Correspondence

A reanalysis of a published Affymetrix GeneChip control dataset

Alan R Dabney and John D Storey

A response to **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset** by SE Choe, M Boutros, AM Michelson, GM Church and MS Halfon. *Genome Biology* 2005, **6:**R16.

Address: Department of Biostatistics, University of Washington, Seattle, WA 98195, USA.

Correspondence: John D Storey. Email: jstorey@u.washington.edu

Published: 22 March 2006

Genome ***Biology*** 2006, **7:**401 (doi:10.1186/gb-2006-7-3-401)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/3/401

In a recent *Genome Biology* article, Choe *et al*. [1] described a control dataset for Affymetrix GeneChips. By spiking RNA at known quantities, the identities of all null and differentially expressed genes are known exactly, as well as the fold change of differential expression. With the wealth of analysis methods available for microarray data, a control dataset would be very useful. Unfortunately, serious errors are evident in the Choe *et al*. data, disproving their conclusions and implying that the dataset cannot be used to validly evaluate statistical inference methods. We argue that problems in the dataset are at least partially due to a flaw in the experimental design.

q-value estimates incorrectly criticized

Figure 8 in Choe *et al*. [1] suggests that estimated *q*-values substantially underestimate the true *q*-values. For example, Figure 8b shows that, when a *q*-value cutoff of 0.20 is used, the true *q*-value is closer to 0.60. This reflects a serious error somewhere in the analysis. The question is whether the error occurs prior to, or as a result of, the *q*-value calculations.

False-discovery rates were originally proposed by Soric [2] and Benjamini and Hochberg [3]. The *q*-value was developed as the FDR analog of the *p*-value [4-7]. There is sound statistical justification behind both FDR and *q*-value methods; that is, there is rigorous mathematical proof for their claimed operating characteristics. For example, conditions have been detailed where the *q*-value estimates are guaranteed to be (simultaneously) conservative [5,7]. This means that, if a gene is assigned an estimated *q*-value of 0.05, then its true, population average, *q*-value is no larger than 0.05.

Note that the *q*-value methodology employed by Choe *et al*. [1] was credited to the SAM method proposed by Tusher *et al*. [8], even though the *q*-value methodology was developed separately from the SAM method in references [4-7]. The SAM software (different from the SAM method [8]) is based on at least four different articles [4,8-10], each of which contributes unique methodology. Failing to differentiate between the SAM method of Tusher *et al*. and the SAM software seems to have caused a good deal of confusion in the literature when evaluating and understanding the operating characteristics of the methods in the software [11].

We now show that the fundamental requirements for employing *q*-values (and even *p*-values) are not satisfied by Choe *et al*.'s dataset, leading to spurious conclusions. These requirements are stated in each of the original papers detailing the *q*-value methodology [4-7]. In particular, the *q*-value methodology draws on the fact that the *p*-values for null genes should be uniformly distributed on the interval (0,1) [12]. As stated by Storey and Tibshirani [5]: "If the null *p*-values are not uniformly distributed, then one wants to err in the direction of overestimating *p*-values (that is, underestimating significance). Correctly calculated *p*-values are an important assumption underlying our methodology." This is not a special requirement for *q*-values - it is a requirement for any type of standard significance criterion [13].

Choe *et al*. [1] identified 10 combinations of pre-inference steps that seemed to work best. Their *q*-value calculations were then done on these "10 best" datasets. We reproduced their analyses of each dataset using standard two-sample *t*-statistics (conclusions did not depend on whether parametric, permutation, or bootstrap null distributions were used). Figure 1 of this correspondence compares the observed quantiles of the null genes' *p*-values to the corresponding quantiles from the uniform distribution. If the null

