

Intra-Slide Correlation and Treatment Significance in Microarray Data: The Impact of Loess Normalization

Eric R. Siegel¹, John Thaden², & Pippa M. Simpson³

Departments of ¹Biostatistics, ²Geriatrics, and ³Pediatrics,
University of Arkansas for Medical Sciences, Little Rock AR 72205-7199

ABSTRACT:

Locally weighted regression (loess, lowess) is often used with demonstrated effectiveness to "normalize" or correct microarray-derived gene expression data for spot intensity differences, spatial effects, and dye bias. Less is known about its impact on systematic variation due to treatment. We used variance components analysis to study the impact of loess normalization on intra-slide correlation and treatment significance. Our data came from two sources: Stanford Microarray Database (SMD) microarray slides probing more than 17,000 *Caenorhabditis elegans* genes, with more than 1,000 probe species spotted at least twice per slide, and UAMS Microarray Core slides probing almost 10,000 *Rattus norvegicus* genes, with all probes spotted twice or more per slide. We subjected the data to no adjustment, to two global adjustments (median centering or mixed-models normalization) or to three loess adjustments (by slide, by print-tip, or by print-tip with rescaling to a common interquartile range). For each adjustment, we estimated the intra-slide correlation of each replicated probe as the ratio of its between-slide variance to the sum of its between-slide and within-slide variances. We then derived Type III p-values for Treatment Effect from a mixed-models analysis of each probe. When we compared the resulting correlation distributions, we found that loess adjustments produced stochastically lower intra-slide correlations than global adjustments or no adjustment. However, when we compared the resulting p-value distributions, we found that the effect of the adjustment depended on both the data set and the quantile of the distribution. We conclude that one cannot use the loess effect on intra-slide correlation to predict its effect on treatment significance.

INTRODUCTION:

Variation is the spice of statistics. But normalization decreases variation, so why do it? In the ideal microarray experiment, differences in red and green fluorescence intensity at each spot would be due solely to treatment-related differences in gene expression between the pair of samples hybridizing to each slide. In the real world, fluorescence intensity differences are also affected by factors arising from the technology. These factors can include differences in the labelling efficiencies and scanning properties of the two fluorophores, differences among print tips on the array printer, and variation over the course of the print run. Additionally, differences in print quality, in ambient conditions when the slides were processed, and changes in scanner settings can (and often do) contribute to the intensity differences. Normalization refers to the process of correcting for such

technical variation, so that intensity difference due to biology can shine through. The assumption behind normalization is that only a few genes on the array will show a treatment effect, whereas most genes won't. Large-scale and/or systematic differences among slides are thus assumed to derive from technical factors; remove them mathematically, and only the biologically driven differences should remain.

The simplest normalization procedures consist of location-scale transformations designed to give each slide the same mean or median, and sometimes also the same variance or other dispersion measure. These methods correct for technical factors that act homogeneously, but not for factors that vary in strength with intensity or with spatial location on the slide. In 2001, Yang *et al.* demonstrated how to use locally weighted regression methods to remove heterogeneous technical variation. Since then, normalization by local regression methods (such as the "loess" and "lowess" procedures) have become a commonplace in microarray analysis.

There is no doubt that normalization by loess and/or lowess regression removes technical variation from microarray data aggressively and effectively. We suspected, however, that such aggressive methods may also remove, or distort, some of the variation due to the treatment of interest. And apart from rare and ultimately unconvincing spike-in experiments, we had seen nothing in the literature to assuage our suspicion. Our own efforts to examine this issue began with a key insight that, in a well-mixed sample, the ratio of red to green fluorescence for a particular probe should be the same for replicate spottings of that probe on the same slide regardless of quality differences among replicates. This is equivalent to saying that replicate spots on the same slide should have red-green ratios that cluster about an expected value, which leads naturally to a variance-components model in which the between-slide variance component contains the treatment effect. Thus, by studying how local-regression normalizations affect between-slide variance, we may gain insight into their effect on treatment significance. In this paper, we report the results of such a study, using a standardized form of the between-slide variance based on the intraclass correlation coefficient (Donner 1986).

