Adaptive Gene Picking with Microarray Data: Detecting Important Low Abundance Signals

Yi Lin[1], Samuel T. Nadler[2], Hong Lan[2],
Alan D. Attie[2] and Brian S. Yandell[1,3]

[1] Statistics Department, University of Wisconsin–Madison
[2] Biochemistry Department, University of Wisconsin–Madison
[3] Horticulture Department, University of Wisconsin–Madison

July 1, 2002

ABSTRACT DNA microarrays to evaluate gene expression present tremendous opportunities for understanding complex biological processes. However, important genes, such as transcription factors and receptors, are expressed at low levels, potentially leading to negative values after adjusting for background. These low-abundance transcripts have previously been ignored or handled in an ad hoc way. We describe a method that analyzes genes with low expression using normal scores, and robustly adapts to changing variability across average expression levels. This approach can be the basis for clustering and other exploratory methods. Our algorithm also assigns p-values that are sensitive to changes in variability with gene expression. Together, these two features expand the repertoire of genes that can be analyzed with DNA arrays.

0.1 Introduction

Microarray technology to measure gene expression is now widespread. The application of microarray analysis to such diverse biological processes as aging (Lee et al., 1999), cancer (Golub et al., 1999; Perou et al., 1999), diabetes (Nadler et al., 2000), and obesity (Nadler et al., 2000; Soukas et al., 2000) have provided important insights. The power of microarrays to simultaneously evaluate the level of expression of thousands of genes creates the challenge of identifying those few genes that demonstrate significant changes in expression from among numerous genes that show little or no change.

Several approaches have been proposed to interpret microarray data. Clustering methods (Eisen et al., 1998; Tamayo et al., 1999) to search for genes showing similar changes in expression across experimental conditions require extensive pre-filtering to eliminate genes with low intensity or modest fold changes. Furthermore, it has become apparent that at different gene expression levels, different thresholds for significant changes are needed (Roberts et al., 2000; Wittes and Friedman, 1999; Hughes et al., 2000). More recent methods model the variability across average expression levels to establish thresholds, but still rely on ad-hoc methods for genes expressed at very low abundance (Newton et al., 2001).

We present a robust statistical approach to pick genes showing significant differential expression across abundance levels from microarray experiments. The application of this method to mouse experiments studying diabetes and obesity uncovered changes in gene expression at low abundance that were missed by other methods. Details of the method are provided, including information on how to obtain public domain software.

0.2 Methods

Our gene array analysis algorithm uses rank order to normalize data for each experimental condition, and estimates the variability at each level of gene expression to set varying significance thresholds for differential expression across levels of mRNA abundance. This procedure can be used to prefilter data in detecting patterns of differential gene expression, for instance using clustering methods. We propose assigning Bonferroni-corrected p-values, which requires only minimal assumptions. While expression data may be acquired from a variety of technologies, we focus attention on the oligonucleotide arrays, in Affymetrix chips used in a mouse experiment on diabetes and obesity.

Our approach was motivated by a series of experiments on diabetes and obesity. Nadler et al., 2000) used Affymetrix MGU74AV2 chips with over 13,000 probes representing about 12,000 genes on mRNA from adipose tissue to examine the relationship between obesity and mouse genotype (B6, BTBR or F1). Further experiments have grown out of this collaboration using replicates and will be reported elsewhere. The primary goal was to find patterns of differential gene expression in mouse tissue between strains. Thus we have a two-factor experiment with possible replication for each chip mRNA.

0.2.1 Background subtraction

Raw microarray measurements are typically normalized to account for systematic bias and noise to attempt to restore expression levels from raw data (Lockhart et al., 1996). One important source of bias is background fluorescence. Other factors that require attention include variations in array, dye, thickness of sample, and measurement noise. Background fluorescence may be measured in several ways, depending on chip technology, and is typically removed by subtraction (*cf.* Chapters 4 and 5 and Lockhart et al., 1996; Li and Wong, 2001; Schadt et al., 2001). Affymetrix chips handle background by comparing perfect match (PM) with mismatch (MM) intensity. We use weighted averages PM and MM across oligo probe pairs using recent "low-level" analysis (Chapter 5, Li and Wong, 2001; Schadt et al., 2001) to reduce measurement variability.

The following model motivates the simple subtraction for background, although it is not required for our methodology; see Hughes et al., 2000) and Li and Wong, 2001) for other approaches. The background intensity b for a gene may be attenuated at some level $\alpha \leq 1$ that could depend on the array. The gene signal g may be affected in a relative way by the degree of hybridization h, blurred by intrinsic noise ϵ (with variance depending on g) and the abundance (β) of material administered to an array. Measurements for gene j are further subject to reading error ω:

$$\begin{aligned} MM_j &= \alpha b_j + \omega_{Mj} \\ PM_j &= \alpha[b_j + \beta \exp(g_j + h_j + \epsilon_j)] + \omega_{Pj} \end{aligned}$$

Notice that the gene signal g is confounded with h unless hybridization efficiency is independent of the experimental condition. Subtracting the background intensity MM from PM yields the adjusted measurements

$$\Delta_j = PM_j - MM_j = \alpha\beta \exp(g_j + h_j + \epsilon_j) + \delta_j = \alpha\beta G_j + \delta_j$$

where the measurement error $\delta = \omega_P - \omega_M$ is symmetric around 0 and $\log(G) = g + h + \epsilon$ is the log expression level. Thus G would be observed if there were no measurement error and no array attenuation. Hence it is natural under this model to consider the log transformation $x_j = \log(\Delta_j)$. This model forms the basis for simulations presented later.

0.2.2 Transformation to approximate normality

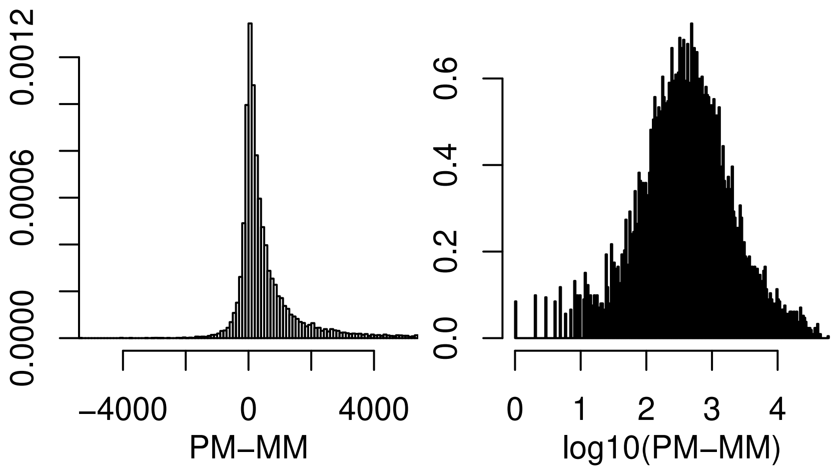

FIGURE 1. Log Transform. Expression data from one chip: (a) relative histogram of raw values $\Delta = PM - MM$, deleting 5% of values beyond ± 5000 for display; (b) relative histogram of $\log(\Delta)$ with 23% of values being negative, hence dropped.

Background-adjusted intensities are typically log-transformed to reduce the dynamic range and achieve normality. Various authors have noted that comparisons based on such log-transformed gene expression levels appear to be approximately normal (*cf.* Kerr and Churchill, 2001). However, negative adjusted values can arise from low expression levels swamped by background noise (Figure 1). Some authors have proposed adding a small value before taking log to recover some of these data (Kerr and Churchill, 2001).

