UNIVERSITY OF WISCONSIN — MADISON
DEPARTMENT OF STATISTICS

Mining for Low-abundance Transcripts in Microarray Data

Yi Lin [1]

Samuel T. Nadler[2]

Alan D. Attie[2]

Brian S. Yandell[1,3*]

[1]Department of Statistics, University of Wisconsin-Madison, 1210 Dayton St., Madison, WI 53706

[2]Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Dr., Madison, WI 53706

[3]Department of Horticulture, University of Wisconsin-Madison, 1575 Linden Dr., Madison, WI 53706

Technical Report # 1031

January 2001

Department of Statistics
University of Wisconsin-Madison
1210 West Dayton Street
Madison, WI 53706-1685

Phone:	608/262-2598
Fax:	608/262-0032
Internet:	yandell@stat.wisc.edu

Summary

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 a data-driven *p*-value that is sensitive to changes in variability with gene expression. Together, these two features expand the repertoire of genes that can be analyzed with DNA arrays.

Introduction

Microarray technology to measure gene expression is becoming widespread . The application of microarray analysis to such diverse biological processes as aging (1), cancer (2,3), diabetes (4), and obesity (4,5) 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 (6,7) search for genes that show similar changes in expression across experimental conditions. These methods do not determine the significance of the changes in gene expression, and they require extensive pre-filtering of the data to eliminate genes with low intensity or modest fold changes. Thus, much information is lost before gene clustering can begin. Furthermore, it has become apparent that at different gene expression levels, different thresholds for significant changes are needed (8,9,10). 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 (11).

We present a robust statistical approach to more accurately assess data from microarray experiments. Simulation studies indicate that our approach is robust. The application of this method to mouse experiments studying diabetes and obesity uncovered changes in gene expression missed by other methods. Details of the method are provided, including information on how to obtain public domain software.

Experimental Procedures

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. It requires only minimal assumptions to assign Bonferroni-corrected *p*-values, and can be used alternatively to prefilter genes to be organized further by clustering methods.

Expression data may be acquired from spotted cDNA arrays or from oligonucleotide arrays. After data acquisition, expression levels are typically adjusted for background, which can lead to negative values. Our procedure rank-orders the adjusted values by intensity, and converts the ranks into normal scores (Figure 1). This normal scores transformation was initially proposed in another setting (12), and has been employed for microarray data in a slightly different approach (13). If expression data are approximately log normal, then this normal scores transformation will be very close to a log transformation. In addition, it readily accommodates negative adjusted expression values that are discarded or set to some arbitrary value by most other procedures that use the log. This transformation, which automatically standardizes the data, may be done by condition or across all conditions, depending on the situation (e.g. whether experimental conditions are at fixed levels or are a random sample of possible conditions).

The average intensity value for each gene across all experimental conditions is then calculated as a mean of these normal scores. Differential expression across conditions of interest is computed by contrasting normal scores. If there are only two conditions, a plot of the mean expression against the difference is just a 45 degree rotation of the plot of the two conditions.

Differential gene expression between experimental conditions may depend on the average level of gene expression. Therefore, we use estimates of the center and spread that can vary across average gene expression to standardize differential expression, specifically smoothed medians and smoothed median absolute deviations, respectively. Differential contrasts standardized by these center and spread should have approximately the standard normal distribution for genes that have no differential expression across the experimental conditions.

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 (11,13). Testing raises the need to account for multiple comparisons, here we use *p*-values derived using a Bonferroni-style genome-wide correction (14). Genes with significant differential expression are reported in order of increasing *p*-value. Further details of this procedure and the software can be found in Supplementary Data.

Results and Discussion

Simulation Studies

Three simulation studies were conducted to examine properties of the normal scores procedure. The first study shows how well the smoothed median absolute deviations can estimate the variability of the uncontaminated part of the data. The second study verifies that our 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 (X,Y) pairs, with X from standard normal and Y normally distributed with mean 0 and standard deviation $\sigma(X)=1/[X/3 + 2.5]$. Then we generated another 500 pairs by adding independent standard normal random numbers to each Y value. Thus given the same X, the standard deviation of the contaminated Y is $[1 + (X/3 + 2.5)^2]^{1/2}$ times that of the uncontaminated Y (1.8 to 3.64 as X goes from −3 to 3). We applied our robust scaling function to the combined data of 10,000 pairs. A typical simulation result is shown in Figure 2. Figure 2(a-b) show scatter plots of the simulated data before and after the addition of contamination. Figure 2(c) 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 2(d) 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 *a* were set at 1. We considered a range of measurement error variances from none to high (δ = 0, 1, 2, 5, 10, 20). The "best" ranking would be based on the true differential expression between the two conditions. Figure 3a compares the top 500 "best" ranks when the true intensities are known with our procedure. In the absence of measurement error, our procedure essentially preserves the true ordering of differential expression (line 0). When a typical level of noise is applied to the simulation, the procedure faithfully captures most of the differentially expressed genes.