METHODS:

The *C. elegans* data were from a biological experiment performed in duplicate (JT). The mRNA was hybridized to three different slide lots, at three times, by three different technicians in the Stanford laboratory of Stuart Kim, resulting in a challenging data set. *R. norvegicus* data were provided by the UAMS Microarray Core Facility with the permission of Dr. Richard Kennedy. Rat mRNA was hybridized using one slide lot in one sitting to create a more typical data set. Slides

for both genomes had 32 print-tip groups of 624 spots each. On the *R. norvegicus* slides, the 16 print-tip groups in the top half of the slide were replicated in the bottom half. Images were converted into spot intensities using the seeded region growing algorithm within SPOT v. 1.0 (CSIRO) for the *C. elegans* slides and the adaptive circles algorithm of ScanArray Express (PerkinElmer) for the *R. norvegicus* slides. Intensities were not adjusted for background. SAS versions 8.2 and 9.1 were used for all normalizations and statistical analyses. Intensities were transformed to their base-2 logarithms prior to normalization. The mixed-models normalization was performed using the approach of Wolfinger *et al.* (2001):

$$y_{gik} = \mu + D_i + S_k + (DS)_{ik} + \gamma_{gik}$$

(see also <http://statgen.ncsu.edu/ggibson/Manual.htm>), in which D_i is dye color, S_k is slide, $(DS)_{ik}$ is the dye-by-slide interaction, and y_{gik} and γ_{gik} are the original and normalized log-intensities, respectively. For the other normalizations, the y_{gik} were combined by dye color D_i into average logs A_{gk} and "logratios" (differences in logs) M_{gk} as described by Yang *et al.* (2001) for their M-vs-A plots. To normalize by median centering, we subtracted from the M_{gk} and A_{gk} their respective median values for each slide. Normalization by loess regression was performed, either by whole slide or by print-tip group, by extracting the residuals η_{gi} from the equation:

$$M_{gk} = f(A_{gk}) + \eta_{gk}$$

in which $f(A_{gk})$ is the loess predictor fitted by local 1st-order polynomial regression with smoothing parameter chosen from within the interval [0.2,0.5] by the AIC_c criterion of Hurvitch *et al.* (1998). The A_{gk} were median-centered after all three loess regressions. Rescaling was accomplished as follows: the η_{gi} from loess regression by print-tip group were divided by their print-tip interquartile ranges and multiplied by the experiment-wide median of those interquartile ranges. The residuals γ_{gik} from the mixed-models normalization were also combined by dye color into averages and differences for comparison with results from the other normalizations. To facilitate further discussion, we refer to all the normalization results as M_{gk}^* and A_{gk}^* , adjusted versions of the average logs and logratios.

Slides were assigned a polarity P_k of +1 or -1 based on how treatments matched to dye color. For each gene g printed more than once per slide, variance components σ_{slide}^* and σ_{ϵ}^2 were extracted using the random-effects model:

$$P_k M_{ik}^* = \mu + slide_i + \epsilon_{ik}$$

in which $P_k M_{ik}^*$ denotes the adjusted logratio multiplied by slide polarity, and in which σ_{slide}^* was allowed to assume negative values. The σ_{slide}^* term for each gene after each normalization was expressed in standardized form as the intra-slide correlation coefficient:

$$\rho = \sigma_{slide}^* / (\sigma_{slide}^* + \sigma_{\epsilon}^2)$$

The impact of the different normalizations on intra-slide correlation was assessed both visually and by Kruskal-Wallis test. For graphical visualization, the set of correlations from each normalization were assigned an empirical distribution via fractional ranks under the ties=high rule.

To examine how the normalizations affected treatment significance, the M_{gk}^* and A_{gk}^* were back-transformed to adjusted log-intensities y_{gik}^* . These were then subjected to mixed-models analysis on a gene-by-gene basis using the approach of Wolfinger *et al.* (2001). Mixed models for each gene were as follows, with fixed effects in bold:

- $y_{hijk}^* = \mu + lot_h + \mathbf{dye}_i + \mathbf{dye} * lot_{hi} + \mathbf{slide}(lot)_{hj} + \mathbf{treatment}_k + spot(slides)_{hj} + \epsilon_{hijk}$, for the replicate-spot genes on the *C. elegans* slides.
- $y_{hijk}^* = \mu + \mathbf{dye}_i + \mathbf{slide}(lot)_{hj} + \mathbf{treatment}_k + \epsilon_{hijk}$, for the singleton-spot genes on the *C. elegans* slides.
- $y_{hijkl}^* = \mu + \mathbf{dye}_h + \mathbf{treatment}_i + \mathbf{rat}_{ij} + \mathbf{slide}(rat)_{ijk} + spot(slides)_{ijkl} + \epsilon_{hijkl}$, for the duplicated spots on the *R. norvegicus* slides.