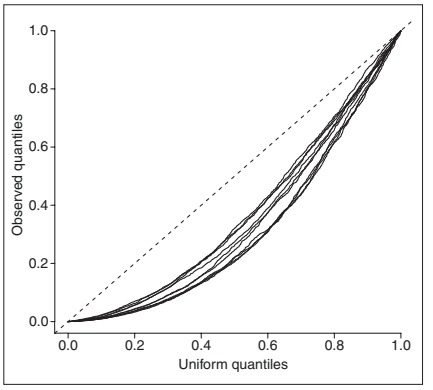

Figure 1
QQ plots of null *p*-values corresponding to null genes. A plot of the observed versus expected quantiles of the null genes' *p*-values are shown for each of the 10 best datasets. The observed trends indicate that the null genes' *p*-values trend are substantially smaller than they should be.

p-values were distributed uniformly, then the observed quantiles should fall along the dashed line of equality. It is clear that the null *p*-values are not uniformly distributed, tending to be much smaller than they should be. In other words, when observing the *p*-value distribution among the null genes (which are known here), they show a substantial amount of significance beyond what would be expected by chance. It is therefore clear that the problems evident in Figure 8 of Choe *et al.* are not due to the *q*-value methodology, but rather to the fact that the calculated *p*-values are not valid.

Problems with the experimental design

The question remains as to the cause of the *p*-values being incorrect. One source of the problem is the experimental design itself. Consider Figure 1 in Choe *et al.* [1]. In column three, there are three aliquots of labeled cRNA clones. Each of these aliquots is divided, and one half is then spiked-in with known concentrations of RNA among the genes selected to be differentially expressed. Because random variability is introduced in the spike-in step (between columns 3 and 4 in Figure 1 of Choe *et al.* [1]), even null genes will have some differences in RNA amount. That is, both halves of each aliquot undergo some modification (even the control half), leading to random variation being introduced to the RNA amounts of all genes. This leaves three matched pairs of independent samples, where some variation exists within pairs for all genes.

A major flaw occurs when the three samples from each treatment (control or spike-in) are combined into two aliquots in column 4. Now, instead of three independent matched pairs, there is only a single matched pair in column 4. Each half of this single matched pair is hybridized to three chips. Therefore, the random variation introduced at the spike-in stage will not be detectable among the six resulting chips. If every chip is treated as an independent observation, then the variation introduced in the previous step among null genes will appear to be true signal. Unfortunately, this is exactly what Choe *et al.* do, leading to the incorrect *p*-values that we observed earlier. This problem cannot be fixed by modifying the statistics.

Consider the following scenario, which suffers from the same problem as Choe *et al.*'s design but is phrased in more familiar terminology. Suppose we are interested in differences in gene expression for two biological groups under two conditions, A and B. To this end, we obtain expression measurements for a single individual (that is, a single biological replicate) under both conditions. On the basis of the sample from this single individual, we then form three replicated chips for each condition.

By chance, there will be small differences in the expression measurements under both conditions A and B. With a proper estimate of the variability, these differences will be identified as being random and the associated tests called null. The problem is that we cannot use our single individual to estimate the true variability of expression measurements under conditions A and B, taking into account all sources of variation. If we treat the six replicated observations as three independent matched pairs, then our estimated variance is due only to random aspects of the hybridization process. This will significantly underestimate the true variances, and, as a

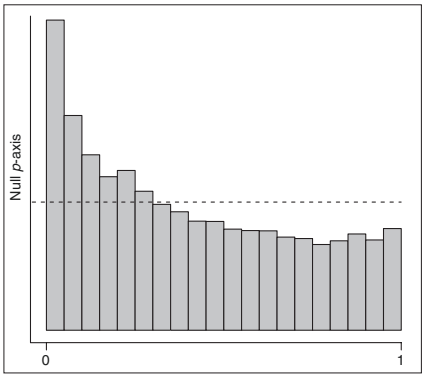

Figure 2
Histograms of null *p*-values from simulation representing the experimental design of Choe *et al.* [1]. The null *p*-values generated from the simulation as described in the text are shown. The dashed line represents the expected height of the bars assuming the null *p*-values are uniformly distributed. The null *p*-values are not uniformly distributed when only technical replicates are used.