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 3b). 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 3c). None of the non-changing genes were identified; there were no false positives. In contrast, an early analytical method assuming a constant coefficient of variation (15) yielded conservative, flat thresholds (Figure 3c, dashed line). The Bayesian approach (11) missed the pattern of changing variation with average gene intensity and misses most of the differentially expressed genes (Figure 3c, dotted line).

Application to obesity

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. Nadler et al. (4) evaluated changes in gene expression between adipose tissue from lean, obese and obese-diabetic mice using oligonucleotide arrays with over 13,000 probes. The obesity experiment had six experimental conditions arranged in a two-way factorial with lean and obese mice from three different genotypes.

Roughly 100 genes were determined to have significant ($p<0.05$) changes in gene expression using the robust normal scores procedure (Figure 4). Almost half of these genes had at least one negative adjusted value in the dataset due to low expression (green), and were missed by other methods.

Table 1 (Supplementary Data) shows new genes identified by the Bonferroni criterion. Some of these genes are transcription factors, including I-κB, a modulator of transcription in connection with inflammatory processes, RXR, a nuclear hormone receptor that forms heterodimers with several nuclear hormone receptors. Other genes in this collection are proteins involved in regulation; *e.g.* protein kinase A and glycogen synthase kinase-3. The correlation of the expression of these genes with obesity raises interesting new questions about the consequences of obesity on adipocyte signaling pathways.

Figure 5 shows the density of standardized differences Z for obesity overlaid on the standard normal density, with good agreement for standardized differences between -2 and 2. The long-dashed line shows an estimate of the density for differentially expressed genes. This was produced by assuming the standard normal is correct for non-changing genes and picking the proportion of changing genes as just large enough so that this differential density is positive (*cf.* 13). This illustrates just how conservative the Bonferroni approach is. We have since begun examining the genes with standardized scores above 2 or below -2 SDs separately using hierarchical clustering. Initial results show clustering that is highly correlated with mean expression level, but with important rearrangement that might suggest functional association among genes in clusters. This work will be reported elsewhere.

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 (16) by adjusting for variability in expression across all conditions relative to the average gene expression. The utility of clustering (6,7) and classification (2) 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.

References

1 Lee, C. K., Klopp, R. G., Weindruch, R., and Prolla, T. A. (1999) *Science*, **285**, 1390-1394

2. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999) *Science*, 286, 531-537

3. Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams. C. F., ZhuS. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 9212-9217

4. Nadler, S. T., Stoehr, J. P., Schueler, K. L., Tanimoto, G., Yandell, B. S., and Attie, A. D. (2000) *Proc. Natl. Acad. Sci. U.S.A.,* **97**, 11371-11376.

5. Soukas, A., Cohen, P., Socci, N. D., and Friedman, J. M. (2000) *Genes & Development*, **14** ,963-980

6. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) *Proc. Natl. Acad. Sci. U.S.A.,* **95**, 14863

7. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999) *Proc. Natl. Acad. Sci. U.S.A.,* **96**, 2907-2912

8. Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D. D., Dai, H. Y., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C., Friend, S. H. (2000) *Science,* **287**, 873-880

9. Wittes, J., Friedman, H. P. (1999) *J. Natl. Cancer. Inst.*, **91**, 400-401

10. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H. Y., He, Y. D. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. (2000) *Cell*, **102**, 109-126

11. Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001) *J. Comp. Biol.,* **8**, 000-000

12. Klaassen, C. A. J., and Wellner, J. A. (1997) *Bernoulli*, **3**, 55-77

13. Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2000) *Tech. Rep.*, Dept. Statist., Stanford U.

14. Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2000) *Tech. Rep. 578*, Dept. Biochem., Stanford U.

15. Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997) *J. Biomed. Optics*, **2**, 364-374

16. Kerr, M. K., Martin, M., and Churchill, G. A. (2000) *Tech. Rep.*, Jackson Laboratory

Figure Legends

Figure 1. Normal scores transformation.

Any data set (a) can be transformed approximately to normal by first (b) replacing each datum by its rank and (b) replacing the rank by its normal score. For each panel, there are 30 observations, and the area for each datum is 1/30 so that the total area is 1.

Figure 2. Simulation to show estimated spread.

Simulated data with 10,000 genes and 5% contamination (see text for details). (a-b) show scatter plots of data before and after addition of contamination. (c) shows how close are the true (solid line) and estimated scale (dotted line) scale. (d) shows a Q-Q plot, which should be straight for normal data; here the middle portion is be almost straight while the tails diverge due to the "contamination" by differentially expressed genes.