Our alternative normalization method leverages this idea while providing comparisons that are more robust to difficulties with the log-normal assumption. For further discussion on normalization, see Chapters 3 and 9.

Our procedure converts the background-adjusted expression values into normal scores without discarding negative values. This normal scores transformation has been employed for microarray data using a different approach (Efron et al., 2001). If expression data are really log normal, then this normal scores transformation is indistinguishable from a log transformation after rescaling. We have found that log-transformed data (Figure 1b) appear roughly normal in the middle of the distribution, while the normal scores (Figure 2) are normal throughout.

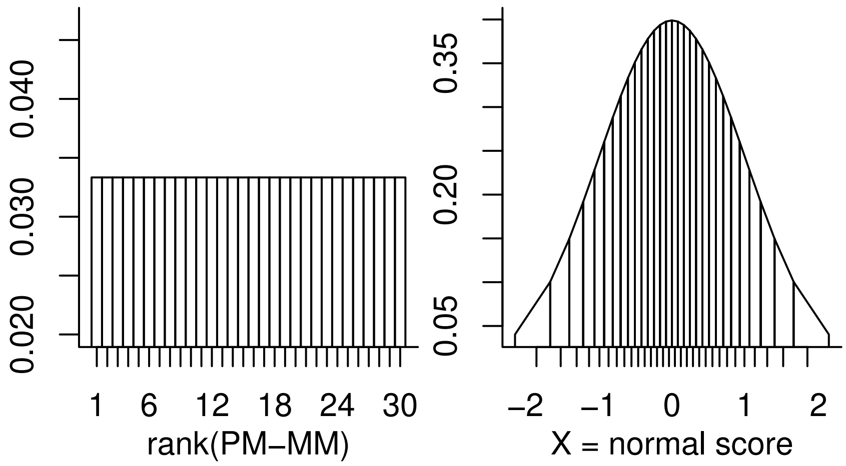

FIGURE 2. Normal Scores Transform. Sample of 30 from one chip for illustration: (a) relative histogram of ranks with equal mass; (b) normal scores transformation with equal probability mass.

Our procedure depends on the existence of some unknown monotone transformation of the data to near multivariate normal. There is always such a transformation in one dimension: let F be the cumulative distribution of adjusted values Δ and Φ be the cumulative normal distribution. Then $\Phi^{-1}(F(\Delta))$ transforms Δ to normal. If F is log-normal, then $\Phi^{-1}(F(\Delta)) = \log(\Delta)$, but we prefer not to make this assumption up front. Instead we approximate the transformation by $\Phi^{-1}(F_J(\Delta))$, where F_J is the empirical distribution of the J adjusted values $\Delta_1, \cdots, \Delta_J$. The difference between this approximate transformation and the ideal one is small, on the order of $1/\sqrt{J}$. This is known as the normal scores transformation, and is readily computed as

$$x = \Phi^{-1}(F_J(\Delta)) = \texttt{qnorm}(\texttt{rank}(\Delta)/(J+1))$$

where $\texttt{rank}(\Delta)$ is the rank order of adjusted gene measurements $\Delta = PM-$

MM among all J genes under the same condition. The normal quantiles, `qnorm`(), transform the ranks to be essentially a sample from standard normal: a histogram of these x is bell-shaped and centered about zero (Figure 2b), with normal scores equally spaced in terms of probability mass (Figure 2a). Thus these normal scores are close to a transformation that would make the data appear normal (Efron et al., 2001). If done separately by condition, this normalization automatically standardizes the center to 0 and the scale (standard deviation) to 1. Alternatively, if the experimental conditions are viewed as a random sample of a broader set of possible conditions, data across all conditions could be transformed together by normal scores. Normal scores are unaffected by monotone transformations of adjusted intensities or by global factors such as array, dye, and thickness of chip sample. Ranks may be disturbed by local noise, but that effect is unavoidable in any analysis of such an experiment.

0.2.3 Differential expression across conditions

Differential expression across conditions of interest can be computed by comparing their transformed expression levels. Information on comparison of two conditions, 1 and 2, is summarized in pairs of normal scores, x_1 and x_2, across the genes; plotting x_1 against x_2 yields points dispersing from the diagonal. However, differential gene expression between experimental conditions may depend on the average level of gene expression, with genes at different average expression having intrinsically different variability. Thus we recommend plotting the average intensity $a = (x_1 + x_2)/2$ against the difference $d = x_1 - x_2$, which involves just a 45 degree rotation similar to (see Chapters 3, 4, 7, 9 and 14, and Roberts et al., 2000). Since our normal scores may be considered a forgiving approximation to the log transform, we prefer to represent the plotting axes as if the data were log transformed–that is use an anti-log or exp scale. Thus the a axis is centered on 1 and suggests fold change in intensity, while the d axis suggests fold change in differential expression.

This method can be extended to experiments with multiple conditions, multiple readings (e.g. dyes) per gene on a chip, and replication of chips (Chapter 14 and Kerr et al., 2001). Consider an anova model

$$x_{ijk} = \mu + c_i + g_j + (cg)_{ij} + \epsilon_{ijk}$$

with $i = 1, \cdots, I$ conditions, $j = 1, \cdots, J$ genes, $k = 1, \cdots, K$ replicate chips per condition, and $\epsilon_{ijk} \sim \Phi(0, \sigma_j^2)$ being the measurement error for the kth replicate, and $c_i = 0$ if there is separate normalization by condition. Both the gene effect g_j and the condition by gene interaction $(cg)_{ij}$ are random effects. In general all variance components may depend on the gene effect g_j. Adding multiple readings per chip introduces a nested structure to the experimental design that we do not develop further here (*cf.* Lee et al., 2000).

The major biological research focus is on differential gene expression, the condition i by gene j interaction. We assume that most genes show no differential expression; thus with some small probability π_1 a particular interaction $(cg)_{ij}$ is non-zero, say from $\Phi(0, \delta_j^2)$. Let $z_j = 1$ indicate differential expression, $\text{Prob}\{z_j = 1\} = \pi_1$. The variance of the expression score is

$$\begin{array}{rcll} \text{Var}(x_{ijk}) & = & \gamma_j^2 + \delta_j^2 + \sigma_j^2 & \text{if } z_j = 1 \text{ (differential expression)} \\ \text{Var}(x_{ijk}) & = & \gamma_j^2 + \sigma_j^2 & \text{if } z_j = 0 \text{ (no differential expression)} \end{array}$$

for $i = 1, ..., I, k = 1, ..., K$, with γ_j^2 the variance for the gene j random effect. This differential expression indicator has been effectively used for microarray analysis (Newton et al., 2001; Kerr et al., 2001; Lee et al., 2000). This anova framework allows isolation of the $(cg)_{ij}$ differential expression from the g_j gene effect by contrasting conditions. Suppose w_i are condition contrasts such that $\sum_i w_i = 0$ and $\sum_i w_i^2 = 1$. The standardized contrast $d_j = (\bar{x}_{1j\cdot} - \bar{x}_{2j\cdot})\sqrt{K/2}$ with $\bar{x}_{ij\cdot} = \sum_k x_{ijk}/K$ compares condition 1 with condition 2. More generally the contrast

$$d_{jk} = \sum_i w_i \bar{x}_{ij\cdot}\sqrt{K} = \sum_i w_i \sqrt{K}[c_i + (cg)_{ij} + \bar{\epsilon}_{ij\cdot}]$$

with $\bar{\epsilon}_{ij\cdot} = \sum_k \epsilon_{ijk}/K$ has $E(d_j) = \sum_i w_i c_i \sqrt{K}$ and

$$\begin{array}{rcll} \text{Var}(d_j) & = & \delta_j^2 + \sigma_j^2 & \text{if } z_j = 1 \text{ (differential expression)} \\ \text{Var}(d_j) & = & \sigma_j^2 & \text{if } z_j = 0 \text{ (no differential expression)}\ . \end{array}$$

Again, condition effects c_i drop out and $E(d_j) = 0$ if each chip is standardized separately, but in general they remain part of the contrast.