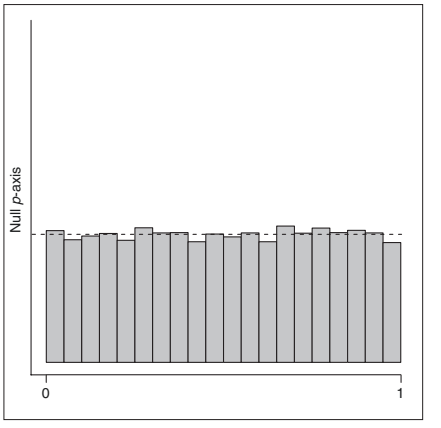

Figure 3
Histograms of null *p*-values from simulation based on independent samples. The null *p*-values using three independently sampled individuals as described in the text are shown. The dashed line represents the expected height of the bars assuming the null *p*-values are uniformly distributed. The null *p*-values are uniformly distributed when biological replicates are used.

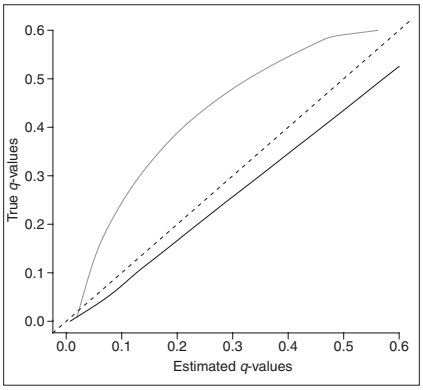

Figure 4
A plot of the true versus estimated *q*-values from simulations described in the text. The solid gray line shows the results averaged over 30 simulations when using a design similar to that of Choe *et al.* [1]. The solid black line is the analogous comparison when using three independent individuals. The dashed line represents equality; conservatively estimated *q* values should fall beneath this line. The Choe *et al.* [1] design produces anti-conservative *q*-values estimates due to the incorrect underlying *p*-values, while the more statistically sound design produces conservative *q*-value estimates. The Monte Carlo variation of the *q*-value estimates is small enough that these conclusions are not affected.

result, we will grossly inflate the significance of differential expression - even among null genes.

We performed a simple simulation of the above scenario; details are given in Additional data file 1. We considered both the case where three technical replicates are formed on a single individual and the case where three independently sampled individuals were used. Figure 2 of this correspondence shows a histogram of the null *p*-values under the first scenario where only technical replicates are used, and Figure 3 shows the results using biological replicates. It is clear that the null *p*-values are not uniformly distributed under Choe *et al.*'s design (Figure 2), tending to be much smaller than they should be. Figure 4 compares estimated and actual *q*-values, analogous to Figure 8b in Choe *et al.* [1]. When a single individual is used, we see that *q*-values substantially underestimate the truth due to the flawed underlying *p*-values. Meanwhile, when three independent individuals are used, the estimated *q*-values are conservative as the theory says they should be [4-7].

A large-scale "spike-in" control dataset would be invaluable for head-to-head comparisons of statistical methods for microarrays. The Choe *et al.* dataset was intended to serve this purpose. Unfortunately, the data set is flawed in that even the null genes appear to be differentially expressed. As a result, the dataset cannot be relied upon for evaluating statistical inference methods. We note further that, when applying statistical methods such as the *q*-value estimates, one must take care to ensure that all necessary assumptions are met.

Additional data file

Additional data file 1 available with this paper provides details of the simulations reported here.