These models were used to generate Type III p-values for treatment effect for each gene after each normalization. The impact on the p-values of the different normalizations was assessed visually. For graphical visualization, the set of p-values from each normalization were assigned an empirical distribution via fractional ranks under the ties=high rule.

RESULTS:

Correlation Distributions, both data sets.

Figure 1 shows the distribution of intra-slide correlations for the ~1,000 replicate spots per *C. elegans* slide, and Figure 2 shows the distribution of intra-slide correlations for the 9984 duplicate spots per *R. norvegicus* slide. Red denotes distributions after loess adjustments, blue denotes distributions after global adjustments, and black denotes distributions after no adjustment.

Figure 1

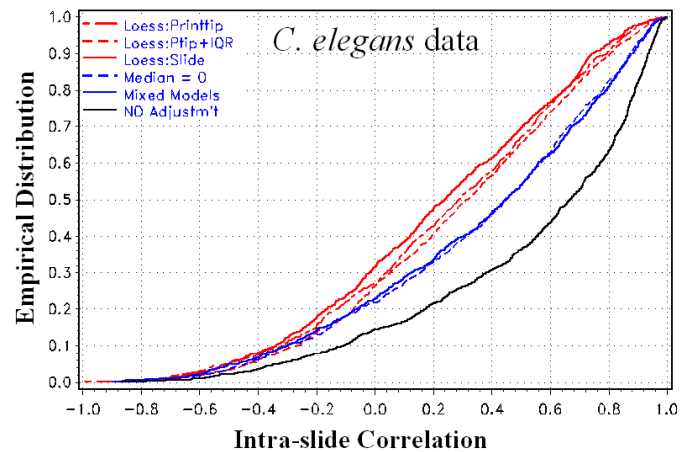
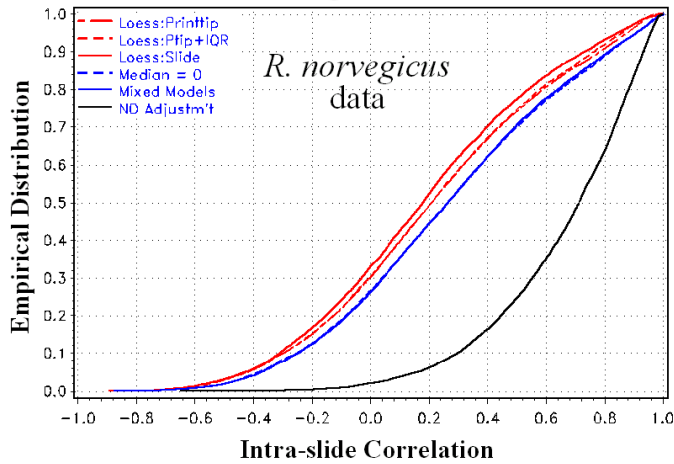


Figure 2



In both figures, the black curves are the most right-shifted, and the red curves are the most left-shifted. This shows that, in both data sets, all the normalization procedures reduced intra-slide correlation, and that the loess adjustments reduced them more than the global adjustments. In paired-difference analysis of any two curves per data set (not shown), some correlations increased even as the majority decreased, showing that reductions were stochastic, not uniform.

In each data set, Kruskal-Wallis tests of simultaneous equality of medians across all six distributions, the five distributions after normalization, and the three distributions after loess adjustment showed that significant differences existed (highest p-value = 0.003), as might be expected from the large number of data points per distribution. Pairwise Kruskal-Wallis tests among the distributions after normalization showed that most of the differences were between the loess and global adjustments, that little stochastic difference was seen between the two global adjustments (p-values = 0.91 in Figure 1 and 0.67 in Figure 2), and that little stochastic difference was seen between the two print-tip loess adjustments with versus without rescaling (p-values = 0.14 in Figure 1 and 0.53 in Figure 2).

Qualitatively, the curves for the global adjustments are about mid-way between unadjusted and the loess adjustments in Figure 1, but much closer to the loess adjustments in Figure 2. This difference may reflect the difficult nature of the *C. elegans* slides compared to the relative cleanliness of the *R. norvegicus* slides.

P-value Distributions, *C. elegans* Replicate-spot Genes.

Figures 3A and 3B show how the normalizations affect the distribution of Type III p-values from the replicate spots of the *C. elegans* data, with Figure 3B showing the lowest decile of distributions in 3A.