While microarray experiments began by contrasting two conditions, this approach adapts naturally to contrasts capturing key features of differential gene expression across design factors. Time or other progressions over multiple levels, such as a linear series of glucose concentrations, might be examined for linear or quadratic trends using orthogonal contrasts (Lentner and Bishop, 1993). For instance, with five conditions the linear and quadratic contrasts are, respectively, (dropping subscripts except for condition)

$$\begin{array}{rcl} d_{\text{linear}} & = & (2x_5 + x_4 - x_2 - 2x_1)\sqrt{K/8} \\ d_{\text{quadratic}} & = & (2x_5 - x_4 - 2x_3 - x_2 + 2x_1)\sqrt{K/14}\ . \end{array}$$

With conditions resolved as multiple factors, such as obesity and genotype in our situation, separate contrasts can be considered for main effects and interactions. Each contrast can be analyzed in a similar fashion to the above. Alternatively, one can examine factors with multiple levels, say three genotypes, by an appropriate anova evaluation (Lee et al., 2000).

0.2.4 Robust center and spread

For the majority of genes that are not changing, the difference d_j reflects only the intrinsic noise. Thus genes that do change can be detected by assessing their differential expression relative to the intrinsic noise found in the non-changing genes. While it is natural to use replicates when possible to assess the significance of contrasts for each gene, microarray experiments have typically had few replicates K, leading to unreliable tests. Some authors have considered shrinkage approaches that combine variance information across genes (Efron et al., 2001; Lönnstedt and Speed, 2001).

Measurement error seems to depend on the gene expression level $a_j = \sum_{ik} x_{ijk}/IK$, and it may be more efficient to combine variance estimates across genes with similar average expression levels (Roberts et al., 2000; Hughes et al., 2000; Newton et al., 2001; Kerr et al., 2001; Baldi and Long, 2001; Long et al., 2001). Further, if there were no replicates, as in early microarray data, then it would be important to combine across genes in some fashion. There may in addition be systematic biases that depend on the average expression level (Dudoit et al., 2000; Yang et al., 2001). We noticed that empirically the variance across non-changing genes seems to depend approximately on expression level in some smooth way, decreasing as a increases due in part to the mechanics of hybridization and reading spot measurements. Here we consider smooth estimates of abundance-based variance to account for these concerns. In a later paper we will investigate shrinking the gene-specific variance estimate using our abundance-based estimate and an empirical Bayes argument similar to (Lönnstedt and Speed, 2001).

Our approach involves estimating the center and spread of differential expression as it varies across average gene expression a_j to standardize the differential expression. Specifically we use smoothed medians and smoothed median absolute deviations, respectively to estimate the center and spread. Smoothing splines (Wahba, 1990) are combined with standardized local median absolute deviation (MAD) to provide a data-adapted, robust estimate of spread $s(a)$. A smooth, robust estimate of center $m(a)$ can be computed in a similar fashion by smoothing the medians across the slices. We use these robust estimates of center and scale to construct standardized values

$$T_j = (d_j - m(a_j))/s(a_j)$$

and base further analysis on these standardized differences.

For convenience, we illustrate with two conditions and drop explicit reference to gene j. Revisiting the motivating model helps explain our specification for spread. Consider again $\log(G) = g+h+\epsilon$ and suppose hybridization error is negligible, or at least the same across conditions. The intrinsic noise ϵ may depend on the true expression level g: for two conditions 1 and 2, the difference d is approximately

$$d \approx \log(G_1) - \log(G_2) = g_1 - g_2 + \epsilon_1 - \epsilon_2 \ .$$

If there is no differential expression, $g_1 = g_2 = g$, then $\mathrm{Var}(d|g) = s^2(g)$, and the gene signal g may be approximated by a. However, the true formula for $\mathrm{Var}(d|a)$ is not exactly $s^2(a)$, and cannot be determined without further assumptions.

Thus differential contrasts standardized by estimated center and spread that depend on a should have approximately the standard normal distribution for genes that have no differential expression across the experimental conditions. Comparison of gene expression between two conditions involves finding genes with strong differential expression. Typically, most genes show no real difference, only chance measurement variation. Therefore a robust method that ignores genes showing large differential expression should capture the properties of the vast majority of unchanging genes.

The genes are sorted and partitioned based on a into many (say 400) slices containing roughly the same number of genes and summarized by the median and the MAD for each slice. That is, with 12,000 genes, the 30 contrasts d for each slice are sorted; the average of ordered values 15 and 16 is the median, while the MAD is the median of absolute deviations from that central value. These 400 medians and MADs should have roughly the same distribution up to a constant. To estimate the scale, it is natural to regress the 400 values of log(MAD) on a with smoothing splines (Wahba, 1990), but other non-parametric smoothing methods would work as well. The smoothing parameter is tuned automatically by generalized cross validation (Wahba, 1990). The anti-log of the smoothed curve, globally rescaled, provides an estimate of $s(a)$, which can be forced to be decreasing if appropriate. The 400 medians are smoothed via regression on a to estimate $m(a)$.

Replicates are averaged over in the robust smoothing approach. That is, contrasts $d_j = \sum w_i \bar{x}_{ij\cdot}\sqrt{K}$ factor out replicates. We are currently investigating shrinkage variance estimates of the form

$$s_j^2 = \frac{\nu_0 s^2(a_j) + \nu_1 \hat{\sigma}_j^2}{\nu_0 + \nu_1}$$

with $\hat{\sigma}_j^2 = \sum_k (x_{ijk} - \bar{x}_{ij\cdot})^2/\nu_1$, $\nu_1 = I(K-1)$ and ν_0 is the empirical-Bayes estimate (*cf.* 20) of the degrees of freedom for $\hat{\sigma}_j^2/s^2(a_j)$.

It should be possible to combine estimates of spread across multiple contrasts, say by using the absolute deviations $|x_{ijk} - a_j|$ for all genes with average intensity a_j within the range of a particular slice to estimate the slice MAD. This is sensible since these absolute deviations estimate the measurement error for most genes and most conditions. Those few genes with large differential effects across conditions would have large absolute deviations that are effectively ignored by using the robust median absolute deviation.

It may be reasonable in some cases to use 'house-keeping genes' that are generally believed to not change over different conditions (*cf.* Baldi and Long, 2001; Chen et al., 1997). However, this may not capture the finer details of the center and scale as average intensity changes over the microarray. We use a robust estimation procedure to guard against the influence of the small proportion of changing genes that "contaminate" microarray data when we estimate the intrinsic noise level. Notice, however, that this contamination is of primary interest in a similar fashion to the problem of outlier detection.

