary with signal intensity. This implies that Stage I algorithms do not always adequately adjust for intensity dependent effects. We provide diagnostic plots which are useful for testing the ability of Stage I analyses to create Stage II data matrices, in the case where the columns of these matrices are adequately centered and scaled with respect to one another. These plots should prove helpful for the analysis of the next generation of controlled spiked-in datasets.

It remains an open research question whether our findings apply more generally to the analyses of other micro-array datasets; they may simply indicate a phenomenon specific to the Golden Spike experiment. We suspect that the underlying problem in the relative variation of the expression values for the null genes might be a consequence of the unbalanced design of the experiment, wherein the spiked-in samples contained a larger amount of genetic materials. It is therefore not unreasonable to assume the possible existence of biological conditions which could engender imbalances similar to those observed in the Golden Spike dataset. For example, such imbalances could occur when comparing different tissue types, in cases of immune challenge or in certain developmental time course studies. Hence, we regard as plausible the assumption that residual intensity related effects of the type observed in the Golden Spike data could be present in other microarray Stage II datasets. We encourage efforts both to extend the work presented here, and to develop diagnostic plots which can provide useful inference on a dataset where the invariant subset of genes is truly unknown. Two notable complications present themselves in that setting: 1) it may not be reasonable to expect the proportion of discoveries to be constant with overall intensity (e.g., higher overall intensities may have a higher percentage of non-nulls), and 2) the logistic regression based test statistic may not be robust to the true correlation structure for the expression values. With respect to the first complication, an obvious extension would be to use weighted logistic regression where the weights depend upon the posterior probability estimates of differential expression.

## References

- Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003), “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias,” *Bioinformatics*, 19, 185–193, Evaluation Studies.
- Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M., and Halfon, M. S. (2005), “Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset,” *Genome Biol*, 6, R16.
- Dabney, A. R. and Storey, J. D. (2006), “A reanalysis of a published Affymetrix GeneChip control dataset,” *Genome Biol*, forthcoming.
- Hothorn, T. (2001), “On Exact Rank Tests in R,” *R News*, 1, 11–12.
- Hothorn, T. and Hornik, K. (2006), *exactRankTests: Exact Distributions for Rank and Permutation Tests*, R package version 0.8-12.
- Koenker, R. (2006), *quantreg: Quantile Regression*, R package version 3.85.
- Koenker, R. W. and D’Orey, V. (1987), “[Algorithm AS 229] Computing Regression Quantiles,” *Applied Statistics*, 36, 383–393.
- Schadt, E. E., Li, C., Ellis, B., and Wong, W. H. (2001), “Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data,” *J Cell Biochem Suppl*, Suppl 37, 120–125.

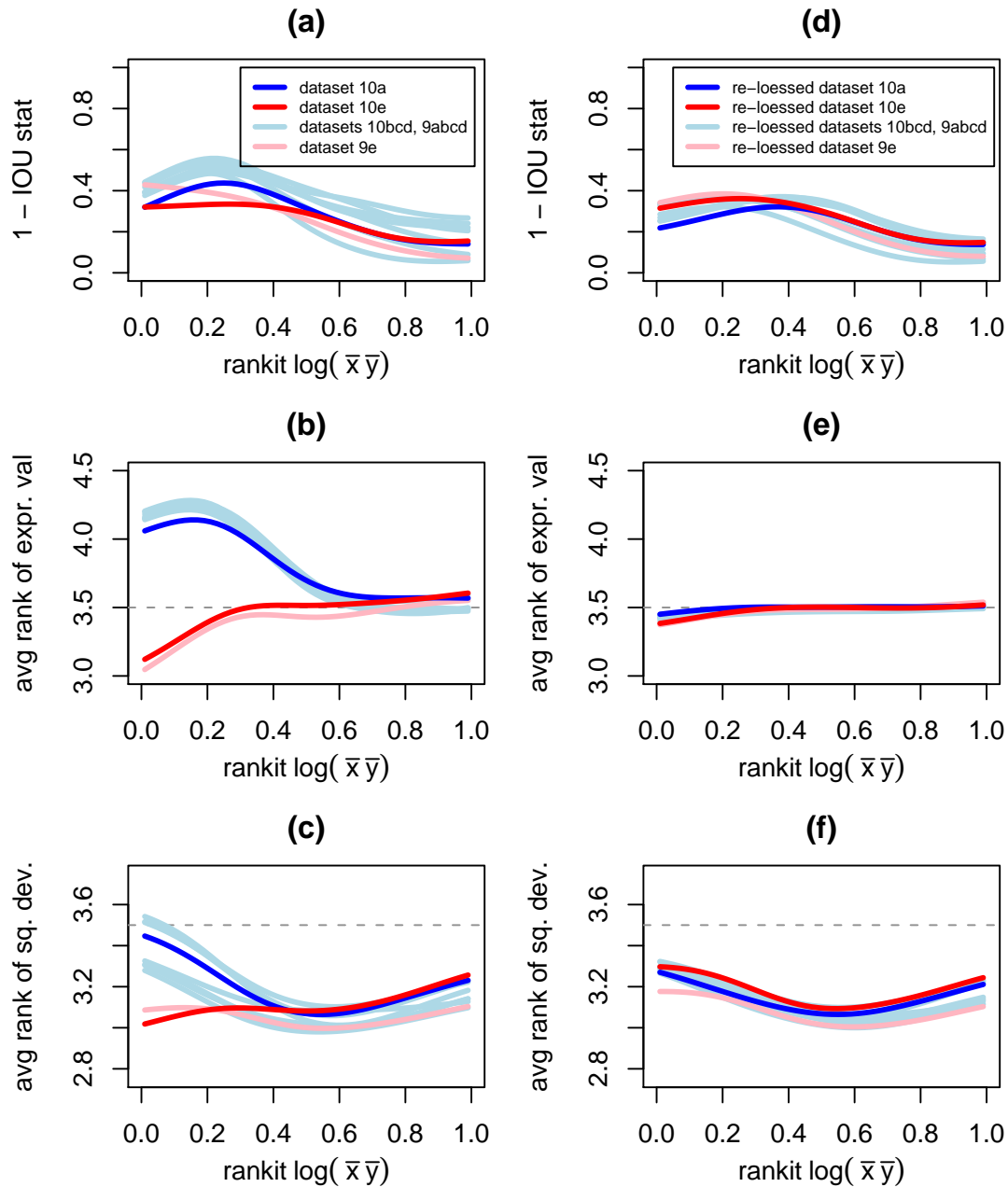


Figure 8: Smoothed estimates of the average rank of expression values and squared deviations (with respect to the appropriate group mean) of the three control replicates for the original (a, b) and re-loessed (c, d) datasets. The x-axes correspond to the rankit of the log of the product of the expression means. The y-axes correspond to the observed ranks and were calculated across all six samples. If the control (“C”) and spike-in (“S”) expression values are interchangeable, the average rank of the control values should be 3.5. (c) Re-loessing adequately re-centers the control expression values relative to the spike-in expression values. (d) However, despite re-loessing, the ranks of the squared deviations for the control replicates remain below those of the spiked-in replicates, suggesting that the expression values for the control replicates are less variable than those for the spiked-in replicates. This difference appears to be intensity dependent.