The unadjusted data produce p-values that are strongly enriched in low values, indicating widespread but artifactual statistical significance. All the normalizations adjust away most of the artifacts, resulting in p-values that are stochastically higher than those from the unadjusted data set.

At quantiles above 0.3, the loess adjustments tend to produce p-values that are slightly lower than those from the global adjustments. However, the curves cross near the 0.2 quantile, such that, below the 0.1 quantile (the lowest decile), the loess adjustments produce stochastically higher p-values than the global adjustments. Such crossover is not apparent in the correlation distributions from the same spots.

Figure 3A: All P-values

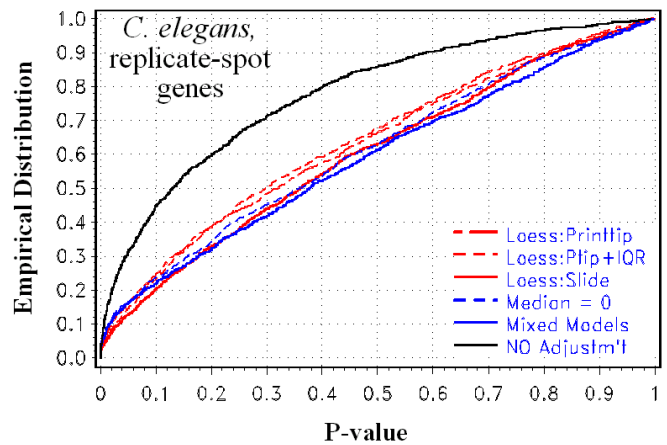
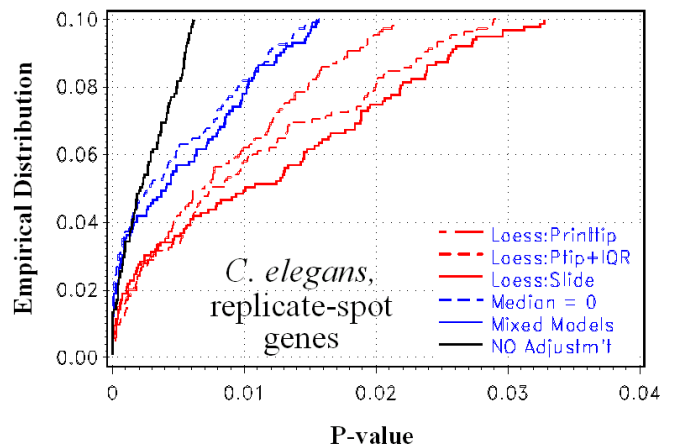


Figure 3B: Lowest decile of P-values



P-value Distributions, *C. elegans* Singleton-spot Genes.

Figures 4A and 4B show how the normalizations affect the distribution of Type III p-values from the singleton spots of the *C. elegans* data, with Figure 4B showing the lowest decile of distributions in 4A. Differences from the replicate-spot p-values are already apparent. The gap between distributions before versus after normalizations is not as drastic. The distributions from the loess adjustments are still stochastically lower than those from the global adjustments above the 0.3 quantile. However, in the lowest decile, the global adjustments tend to run in between the slide loess and print-tip loess adjustments. These differences in distribution behavior between single-spot and replicate-spot p-values suggest that it is risky to generalize from one set to both.