0.2.5 Formal evaluation of significant differential expression

Formal evaluation of differential expression may be approached as a collection of tests for each gene of the "null hypothesis" of no difference, or alternatively as estimating the probability that a gene shows differential expression (Newton et al., 2001; Kerr et al., 2001). Testing raises the need to account for multiple comparisons, here we use p-values derived using a Bonferroni-style genome-wide correction (Dudoit et al., 2000). Genes with significant differential expression are reported in order of increasing p-value. Further details of this procedure and the software can be found below.

We can use the standardized differences T to rank the genes. The conditional distribution of these T given a is assumed to be standard normal across all genes whose expressions do not change between conditions. Hypothesis testing here amounts to comparing the standardized differences with the intrinsic noise level. Since we are conducting multiple tests, we should adjust the test level of each gene to have a suitable overall level of significance. We prefer the conservative Zidak version of the Bonferroni correction: the overall p-value is bounded by $1 - (1 - p)^J$, where p is the single test p-value. For example, for 13,000 genes with an overall level of significance of 0.05, each gene should be tested at level $1.95 * 10^{-6}$, which corresponds to 4.62 score units. Testing for a million genes would correspond to identifying significant differential expression at more than 5.45 score units. Guarding against overall type one error may seem conservative. However, a larger overall level does not substantially change the normal critical value (from 4.62 to 4.31 with 13,000 genes for a .05 to .20 change in p-value). This test can be made one-sided if preferred. We report gene-by-gene results in the obesity data analysis below in terms of the overall significance level rather than the single test level in order to avoid confusion.

Apparently less conservative multiple comparison adjustment to p-values are proposed in (Yang et al., 2001). However, the results are essentially the same with all such methods, except when more than 5–10% of the genes show differential expression across conditions. Figure 3 shows p-values from a typical mouse experiment with about J =13,000 genes. The Bonferroni (p/J) and Zidak $(1 - (1 - p)^J)$ significance levels are virtually identical, 3.64 and 3.74 $\times 10^{-6}$, respectively. Further, the Holms methods discussed in Dudoit et al., 2000), which adjusts for the number of genes remaining to

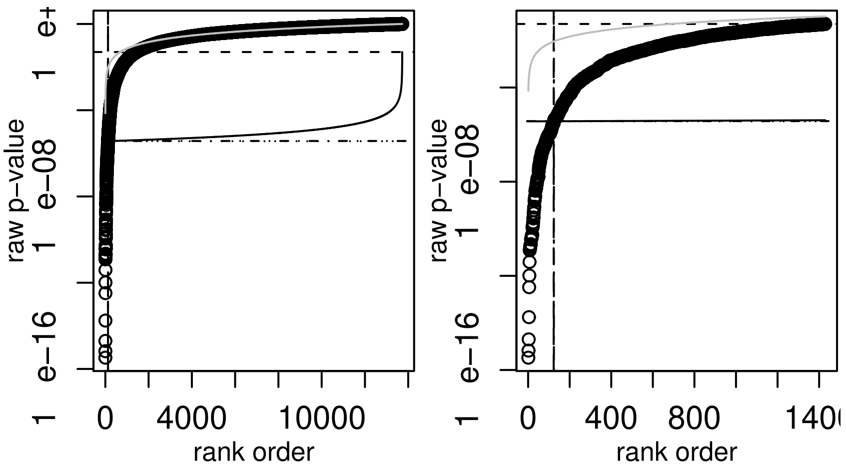

FIGURE 3. Multiple Comparisons Criteria. Typicial mouse chip evaluation with roughly 13,000 genes. Circles are raw p-values on semi-log plot in rank order. Grey curved line at top for uniform p-value ideal under null hypothesis. Dotted line at nominal 5% level. Dashed lines at about 3.7×10^{-6} where Bonferroni, Zidak and Holms methods meet. Curved solid line is Holms. (a) all p-values; (b) top 5% of p-values.

be tested, agree with these two for the most significant 5–10% of the genes, picking the top 124 while Bonferroni picks only 123. The Westfall-Young method recommended by (Dudoit et al., 2000) is not shown but should have similar properties to Holms. Thus we are quite comfortable using the Zidak method in situations where only a small portion of the genes show differential expression.

It may be appropriate to examine a histogram of standardized differences T, using these critical values as guidelines rather than strict rules. The density f of all the scores is a mixture of the densities for non-changing f_0 and changing f_1 genes,

$$f(T) = (1 - \pi_1) f_0(T) + \pi_1 f_1(T) \ .$$

By our construction, f_0 is approximately standard normal. Following Efron et al., 2001), set π_1 just large enough so that the estimate

$$f_1(T) = [f(T) - (1 - \pi_1) f_0(T)]/\pi_1$$

is positive. This in some sense provides a 'liberal' estimate of the distribution of differentially expressed genes. It lends support to examination of a wider set of genes, with standardized scores that are above 3 or below –3. We suggest using this set as the basis for hierarchical clustering. Notice also that this provides an estimate of the posterior probability of differential expression ($z_j = 1$) for each mRNA,

$$\text{Prob}\{z_j = 1 | T_j\} = \pi_1 f_1(T_j)/f(T_j).$$

Gross errors on microarrays can be confused with changing genes. Replicates can be used to detect outliers in a similar fashion to the approach for differential gene expression. Residual deviations of each replicate from the condition*gene mean, $x_{ijk} - \bar{x}_{ij\cdot}$, could be plotted against the average intensity, a_j. Robust estimates of center and scale could be used as above in formal Bonferroni-style tests for outliers. Separate smooth robust estimates of center and scale are needed for each contrast. Perhaps an additional Bonferroni correction may be used to adjust for multiple contrasts.

0.2.6 Simulation Studies

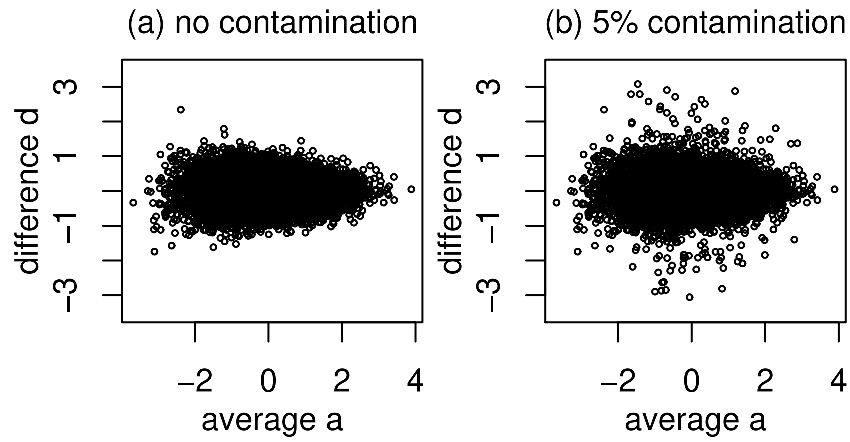

FIGURE 4. Simulation of expression data. Scatter plots of difference d against average intensity a (a) before and (b) after adding 5% contamination.