Figure 3. Simulation of differential expression and effect of noise.

Data were simulated with intrinsic variability due to gene-specific hybridization efficiency, and varying amounts of measurement error. Five percent of the 10,000 simulated genes were assigned to display differential expression; the mean fold change for differentially expressed genes decreased as average intensity increased. With no measurement error, differential expression has constant variance on a log scale. As measurement error increases, the variance of differential expression decreases as average intensity increases. Axes are antilog of normal scores, which approximates fold change. Details of the simulation model can be found in the Supplementary Data. (a) Number of changed genes captured as measurement noise increases (0 to 20). Note almost perfect recovery of rank order with no noise, followed by gradual degradation as noise increases.

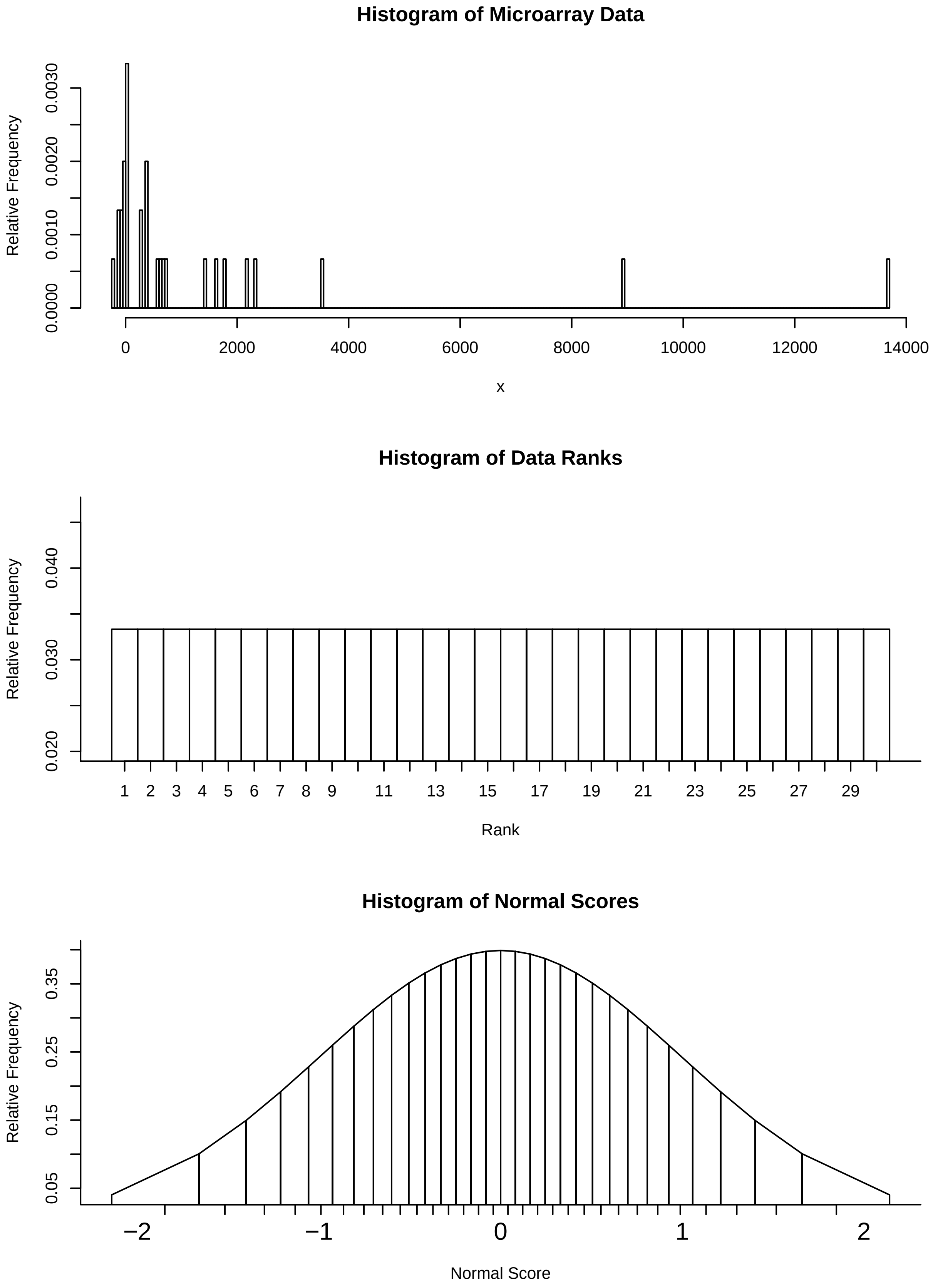
(b) Detection of fold change with no measurement error. Blue points represent genes with differential expression. Dashed lines are from Chen et al. (12). Dotted curves are from Newton et al. (11). Solid curves are from our procedure. Note that Chen and Newton provide similar results, while our procedure is more conservative. (c) Detection of fold change with high measurement error. Here lines for Chen and Newton methods are more conservative. Newton method does not appear to adequately capture the pattern of variability.