Sung E Choe, Michael Boutros, Alan M Michelson, George M Church and Marc S Halfon respond:

One of the main purposes of our paper [1] was to challenge the community to improve on existing microarray analysis methods and to promote a better understanding of the experimental conditions for which these methods are appropriate. Dabney and Storey make a valuable contribution to this effort with their clarification as to why our *q*-value calculations underestimate the false-discovery rate by noting that the underlying *p*-value distributions for the "null genes" - those with no expected fold change between the S spike (S) and control (C) samples - are non-uniform in each of the 10 best datasets considered in our original manuscript [1]. Dabney and Storey also provide simulation results to suggest that the problem is due to our use of technical replicates. However, we demonstrate

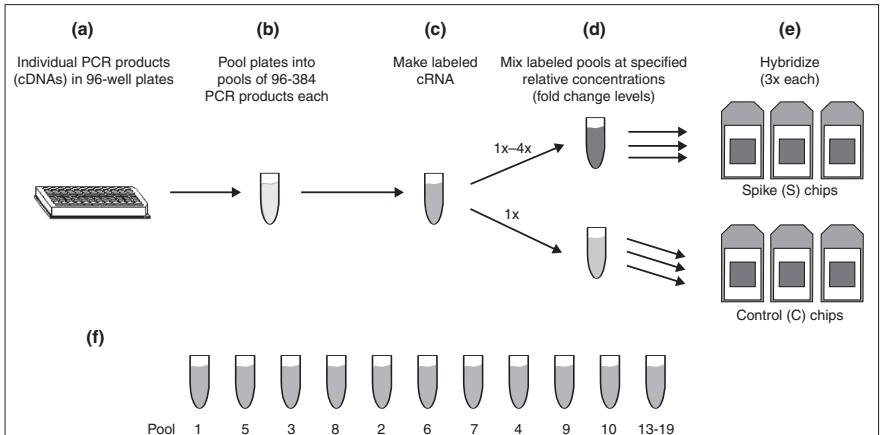

Figure 5
A detailed description of the Choe *et al.* [1] experiment. Individual PCR products **(a)** were pooled together **(b)** and converted to labeled cRNA **(c)**. Note that all mixing and labeling within each pool was performed at this stage, before splitting the pools into C and S samples. Therefore, relative concentrations of individual cRNA species are identical for all cRNAs in a given pool. **(d)** The labeled pools were then divided into the C and S samples. Poly(C) RNA (20 μg) was added to the C sample at this step to equalize the amount of nucleic acid present in each hybridization. **(e)** Each sample contained enough labeled cRNA for three hybridizations. Relative concentrations for each pool are shown in **(f)**. Note that the 1x ("null gene") pools 13-19 were combined together at step (b), before labeling at step (c), creating a single 1x pool before labeling and splitting. The '1x' concentration of RNA used for this pool was approximately 6x greater than the 1x concentrations of the other pools to reflect the greater number of individual RNAs (that is, so that the 1x concentrations of all RNAs were approximately equal).