Three simulation studies were conducted to examine properties of our procedure. The first study shows how well the smoothed median absolute deviations can estimate the variability among unchanging genes. The second study verifies that the normal scores procedure can essentially extract the "true differential expression" that would be observed if there were no measurement error. Simulated data from the second study was used to compare our procedure with other procedures that have been previously proposed.

The following simulation demonstrates the effectiveness of the robust standardization. We generated 9,500 (a, d) pairs, with a from standard normal and d normally distributed with mean 0 and standard deviation $\sigma(a) = 1/[a/3 + 2.5]$. Then we generated another 500 pairs by adding independent standard normal random numbers to each d value. Thus given the same X, the standard deviation of the contaminated d is $[1 + (a/3 + 2.5)^2]^{1/2}$ times that of the uncontaminated d (1.8 to 3.64 as a goes from

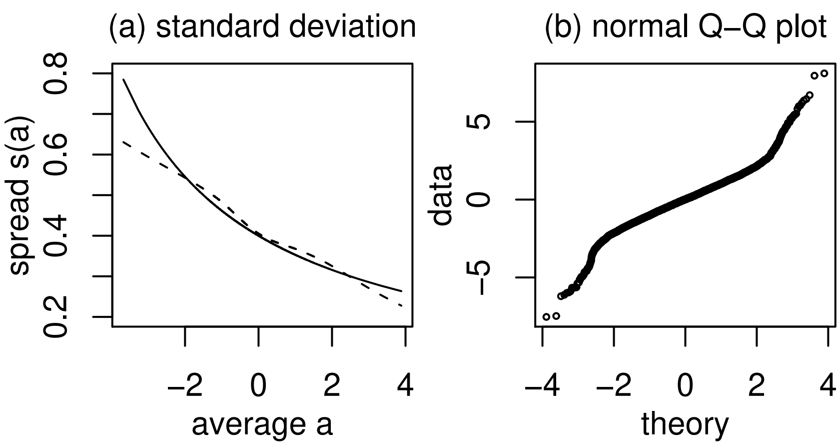

FIGURE 5. Simulation study of spread. Simultated data has 5% contamination shown in Figure 4b. (a) True (solid line) and estimated (dotted line) spread $s(a)$; (b) Q-Q plot reveals "contamination" by differentially expressed genes.

–3 to 3). We applied our robust scaling function to the combined data of 10,000 pairs. Figure 4 show scatter plots of the simulated data before and after the addition of contamination. Figure 5a shows how close are the true (solid line) and estimated scale (dotted line) scale. While there is always some bias with non-parametric estimation, the key bias problem arises in estimating spread in the presence of differentially expressed genes. The robust procedure reduces the influence of this contamination. The normal quantile plot of Y/s(X) in Figure 5b shows the middle portion to be almost straight, as expected with normal data, while the tails diverge due to the "contamination" by differentially expressed genes.

We tested the normal scores procedure on simulated data with two conditions and constant intrinsic variance across average expression levels. We generated samples with 10,000 genes and 5% differential expression and increasing amounts of measurement error. First, we randomly generated 9,500 normal variates with mean 4 and variance 2. Next, we generate 500 random numbers from the same distribution and added normal "contamination" which was either up regulated or down regulated with probability 1/2. This contamination had variance 1/2 and mean tending from 3 to 2 as average expression level ranged from low to high abundance. The intrinsic noise ϵ was generated with variance 0.5, attenuations β were set at 1. We considered a range of measurement error variances from none to high ($\delta = 0, 1, 2, 5, 10, 20$). In the ideal situation of no measure error, the 'best' ranking would be based on the differential expression between the two conditions. We use the performance of this ideal 'best' ranking as the benchmark against which to test our procedure. Figure 6 compares the top 500 'best' ranks when there is no measurement error with the ranks

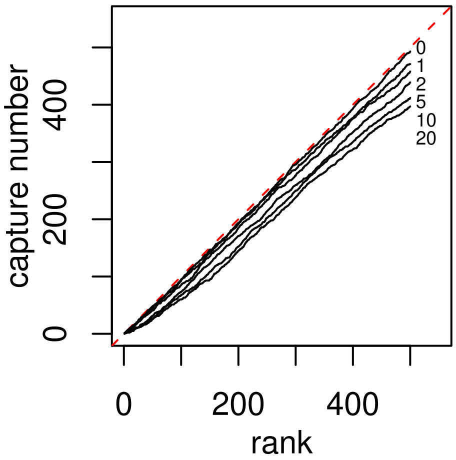

FIGURE 6. Capture efficiency. The top 500 'best' ranks when there is no measurement error with the ranks produced by our procedure under different level of measurement error ($\delta = 0, 1, 2, 5, 10, 20$).

produced by our procedure under different level of measurement error. In the absence of measurement error, our ranks essentially matches the 'best' ranking (line 0). When a typical level of noise is applied to the silmulation, the ranks produced by our procedure still comes close to the 'best' ranking.

In practice, analysis of low-abundance mRNA's leads to negative adjusted values, which are ignored or set to an arbitrary value by most other procedures. In the absence of measurement error, previously proposed methods perform well when they are first rank-ordered as done in our algorithm (Figure 8a). In practice, measurement error becomes high with genes of low abundance and therefore background correction masks changes in gene expression. Despite a high level of noise, our method successfully detected numerous differentially expressed low-abundance mRNA's (Figure 8b). None of the non-changing genes were identified in our simulations. In contrast, an early analytical method assuming a constant coefficient of variation (Chen et al., 1997) yielded conservative, flat thresholds (Figure 8b, dashed line). The Bayesian approach (Newton et al., 2001) missed the pattern of changing variation with average gene intensity and misses most of the differentially expressed genes (Figure 8b, dotted line).

0.2.7 Comparison of methods with E. coli *data*

We reexamined some of the *E. coli* data reported in Newton et al., 2001). Figure 9 shows (a) log-transformed and (b) normal scores transformed data, with decisions based on Newton et al., 2001) as red dotted lines and our method as purple solid lines. Both methods and both figures agree in choosing the three extreme genes associated with IPTG-b. However, the normal scores (b) handles these in a natural way while the log-transform (a) uses an ad hoc approach–leaving negative values out for computation and then including them at the margin for inference. Note in addition that our line based on a non-parametric estimate of variance better reflects the variability in the data, particularly at both extremes of mean intensity.

0.3 Software

The analysis procedures are written as an R language module. The R system is publicly available from the R Project, and our code is available from the corresponding author as the R `pickgene` library. The function `pickgene()` plots d against a, after backtransforming to show fold changes, and picks the genes with significant differences in expression. Examples include the simulations and graphics presented here. This library can be found at `www.stat.wisc.edu/~yandell/statgen`.

In its simplest form, `pickgene()` takes a data frame (or matrix) of microarray data, one column per array. We assume housekeeping genes have already been removed. Columns are automatically contrasted using the prevailing form of orthonormal contrast (default is polynomial, `contrasts = "contr.poly"`).

```
library( pickgene )
result <- pickgene( data )
```

This produces a scatter plot with average intensity a along the horizontal axis and contrasts d along the vertical, with one plot for each contrast (typically one fewer than the number of columns of `data`).