Figure 4A: All P-values

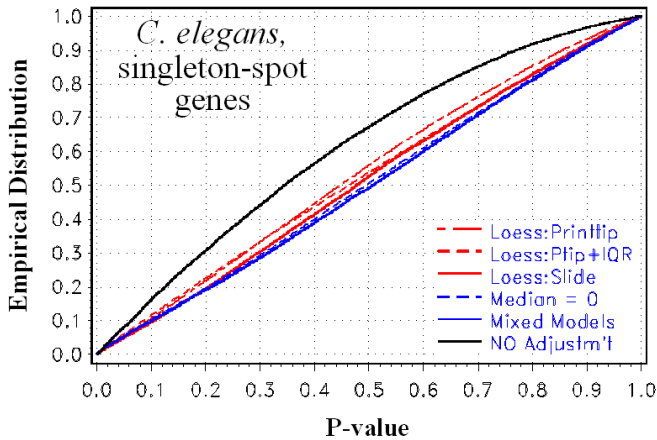


Figure 5B: Lowest decile of P-values

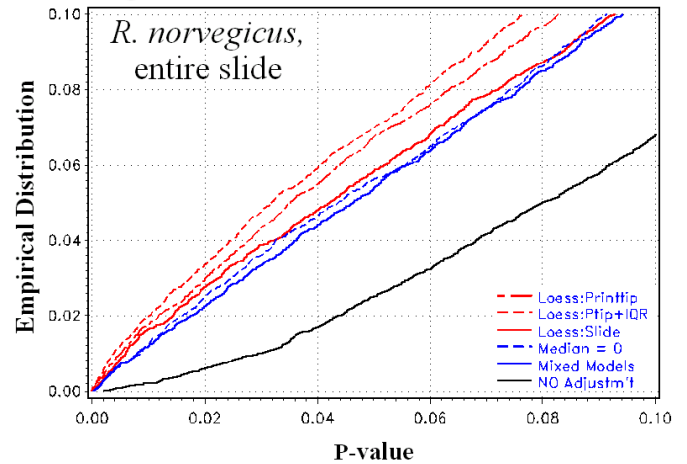
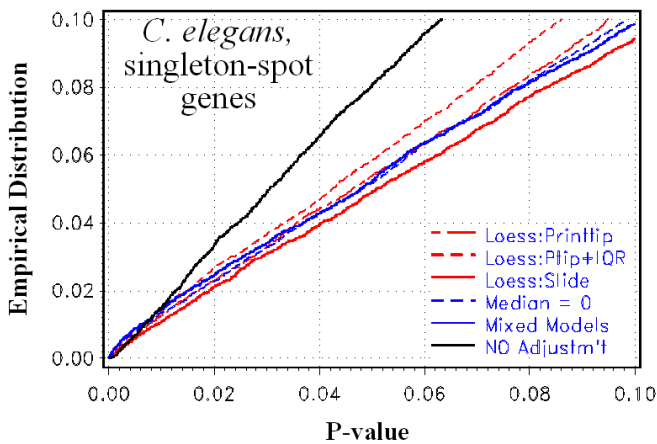


Figure 4B: Lowest decile of P-values



Here, the p-value distribution from unadjusted data show a deficit of low values, in marked contrast to what is seen in the *C. elegans* data. After adjustments, the deficit of low p-values is filled in, such that the results track the uniform distribution rather closely. As with the *C. elegans* data, p-values above the 0.3 quantile were stochastically lower when from the loess adjustments than when from the global adjustments. But unlike the *C. elegans* case, the two distributions do not cross over (except when no normalization is done). Figure 5B shows that, even at the lowest quantiles, p-values from the loess adjustments are stochastically less than those from the global adjustments.

DISCUSSION:

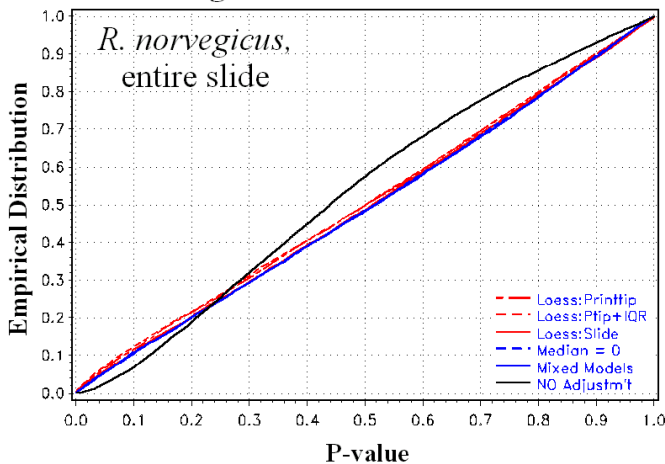
We undertook this avenue of research because we did not trust local-regression methods of normalizing microarray data. Although many researchers rely on loess and/or lowess normalizations to straighten out their curvilinear data clouds, we suspected that such chirpraxis is too aggressive on the data, and makes the data's connection to reality rather tenuous. Since many microarray slides include replicated spots for some or many probes, we investigated the possibility of using intra-slide correlation to track how different normalizations affect the data.

However, our most important finding was that intra-slide correlation did not pan out as a reliable proxy for treatment significance. This negative finding had two aspects to it. First, when we look at Figures 1 and 2, we see that the stochastically highest correlations are obtained from doing no normalization. In retrospect, we should have expected this result. The intra-slide correlation coefficient is interpretable as a standardized version of the between-slide variance component. This component captures the variation due to the treatment of interest, but also captures all the unwanted variation between slides due to technical factors. It thus stands to reason that a normalization method that reduces unwanted variation would tend to reduce intra-slide correlation regardless of what it does to treatment variation.