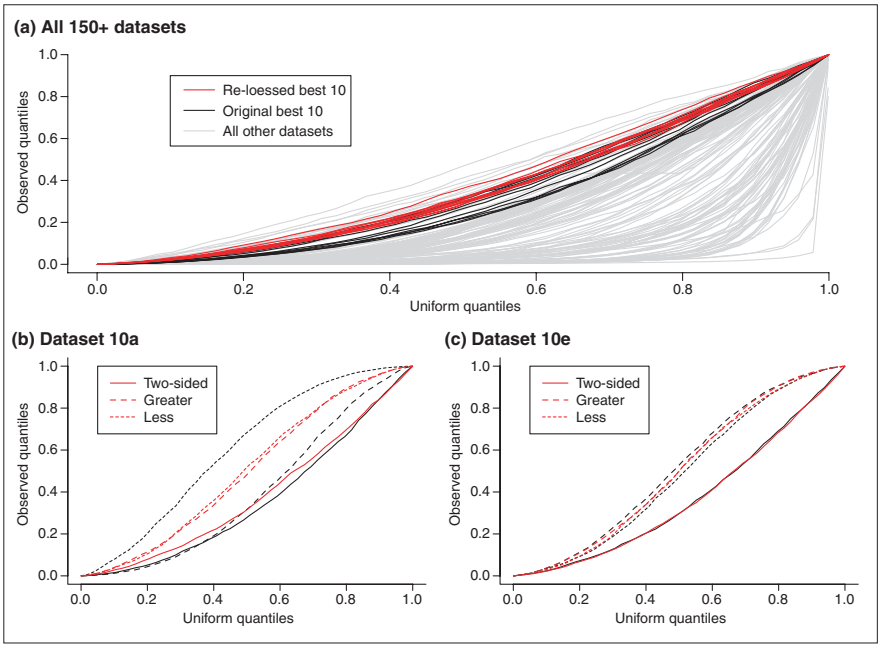

Figure 6
Sample quantile plots for the *p*-values of the observed test statistics for the "null genes". The *x*-axes correspond to the expected quantiles for a uniform distribution and the *y*-axes correspond to the observed (sample) quantiles. **(a)** Sample quantile plots for the *t*-test *p*-values associated with the 150 preprocessing combinations described by Choe *et al.* [1]. Black lines correspond to the 10 best datasets and are consistent with the curves presented in Figure 1 of this correspondence. 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) probe sets (versus both the 'present null' and 'empty null' probe sets used in [1]). 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. Dabney and Storey's model 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. Used with permission from [15].

here that this aspect of our design is sound, and that their model is consistent neither with our actual experimental design nor with the observed distribution of the null *p*-values. We also offer a more likely explanation for the problems that they note.

Dabney and Storey claim that "serious errors are evident in the Choe *et al.* data" due to a "flaw in the experimental design" concerning technical instead of independent replication, and they provide simulation results to support their view. They fail, however, to provide justification for their simulation parameters, which appear to be greatly exaggerated and inconsistent with our actual experimental design. This is especially true with respect to their value for ρ (the correlation between the S and C concentrations). Whereas Dabney and Storey place ρ at 0.85 (see their Additional data file 1), in reality it should be close to 1.0, as our design ensures that cRNAs from the same pool have virtually identical fold changes between the S and C samples. The "null genes", listed in the original paper as pools 13-19, were combined into a single pool before labeling and division into the S and C samples (Figure 5). In other words, all of the null genes were partitioned as a group into either the S or C samples. Common intuition as well as standard experimental practice tell us that the variation in fold change for genes from the same pool will be negligible, but we can also estimate it as follows. Consider the set of RNAs at the detection limit of the assay, which Affymetrix puts at 1.5 pM [14]. Assuming that equal partitioning from the null gene pool to the S and C samples follows a binomial distribution with $p = 0.5$, the standard deviation in RNA amount is only slightly more than 10^4 molecules for the approximately 10^8 molecules in the pool. Thus there is no appreciable variation in the amount of any given RNA species (nor in the fold changes between the S and C samples for cRNAs within the same pool). As there was a single pipetting event from the null gene pool to each of the S and C samples, there is some uncertainty as to the exact ratio of allocation to S versus C; however, this degree of uncertainty is low (less than 0.3% according to the pipette manufacturer's specifications). Importantly, this is manifest as a scaling error that affects the pool of null genes as a whole - for example, the true ratio might be 1:0.997 for all null genes, rather than 1:1. Any such differences, if actually detectable above the variation introduced by hybridization itself, were resolved by using our knowledge of the null genes to normalize between arrays whenever applicable. The dominant source of variation in our experiment is indeed, by design, that due to "aspects of the hybridization process". When parameters in keeping with a greatly reduced amount of variation are used in Dabney and Storey's model, the *p*-value distribution is uniform (data not shown). The observed non-uniformity therefore cannot be attributed to our use of technical replication.