With two columns, we are usually interested in something analogous to the log ratio, which can be achieved by renormalizing the contrast. If desired, log transform can be specified by setting `rankbased = F`. Gene ideas can be preserved in the results as well.

```
result <- pickgene( data, geneID = probes,
                    renorm = sqrt( 2 ), rankbased = F )
print( result$pick[[1]] )
```

The `pick` object is a list with one entry for each contrast, including the `probe` names, `average` intensity a, `fold` change $(exp(d)$, as if $\Phi^{-1}(F(\Delta)) = \log(\Delta))$, and Bonferroni-adjusted p-value. The `result` also contains a `score` object with the average `intensity` a, `score` T, `lower` and `upper` Bonferroni limits, and `probe` names.

The `pickgene()` function relies on two other functions. The function `model.pickgene()` generates the contrasts, although this can be bypassed. More importantly, the function `robustscale` slices the pairs (a, d) into 400 equal-sized sets based on a, finds medians and log(MAD)s for each slice, then smooths them using splines (Wahba, 1990) to estimate the center, $m(a)$, and spread, $s(a)$, respectively.

Estimates of density are based on the `density()` function, packaged in our `pickedhist()` routine.

```
pickedhist( result, p1 = .05, bw = NULL )
```

We pick a bandwidth `bw` that provides smooth curves and then adjust π_1 = `p1` so that f_1 is positive.

The standard deviation $s(a)$ is not returned directly in `result`. However, it is easily calculated as `log( upper / lower ) / 2`.

0.4 Application

0.4.1 Diabetes and Obesity studies

Our approach was motivated by a series of experiments on diabetes and obesity (Nadler et al., 2000) using Affymetrix mouse MGU74AV2 chips with over 13,000 RNA probes representing roughly 12,000 genes or ESTs. Here obesity was controlled by the leptin gene. Each chip was treated with RNA extracted from adipose tissue combined over sets of four mice. Three chips were assigned to lean mouse sets from strains of B6, BTBR or F1 mice; the other three chips were for obese mice from the same strains. Thus we had a 3x2 factorial design with genotype and obesity as the main effects. The primary goal was to find patterns of differential gene expression in adipose tissue between obese and lean mice.

The majority of individuals with Type 2 diabetes mellitus are obese. Adipose tissue is thought to influence whole-body fuel partitioning and might do so in an aberrant fashion in obese and/or diabetic subjects. Almost half of these genes had at least one negative adjusted value in the dataset at low expression (Figure 10a, green dots), and were missed by other methods. Our earlier study using clustering methods found interesting genes at high expression levels, including many at smaller fold changes (purple dots). Clearly fold change is not the whole story.

However, this earlier analysis ignored genes at low expresssion, which we detected here. Roughly 100 genes were determined to have significant ($p < 0.05$) fold changes in gene expression using the robust normal scores procedure (blue circles). A handful of significant genes with low expression (green dots in blue circles) were evaluated using RT-PCR, with a false positive rate of about 50%, reflecting high noise in Affymetrix data at low intensity. Nevertheless true positives detected by this method corresponding to transcription factors or receptors have potential to shed important light on adipocyte signaling pathways.

We also examined genotype effects (Figure 10b), finding numerous genes that seemed to have either additive or dominance effect, but not both. Red lines in Figure 10 correspond to Bonferroni limits for one contrast; the blue circle in Figure 10b is a simultaneous limit for the two genotype contrasts. The 5% criteria is based on the Zidak adjustment for multiple comparisons, which essentially agrees with both the Bonferroni and Holms when only a small fraction of genes show differential expression. We employed a permutation check of our method to verify the size of the test. That is, we took two of the chips (B6 and BTBR), randomly permuted the data for one chip relative to the other, and applied our method on the permuted data. This procedure was repeated 100 times. Our method picked out one significant gene in four permutations, plus a fifth permutation had one gene right on the boundary; this agrees well with the expected 5% error rate.

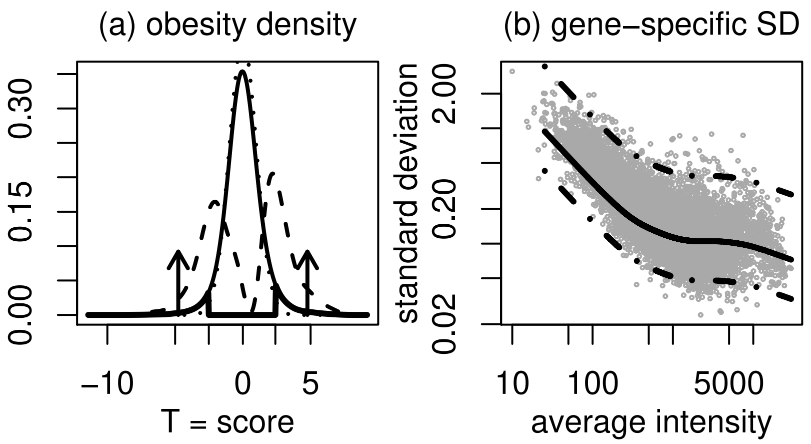

FIGURE 7. Investigation of differential density and gene-specific variance. (a) Density f of obesity T scores: all values (solid) overlaid on standard normal f_0 (dotted), with good agreement to 2 SD; dashed for f_1 differentially expressed genes; arrows at Bonferroni critical value. Thick lines near ± 2.5 show where $f = 2f_0$. (b) Gene-specific SDs for experiment with 5 replicates per two conditions: solid line is our smoothed MAD estimate of $s(a)$; dashed lines are upper and lower limits based on χ^2_8.

Figure 7a shows the density f of standardized differences T for obesity overlaid on the standard normal density f_0 , with good agreement for standardized differences between –2 and 2. The long-dashed line shows an estimate of the density f_1 for differentially expressed genes as described above, picking the proportion of changing genes just large enough to ensure the density is positive (*cf.* Efron et al., 2001). This illustrates just how conservative the Bonferroni approach is. We have since begun examining the genes with differential expression exceeding 3 SDs using hierarchical clustering after removing mean expression level. Initial results suggest important rearrangement that might imply functional association among genes in clusters.

While this early experiment had no replication, subsequent experiments have employed more chips to increase power. One study, to be reported elsewhere, had 10 mice, five from each of two strains, that were separately applied to 10 chips. Using our procedure on the mean differences, we computed the spread estimate $s(a)$ described above. Figure 7b shows $s(a)$ overlaid on gene-specific estimates of standard deviation. That is, for each gene, an SD was estimated using 8 d.f. (10 chips – two groups). The abundance-based $s(a)$ appears to capture the central tendency in spread over the range of average intensity for this experiment. Notice that the SD confidence band based on χ^2_8 variability of gene-specific SDs covers most of the SDs, suggesting that our abundance-based approximation may not be too bad. However, since this band spans an almost 10-fold difference in SD, a shrinkage-based estimator that combines gene-specific and average intensity based estimates of spread is adviseable.

In conclusion, this novel method adapts to the dynamic range of expression data while handling low intensity signals, including negative adjusted values. No data need be ignored, as the method finds a transformation to identify differentially expressed genes from large microarray data sets. Further, we have demonstrated the feasibility of putting p-values on differential gene expression without making many of the assumptions other methods require.

This method can be extended to general experimental designs (Lee et al., 2000) by adjusting for variability in expression across all conditions relative to the average gene expression. The utility of clustering (Eisen et al., 1998; Tamayo et al., 1999) and classification (Golub et al., 1999) methods can be extended by relying on the standardized normal scores rather than log-transformed values. This can uncover novel relationships, particularly involving low-abundance transcripts. The p-values proposed here can further refine relationships uncovered by these omnibus methods.