Figure 4. Gene expression in obesity.

Solid red line is our 5% confidence limit. Points in green have at least one of six readings with negative adjusted values. Purple points were detected by Nadler et al. (4). Additional blue points, and all points beyond 5% line, were detected by our procedure. Axes are antilog of normal scores, which approximates fold change. Note that methods based on fold change, even adjusting for changing variability across average intensity, may miss important genes uncovered by other methods.

Figure 5. Density of standardized differences.

The density for the standardized differences Z (solid line) for obesity is overlaid on the standard normal density (dashed line), with good agreement between -2 and 2. The long-dashed line shows an estimate of the density for differentially expressed genes. This was produced by picking the proportion of changing genes just large enough so that the differential density is positive, assuming the standard normal for non-changing genes. The dot-dashed line shows where the density for the standardized differences is twice that of the standard normal.


Histogram of Microarray Data

Histogram of Data Ranks

Histogram of Normal Scores

Figure 1

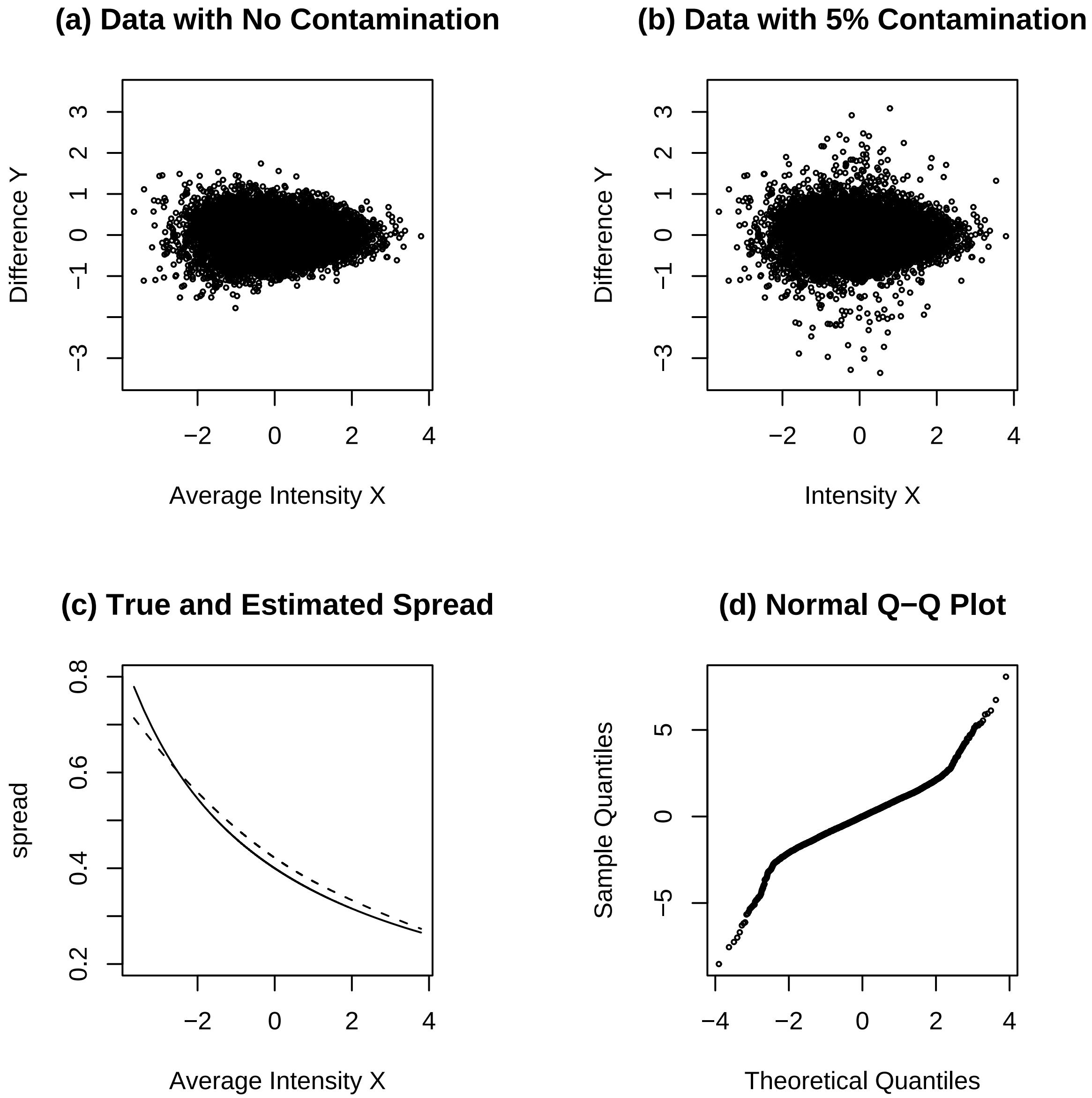
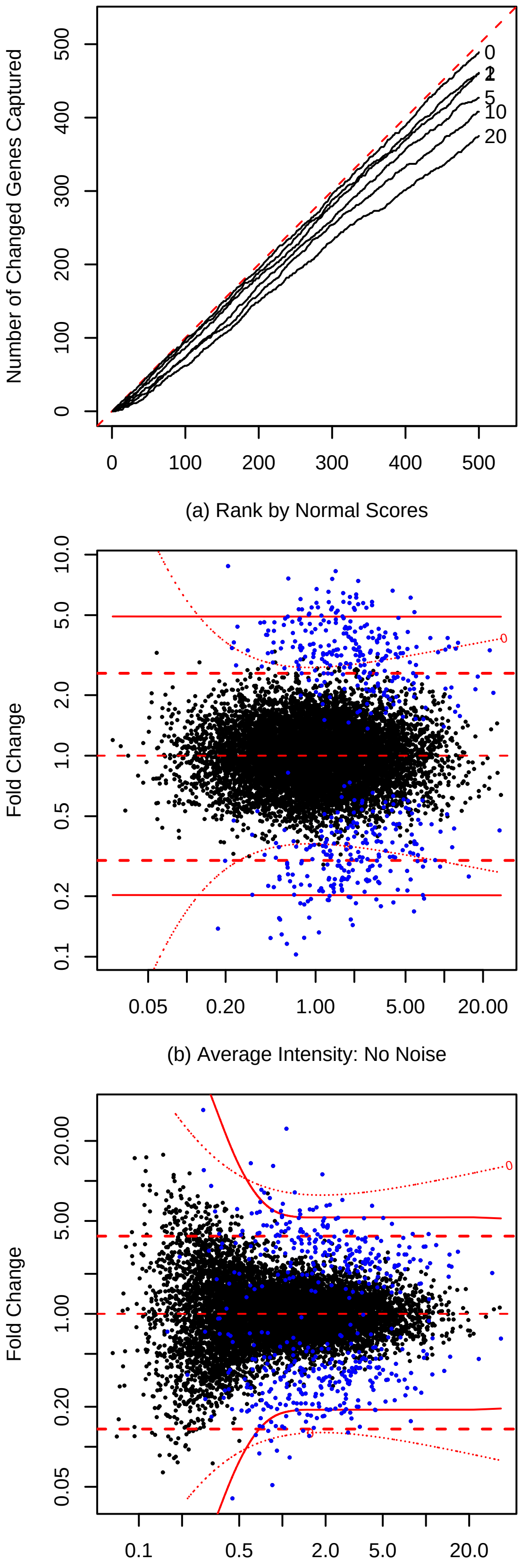

(a) Data with No Contamination
Difference Y
3
2
1
0
−1
−3
−2
0
2
4
Average Intensity X
(b) Data with 5% Contamination
Difference Y
3
2
1
0
−1
−3
−2
0
2
4
Intensity X
(c) True and Estimated Spread
spread
0.8
0.6
0.4
0.2
−2
0
2
4
Average Intensity X
(d) Normal Q−Q Plot
Sample Quantiles
5
0
−5
−4
−2
0
2
4
Theoretical Quantiles