P-value Distributions, *R. norvegicus* Genes.

Figures 5A and 5B show how the normalizations affect the distribution of Type III p-values from the spots of the *R. norvegicus* data, with Figure 5B showing the lowest decile of distributions in 5A.

Figure 5A: All P-values



Second, we found remarkably little agreement between the way intra-slide correlations changed with normalization and the way that p-values changed with normalization. Whereas the different normalizations affected intra-slide correlations in the same way in both data sets, they affected treatment p-values differently in the two. Loess adjustments produced stochastically lower intra-slide correlations than global adjustments in both data sets, with no crossing of the empirical distribution curves. And for quantiles above 0.3, loess adjustments produced stochastically lower p-values than global adjustments in both data sets. However, at quantiles below 0.3, the empirical distribution curves from normalizations cross for the *C. elegans* data and not for the *R. norvegicus* data. This lack of concordant behavior between the two measures argues against being able to use intra-slide correlation to track how different normalizations affect treatment significance.

An additional finding of interest was the considerable proportion of intra-slide correlation coefficients that assumed negative values. One way that negative correlation could happen is if one duplicate spot tends to be up when its partner is down on some slides, and vice versa on other slides, but it is difficult to visualize either a biological or a technical mechanism that could bring this about between supposedly replicated spots containing the same species of probe cDNA. Negative correlation has been seen by other researchers, however, and at least one group (at the National Center for Toxicological Research in Jefferson, AR) has poster presentation results showing that negative correlations occur primarily among low-intensity spots. We currently believe that the phenomenon of negative correlation is closely connected to the phenomenon of "flipping", in which some spots do not change color when slide polarity changes.

Our method of testing for treatment significance was identical to Stage Two of the two-stage mixed-models approach of Wolfinger et al (2001). This approach looks on a gene-by-gene basis only at the information for that gene, and is based, moreover, on a parametric model. Essentially, it is just a sophisticated version of the t-test. Rank-based methods for testing significance also exist, some of which use the gene's rank relative to all the others on the array (see Breitling *et al.*, 2004, for a recent example). We did not look at rank-based methods with and without normalization, but it could be hypothesized, whatever, the form of normalization and test, that the value of the normalization is dependent on the type of noise present and its relationship to the test used.

Unfortunately, this is never known with anything near certainty...unless, perhaps, one performs a "spike-in" experiment using the introduction of controlled types and amounts of xenogenous transcript. While we were writing up our results, we became aware of a recent publication by Qin and Kerr (2004) that used spike-ins to assess the value of background subtraction and loess normalization, and the robustness of the t-test relative to other testing methods such as SAM. They concluded that background subtraction was harmful, that loess normalization was slightly better than median subtraction, and that the t-test was far less robust than

other testing methods. Interestingly, their Figures 2 and 3 seem to show, when background subtraction is not employed, that loess normalization affects only the t-test. However, their spike-ins were at levels that, while not unphysiological, were extreme enough to get ranked ahead of most (and often all) the endogenous genes. This is both a virtue and a vice.

In conclusion, we observe that loess normalization appeared to help treatment significance in the relatively clean *R. norvegicus* slides, but appeared to hurt it in the much more challenging *C. elegans* slides. Based on this admittedly small sample size of N=1 per group, we wish to speculate that, as far as enhancing treatment significance goes, loess normalization will tend to do the most good in precisely those situations where its aggressiveness is needed the least. We continue to suspect that, in situations demanding an aggressive normalization, the price to be paid for success will be paid in the coin of significance.

ACKNOWLEDGEMENTS:

We would like to thank Rich Kennedy, Charlotte Peterson, and the UAMS Microarray Core Facility for access to the *R. norvegicus* slide data and for the constructive feedback on findings during the development of this research.

REFERENCES:

- Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* 2004 Aug 27;573(1-3):83-92.
- Donner A. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* 54:67-82.
- Hurvich CM, Simonoff JS, Tsai CL. Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. *J Roy Stat Soc B* 1998, 60:271-293.
- Qin LX, Kerr KF; Contributing Members of the Toxicogenomics Research Consortium. Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res.* 2004 Oct 12;32(18):5471-9.
- Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol.* 2001;8(6):625-37.
- Yang YH, Dudoit S, Luu P, and Speed TP. Normalization for cDNA Microarray Data. *SPIE BiOS* 2001, San Jose, California, January 2001. Available as Tech Report #589 from <http://www.stat.berkeley.edu/tech-reports/index.html>.