Transcriptional regulation plays a particularly important role in the biology of low-abundance mRNA transcripts. This new algorithm now extends the powerful techniques of DNA array analysis to the world of low-abundance mRNA's.

0.4.2 Software Example

Data analysis was based on a data frame from Affymetrix chip processing, which after some manipulation has six columns of Affymetrix 'Log.Avg' values plus the probe set name. These data have the 'house-keeping genes' and other Affymetrix references already removed. Suppose the data frame has the probe names in column 1 and the next six columns contain the three lean chips, followed by the three obese chips, with B6, F1 and BTBR genotypes within each obesity class. The `pickgene` routine will show the three main effects plots and return a data analysis list with the following command:

```
> library( pickgene )
> Leanob.pick <- pickgene( data[,-1], data[,1],
    faclevel = c(2,3), facnames = c("Obese","Genotype"),
    marginal = T, mfrow = c(2,2),
    renorm = c( sqrt(2)/3, sqrt(2)/2, sqrt(6)/4 ) )
```

Contrasts for the experimental design are automatically created with `faclevel` and `facnames` through a call to `model.pickgene`. That is, there are 2 levels of factor `Obese` and 3 levels of factor `Genotype`. The default normalization is for the sum of squared weights $\sum_i w_i$ to be K times the product of levels of other model factors. Often it is useful to change this using the `renorm` option.

```
> apply( model.pickgene( faclevel = c(2,3),
+    facnames = c("Obese","Genotype")),
+    2, function(x) sum( x^2 ) )
(Intercept)     Obese.L  Genotype.L  Genotype.Q
          6           3           2           2
## renormalize so contrasts have natural meaning
> model.pickgene( faclevel = c(2,3),
+    facnames = c("Obese","Genotype"),
+    renorm = c( sqrt(2)/3, sqrt(2)/2, sqrt(6)/4 ) )
  (Intercept)   Obese.L Genotype.L Genotype.Q
1           1 -0.333333       -0.5       0.25
2           1 -0.333333        0.0      -0.50
3           1 -0.333333        0.5       0.25
4           1  0.333333       -0.5       0.25
5           1  0.333333        0.0      -0.50
6           1  0.333333        0.5       0.25
attr(,"assign")
[1] 0 1 2 2
attr(,"contrasts")
attr(,"contrasts")$Genotype
[1] "contr.poly"
```

It is also possible to explictly enter a design matrix for conditions using the `model.matrix` option to `pickgene`

The overlay in Figure 4a of green dots is found by simple use of `apply` to identify genes with `any( x <= 0 )`. In order to plot these values, one has to explicitly recompute the normal scores and find the average and contrast:

```
> datanorm <- apply( ddata[,-1], 2, function( x )
+   qnorm( rank( x ) / ( 1 + length( x ))))
> datamean <- apply( datanorm, 1, mean )
```

```
> dataleanob <- apply( datanorm, 1, function( x )
+   mean( x[1:3] ) - mean( x[4:6] ) )
```

Figure 4b comes directly from the `pickgene` results. Note that all three sets of scores are ordered by the same average intensity per gene.

```
> names( Leanob.pick$score )
[1] "Obese.L"    "Genotype.L" "Genotype.Q"
> plot( Leanob.pick$score[[2]]$score,
+    Leanob.pick$score[[3]]$score )
```

The red lines come from the Zidak adjusted value,

```
> zidak <- qnorm( 1 - exp(( log( 1 - .05 ))/ nrow( data )) / 2,
+   lower.tail = F)
```

The density plot in Figure 4c was constructed with the `pickedhist` command. This can be applied to a single contrast contrast using the `show` option, or to all simultaneously. Since this is an ad hoc procedure (*cf.* Efron et al., 2001), a bit of trial and error is required. First it is advisable to increase the smoothing by raising the bandwidth (`bw`). Then adjust the prior proportion of changed genes ($\pi_1 =$ `p1`) until density estimates do not cross negative.

```
## automatic bandwidth selection with p1 = .05
pickedhist( picked )
## refined selection by hand
pickedhist( picked, bw=.5, p1 = .1145)
```

Figure 4d came from another experiment. Gene-specific variances were computed by using `apply` on rows of the data frame and then overlaying the abundance-based variance. Details are left to the reader.

Bibliography

Baldi P, Long AD (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17:509–519

Chen Y, Dougherty ER, Bittner ML (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2:364–374

Dudoit S, Yang YH, Callow MJ, Speed TP (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, Dept. Biochem., Stanford U

Efron B, Tibshirani R, Goss V, Chu G (2001). Microarrays and their use in a comparative experiment. *J. Amer. Statist. Assoc.* 96:1151–1160

Eisen MB, Spellman PT, Brown PO, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863–14868

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, a. Caligiuri M, Bloomfield CD, Lander ES (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286:531–537

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai HY, He YDD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH (2000). Functional discovery via a compendium of expression profiles. *Cell* 102:109–126

Kerr MK, Churchill GA (2001). Statistical design and the analysis of gene expression microarrays. *Genet. Res.* 77:123–128

Kerr MK, Martin M, Churchill GA (2001). Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7:819–837

Lee CK, Klopp RG, Weindruch R, Prolla TA (1999). Gene expression profile of aging and its retardation by caloric restriction. *Science* 285:1390–1394

Lee MLT, Kuo FC, Whitmore GA, Sklar J (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* 97:9834–9839

Lentner M, Bishop T (1993). *Experimental Design and Analysis.* Blacksburg, VA: Valley Book Company, 2nd edition

Li C, Wong WH (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* 98:31–36

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotech.* 14:1675–1680

Long AD, Mangalam HJ, Chan BYP, Tolleri L, Hatfield GW, Baldi P (2001). Gene expression profiling in *escherichia coli* K12: Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *J. Biol. Chem.* 276:19937–19944

Lönnstedt I, Speed T (2001). Replicated microarray data. *Statistica Sinica* 0:000–000

Nadler ST, Stoehr JP, Schueler KL, Tanimoto G, Yandell BS, Attie AD (2000). The expression of adipogenic genes is decreased in obesity and *diabetes mellitus. Proc. Natl. Acad. Sci. USA* 97:11371–11376

Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol.* 8:37–52

Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JCF, Lashkari D, Shalon D, Brown PO, Botstein D (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* 96:9212–9217

Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YDD, Dai HY, Walker WL, Hughes TR, Tyers M, Boone C, Friend SH (2000). Signaling and circuitry of multiple mapk pathways revealed by a matrix of global gene expression profiles. *Science* 287:873–880

Schadt EE, Li C, Su C, Wong WH (2001). Analyzing high-density oligonucleotide gene expression array data. *J. Cellular Biochemistry* 80:192–202

Soukas A, Cohen P, Socci ND, Friedman JM (2000). Leptin-specific patterns of gene expression in white adipose tissue. *Genes & Development* 14:963–980

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907–2912

Wahba G (1990). *Spline Models for Observational Data*. Philadelphia, PA: Soc. Indust. Appl. Math.