Figure 2



(a) Rank by Normal Scores

(b) Average Intensity: No Noise

(c) Average Intensity: High Noise

Figure 3

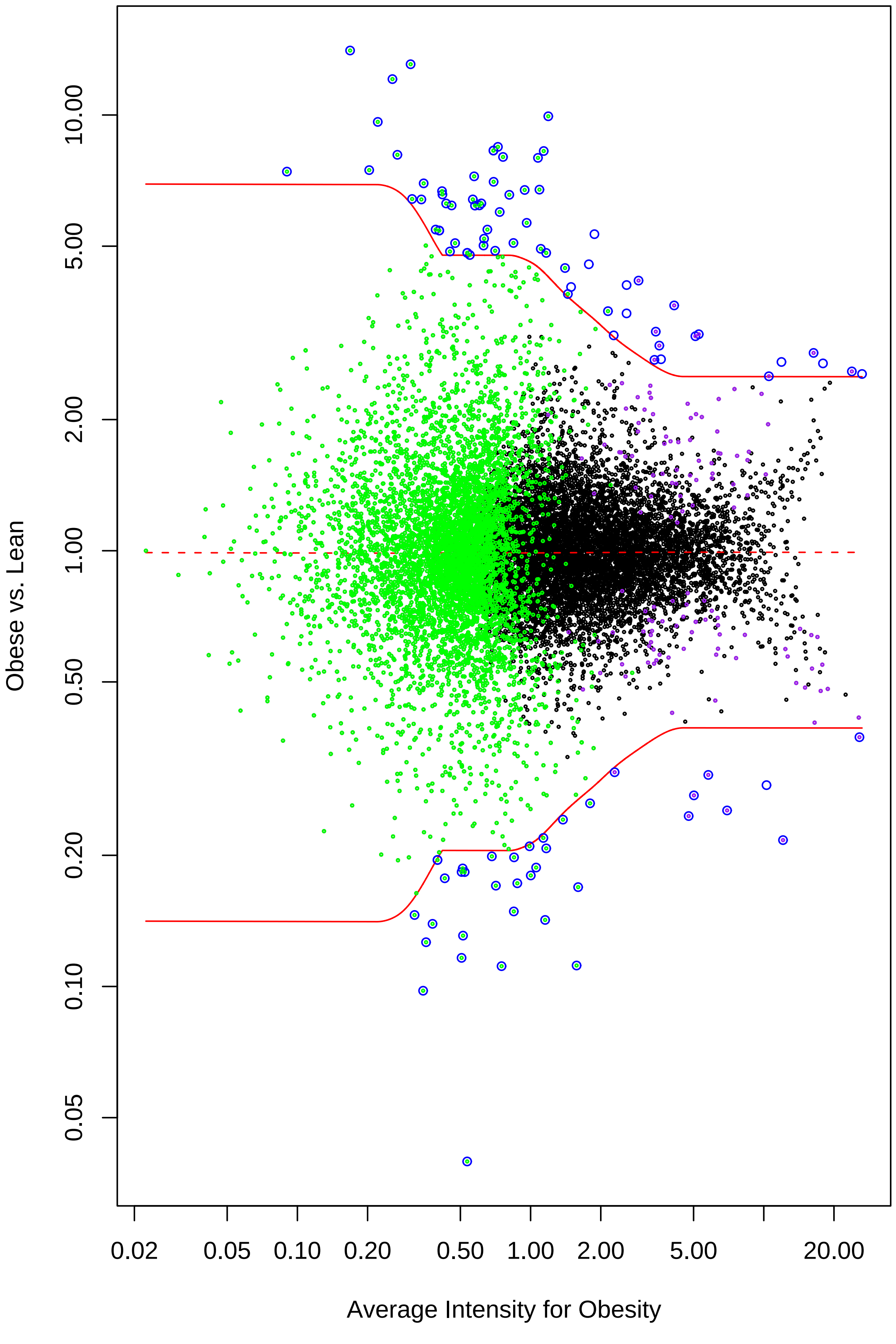

Obese vs. Lean
10.00
5.00
2.00
1.00
0.50
0.20
0.10
0.05
0.02
0.05
0.10
0.20
0.50
1.00
2.00
5.00
20.00
Average Intensity for Obesity