A non-uniform null *p*-value distribution does not necessarily invalidate our dataset; it may simply indicate that the analysis methods we applied do not adequately model the hybridization process. In this regard, we note that the models proposed by Dabney and Storey are not truly consistent with the observed null gene *p*-values in our

data. Their model cannot explain unexpected discrepancies present in the distributions of the one-sided *p*-values, and neither can it explain the fact that the actual distributions of the null gene *t*-statistics and *p*-values appear to depend upon signal intensity. Figure 6a of this correspondence shows quantile plots for the *t*-test *p*-values associated with the 150 preprocessing combinations we used. The lines highlighted in black correspond to the "10 best datasets" and are consistent with the curves presented in Figure 1 of this correspondence. The black curves in Figures 6b and 6c contain sample quantile curves; the solid lines correspond to the two-sided *p*-values and the dashed and dotted lines correspond to the *p*-values from the one-sided tests. Note the discrepancy in the one-sided *p*-values in Figure 6b. This was observed for eight of the 10 datasets and is not accounted for in the Dabney and Storey model, under which the distributions of these *p*-values should be similar. This discrepancy appears to follow from the fact that each of the 10 best Choe *et al*. datasets were obtained using preprocessing steps that included loess correcting the observed intensity values from a set of "null genes" that included both "empty null" (not present in either sample) and "present null" (present with fold change = 1) probe sets. We have calculated a new set of 10 best datasets in which the correction is based only on the present null probesets (Figure 5, red curves). This 're-loessing' of the data eliminates the discrepancy in the one sided *p*-values (Figure 6a,b) and provides proof of principle that preprocessing algorithms can have a substantial effect on the *p*-value distribution.

Although re-loessing the data improved the null distributions, they are still non-uniform. If the model proposed by Dabney and Storey is neither consistent with the actual experimental design (their parameter values are not realistic) nor consistent with the observed data (the real one-sided *p*-values are dissimilar), what does account for the non-uniform distribution of the