Wittes J, Friedman HP (1999). Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. *J. Natl. Cancer. Inst.* 91:400–401

Yang YH, Dudoit S, Luu P, Speed TP (2001). Normalization for cDNA microarray data. spie. Technical report, SPIE BiOS 2001, San Jose, California

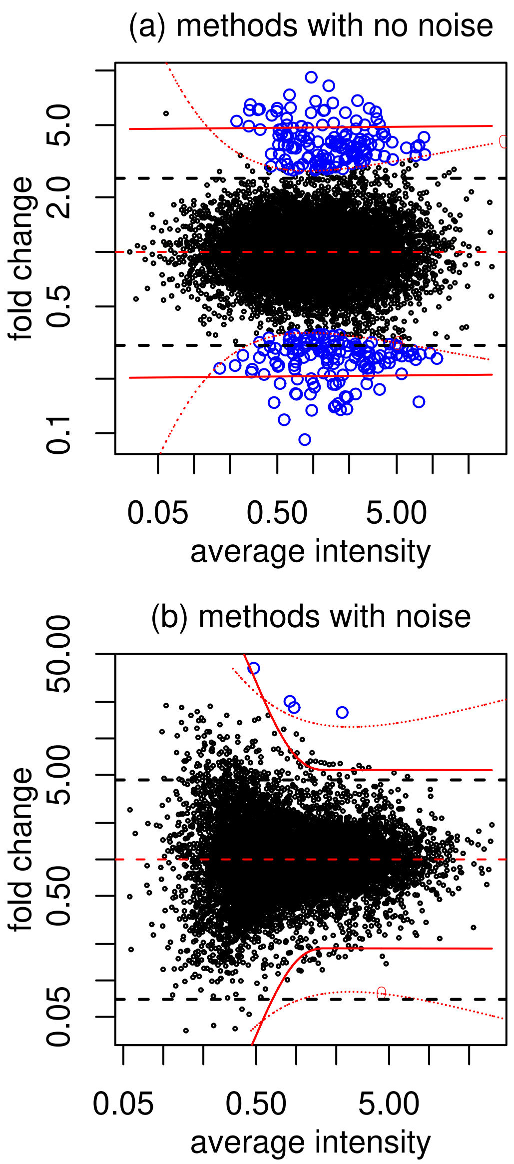

FIGURE 8. Effect of measurement error on shape of differential expression. Flat black dashed line for Chen *et al.*, (1997); curved red dotted line for Newton *et al.*, (2001) odds ratio of 1; smooth red solid line for our method. Blue circles are genes beyond odds ratio of 1. Horizontal red dashed line at 1 = no fold change. (a) no measurement error; (b) high measurement error ($\delta = 20$).

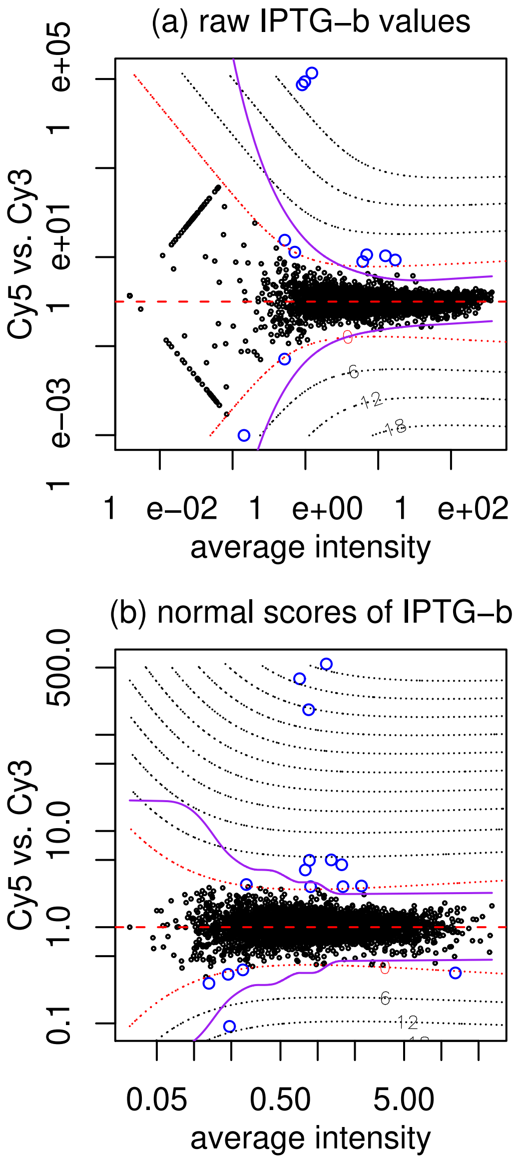

FIGURE 9. *Escheria coli* IPTG-b analysis to compare gamma-gamma Bayesian to normal scores method. Each point is a gene, with blue circles for genes with odds ratio above 1 (red dotted at odds ratio of 1, black dotted lines at 10^6 incrementes in odds; see Newton et al., 2001). Solid purple line for our Bonferroni 5% criteria. (a) log-log scale of original expression levels, with negative values plotted along diagonals at far left; (b) normal score transformed data, anti-logged and plotted on log-log scale. Note three genes with huge IPTG-b signal. Our procedure is more conservative with no noise, but detects pattern of variability when noise present.

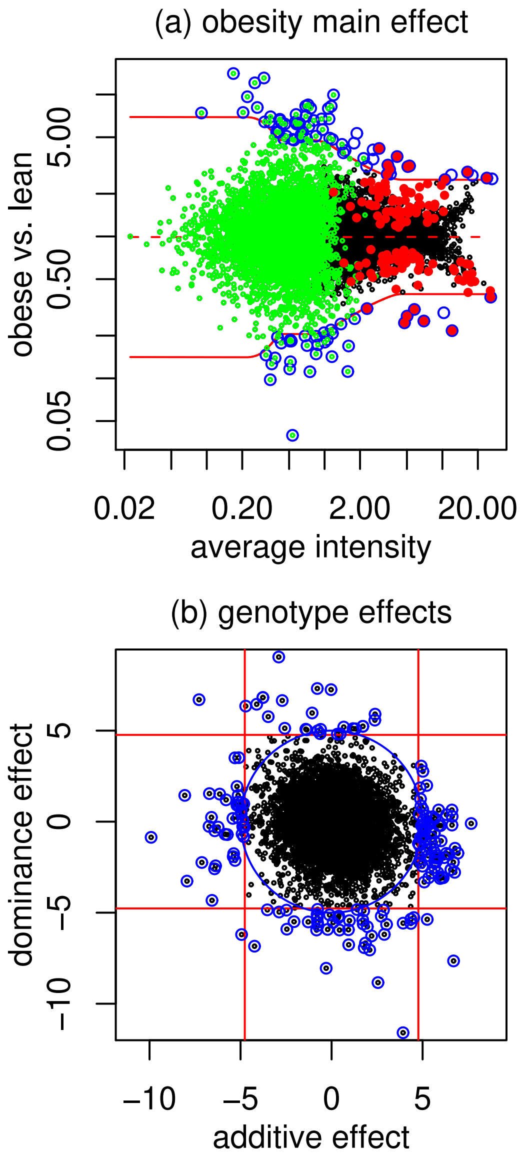

FIGURE 10. Diabetes and obesity study. Solid red line is our 5% Bonferroni limit. (a) obesity main effect: green points have 1–6 negative adjusted values; purple points detected in Nadler *et al.* (2000); blue points beyond 5% line detected by our procedure. Clearly methods based only on fold cannot detect all interesting patterns. (b) genotype main effects: T scores scatter plot with blue circles on detected genes.