Figure 4

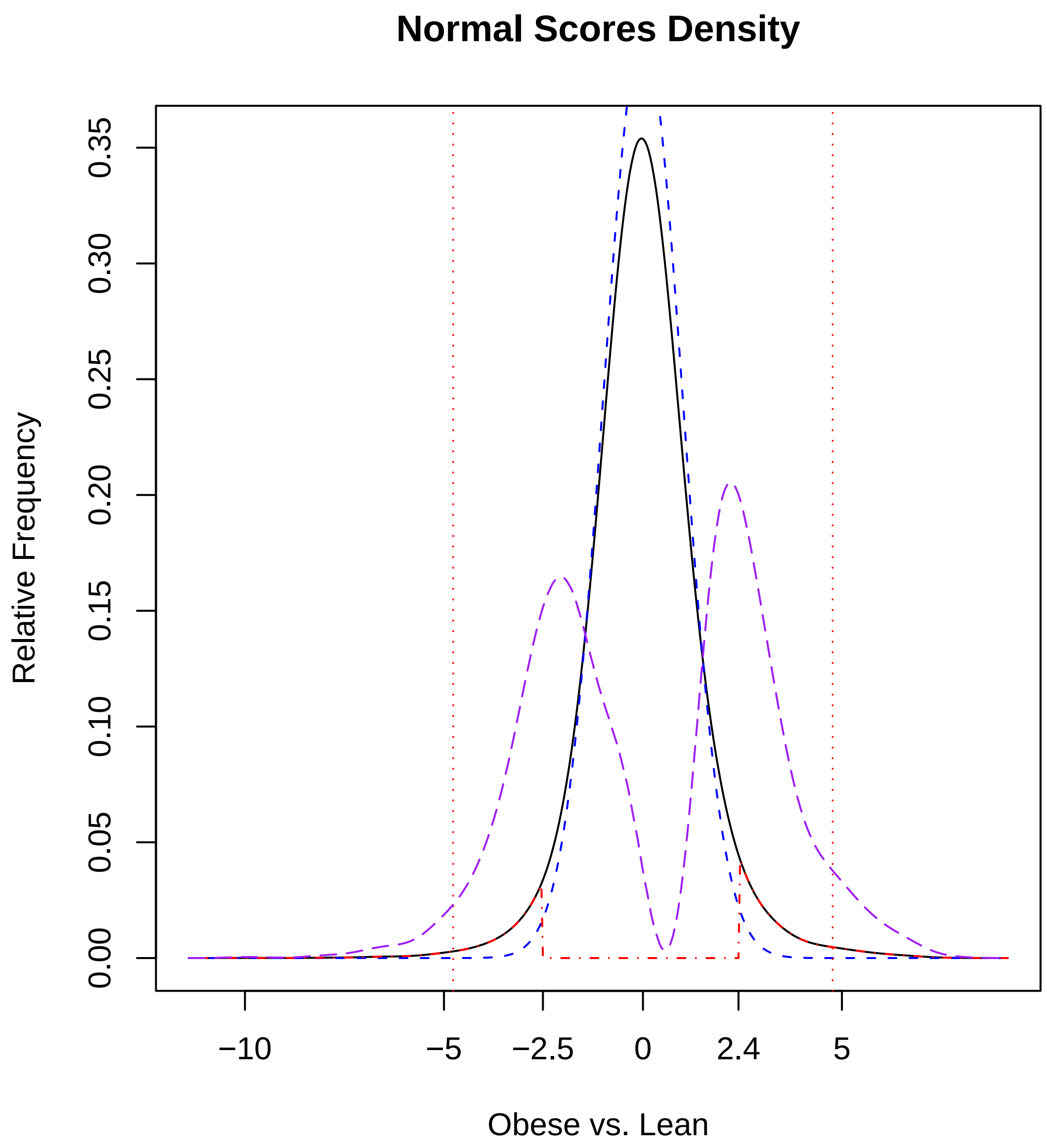

Normal Scores Density
Relative Frequency
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
−10
−5
−2.5
0
2.4
5
Obese vs. Lean

Figure 5

Supplementary Data

Raw microarray measurements are typically normalized to account for systematic bias and noise to attempt to restore expression levels from raw data. 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 are typically removed by subtraction or division. Background-adjusted intensities are typically log-transformed to reduce the dynamic range and achieve normality. However, negative adjusted values, arising from low expression swamped by background, are either dropped or adjusted upwards by a small constant before taking the log. Various authors have noted that comparisons based on such log-transformed gene expression levels appear to be approximately normal. Our alternative normalization method leverages this idea while providing comparisons that are more robust to difficulties with the log-normal assumption.

Our rank-based procedure depends on the existence of a monotone transformation (*e.g.* log of expression plus a constant, or square root) that is that transforms the data to near normal, but does not actually require that we know the transformation. Consider measurements under one condition. Let Q be the raw expression for a gene, and denote the background by B. The adjusted intensity is the difference $A=Q-B$ (or the relative difference $A=Q/B$). This is normalized by some transformation, although its form may be unknown. We prefer to approximate this transformation using the normal scores

$$N = \text{qnorm}[\ \text{rank}(A) \ / \ (n+1)\ ]$$

where rank(A) is the rank among all n adjusted gene measurements under the same condition. The normal quantiles, qnorm(), transform the ranks to be essentially a sample from standard normal: a histogram of these N would be bell-shaped and centered about zero (Figure 1). Thus these normal scores are close to a transformation that would make the data appear normal (1). If done separately by condition, this normalization automatically standardizes the scale and center. 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.