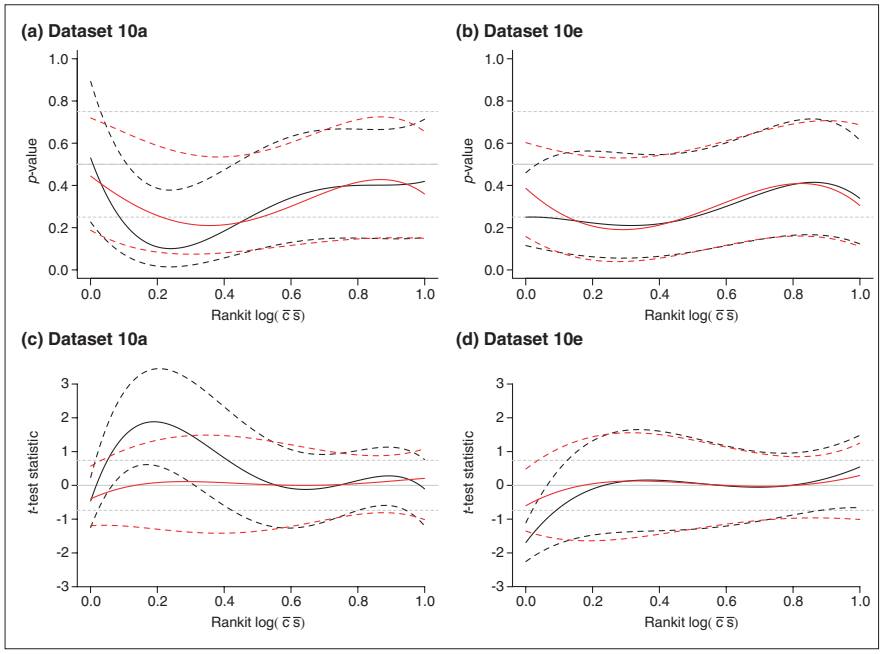

Figure 7
Smoothed estimates of the quartiles as a function of signal intensity for the *p*-values and observed *t*-test statistics for **(a,c)** dataset 10a and **(b,d)** dataset 10e. Distributions of the null *p*-values and test statistics vary with intensity, although less so for the re-loessed datasets (red lines). Even though the medians for the *t*-statistics seem to be properly centered after re-loessing the data, the quartiles (and hence the variation) are greatly inflated and appear to be intensity dependent. The *x*-axes show the rankit of the log of the product of the expression means. The *y*-axes show the observed two-sided *p*-values (a,b) or the observed *t*-statistics (c,d). Solid and dashed gray lines indicate the theoretical medians and quartiles, respectively. Black curves correspond to the original datasets and red curves correspond to the re-loessed datasets. Used with permission from [15].

p-values? We find that the distributions of the null gene *t*-statistics and *p*-values depend upon signal intensity. Figure 7 of this correspondence demonstrates this point using smoothed estimates of the *p*-value and *t*-score quartiles as a function of signal intensity in representative datasets. Figure 7c shows that the re-loessed dataset 10a provides for *t*-tests that are better centered than the original dataset, an observation consistent with the results depicted in Figure 6b. While the medians for the *t*-statistics seem to be properly centered after re-loessing the data, the quartiles (and hence the amount of variation) remain greatly inflated, however, and change with intensity.

We speculate that the preprocessing algorithms are unable to properly adjust for systematic differences in overall signal strength that exist between the control and spike-in samples. Figure 8 of this correspondence contains smoothed estimates of the average rank of expression values and squared deviations (with respect to the appropriate group mean) of the three control ("C") replicates for the original (Figure 8a,b) and re-loessed (Figure 8c,d) datasets. Comparison of Figures 8a and 8c reveals that re-loessing appears to adequately recenter the control expression values relative to the spike-in (S) expression values (the average rank over six arrays should be 3.5). The ranks of the squared deviations for the control replicates, however, remain below those of the spike-in replicates, suggesting that the control expression values are less variable than the spike-in values. This difference again appears to be intensity dependent.

The preceding analysis suggests that the observed non-uniformity of the *p*-values is not easily explained by a

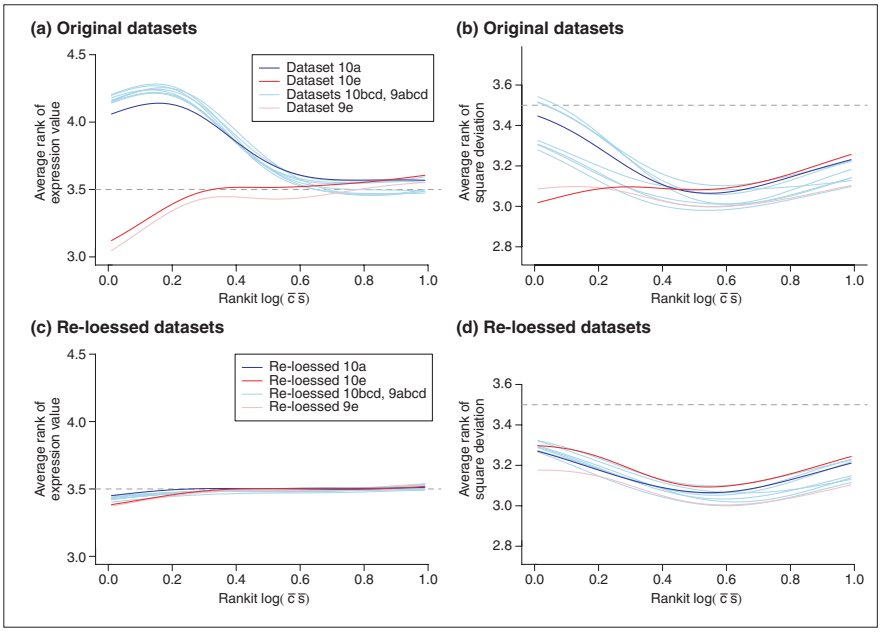

Figure 8
Smoothed estimates of the average rank of expression values and squared deviations (with respect to the appropriate group mean) of the three control replicates for the **(a,b)** original and **(c,d)** re-loessed datasets. The *x*-axes correspond to the rankit of the log of the product of the 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) Despite re-loessing, however, 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. Used with permission from [15].

simple model. Although relevant mainly to just one aspect of our study (regarding *q*-value estimates), this is an important and complex issue in need of further investigation, and we are grateful to Dabney and Storey for bringing it to our attention. Whether these non-uniform *p*-values are manifest in other datasets or are merely a byproduct of the unbalanced signal strengths present in our experiment also remains an open question to be addressed by future studies. However, in most microarray experiments it is impossible to check if the null *p*-values are uniformly distributed (as the true set of null genes is not known) and it is therefore impossible to determine whether or not the requirements for accurate estimation of *q*-values are met. We therefore caution that the preprocessing issues seen here might be relevant in other microarray experiments. We can easily conceive of biological conditions in which imbalances similar to those in our original study could exist - for example, when comparing different tissue types, in certain developmental time courses, or in cases of immune challenge.

Acknowledgements

We are indebted to Daniel Gaile and Jeffrey Miecznikowski of the University at Buffalo Department of Biostatistics for the analysis presented in Figures 6-8 and for their permission to use material from [15] in this response.

Sung E Choe, Michael Boutros, Alan M Michelson, George M Church and Marc S Halfon

Correspondence should be sent to Marc S Halfon: Department of Biochemistry and Center of Excellence in Bioinformatics and the Life Sciences, State University of New York at Buffalo, Buffalo, NY 14214, USA.
E-mail: mshalfon@buffalo.edu

References

1. Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS: **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.** *Genome Biol* 2005, **6:**R16.
2. Soric B: **Statistical discoveries and effect-size estimation.** *J Am Stat Ass* 1989, **84:**608-610.
3. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *J Roy Stat Soc, Ser B* 1995, **57:**289-300.
4. Storey JD: **A direct approach to false discovery rates.** *J Roy Stat Soc, Ser B* 2002, **64:**479-498.
5. Storey JD, Tibshirani R: **Statistical significance for genome-wide studies.** *Proc Natl Acad Sci USA* 2003, **100:**9440-9445.
6. Storey JD: **The positive false discovery rate: A Bayesian interpretation and the *q*-value.** *Annls Stat* 2003, **31:**2013-2035.
7. Storey JD, Taylor JE, Siegmund D: **Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach.** *J Roy Stat Soc, Ser B* 2004, **66:**187-205.
8. Tusher V, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98:**5116-5124.
9. Efron B, Tibshirani R, Storey JD, Tusher V: **Empirical Bayes analysis of a microarray experiment.** *J Am Stat Ass* 2001, **96:**1151-1160.
10. Taylor J, Tibshirani R, Efron B: **The miss rate for the analysis of gene expression data.** *Biostatistics* 2005, **6:**111-117.
11. Dudoit S, Shaffer JP, Boldrick JC: **Multiple hypothesis testing in microarray experiments.** *Stat Sci* 2003, **18:**71-103.
12. Lehmann EL: *Testing Statistical Hypotheses.* Berlin: Springer; 1997.
13. Rice JA: *Mathematical Statistics and Data Analysis* 2nd edition. Belmont California: Duxbury Press, 1995.
14. Affymetrix: *GeneChip Expression Analysis: Technical Manual.* Santa Clara, CA: Affymetrix; 2004.
15. Gaile DP, Miecznikowski JC, Choe SE, Halfon MS: **Putative null distributions corresponding to tests of differential expression in the Golden Spike dataset are intensity dependent. Technical report 06-01.** Buffalo, NY: Department of Biostatistics, State University of NewYork; 2006. [http://sphhp.buffalo.edu/biostat/research/techreports/UB_Biostatistics_TR0601.pdf]