A.1. Motivating Model for Expression Data

The following model motivates the normal scores in the case of simple subtraction for background; see (2) for another approach. Here, the natural transformation to normality is the logarithm. The observed background intensity B for a gene is measured with error ω, with perhaps some attenuation d that may depend on the condition: $B = bd + \omega_B$. The observed raw measurement Q has in addition the gene signal g, which may be affected in a relative way by the degree of hybridization h: $Q = [a\exp(g+h+\varepsilon)+b]d+\omega_Q$. Here ε is the intrinsic noise (whose variance may depend on g) and a is the attenuation effect of the array and the thickness of the sample. Notice that g is confounded with h unless hybridization efficiency is gene specific with no dependence on the experimental condition.

Subtracting the background intensity B from Q yields the adjusted measurements $A = aG+\delta$ where $\delta=\omega_B-\omega_Q$ is symmetric around 0 and $\log(G) = g+h+\varepsilon$ is the log expression level. Thus G is the measurement if there were no measurement error, no dye or array effect, and no background intensity. Hence it is natural under this model to consider N=log(G), although the normal scores would be almost as good. This model forms the basis for simulations presented later.

A.2. Comparing Two Conditions

Comparison of gene expression between two conditions involves finding genes with strong differential expression. Typically, most genes show no real difference, except for that due to measurement variation. We propose a data-driven, robust standardization to assess differential expression that accounts for changing variance in differential expression with average intensity noted by other authors.

Genes at different average expression may have intrinsically different variability. Information on comparison of conditions is summarized in N_1 and N_2. However, the joint distribution of normalized values from two conditions across the n genes is in general not normal. Consider plotting the average intensity $X = (N_1+N_2)/2$ against the difference $Y = N_1\text{-}N_2$. The difference Y measures the change in gene expression level relative to the average intensity X. Our procedure standardizes Y by the variability of the intrinsic noise at X. For those genes with no change, the variance of Y depends approximately on X in some smooth way. The decreasing variance in Y as X increases comes in part from normalization: the same noise is more likely to disturb expression ranks at lower intensities.

A.3. Robust Center and Spread

Smoothing splines are combined with standardized local median absolute deviation (MAD) to provide a data-adapted, robust estimate of spread $s(X)$. A smooth, robust estimate of center $m(X)$ 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 normal scores $Z = [Y-m(X)]/s(X)$.

The genes are sorted and partitioned based on X into many (about 400) slices containing roughly the same number of genes and summarized by the median and the MAD for each slice. These should have roughly the same distribution up to a constant. To estimate the scale, it is natural to regress log(MAD) on X with smoothing splines, but other non-parametric smoothing methods would work as well. The smoothing parameter is tuned automatically by generalized cross validation (3). The anti-log of the smoothed curve, globally rescaled, provides an estimate of $s(X)$, which can be forced to be decreasing if appropriate.

It may be reasonable in some cases to use "house-keeping genes" that are generally believed to not change over different conditions (*cf*. 4). However, this approach 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. Notice, however, that this contamination is of primary interest.

The following model may help motivate our specification for spread. Consider again log(G) = $g+h+\varepsilon$ and suppose hybridization is negligible, or at least the same across conditions. However, the intrinsic noise ε may depend on the true expression level g. For two conditions 1 and 2, the difference Y is approximately $\log(G_1)-\log(G_2)= g_1-g_2 + \varepsilon_1-\varepsilon_2$. If there is no differential expression, $g_1=g_2=g$, then $\text{Var}(Y) = s^2(g)$, and g may be approximated by X. However, the true formula for $\text{Var}(Y\,|\,X)$ is not exactly $s^2(X)$, and cannot be determined without further assumptions.

A.4. Formal Testing Procedure

We can use the standardized scores Z to rank the genes. The conditional distribution of these Z given X are assumed to be standard normal across all genes whose expressions do not change between conditions. 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 $1-(1-p)^n$, 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.

Less conservative multiple comparison adjustment to p-values are proposed in (5). However, the results are essentially the same with all such methods, except when more than 5-10% of the genes change across conditions.

It may be appropriate to examine a histogram of standardized scores Z, using these critical values as guidelines rather than strict rules. Figure 5 shows a smooth density estimate of the histogram compared to the standard normal. Following (6), we recognize that the density f of all the scores is a mixture of the densities for nonchanging f_0 and changing f_1 genes:

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z).$$

We assume that f_0 is standard normal, and set π_1 just large enough so that the estimate

$$f_1(z) = [f(z) - \pi_0 f_0(z)] / \pi_1$$

is positive. This in some sense provides a "liberal" estimate of the distribution of differentially expressed genes, as shown in Figure 5. 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.

A.5. Experimental Design Considerations

This method can be extended to designed experiments, with multiple conditions, multiple readings (*e.g.* dyes) per gene on a chip, and replication of chips (7). The development for two conditions adapts naturally to contrasts capturing key features of differential gene expression across design factors.

Robust estimation methods may overlook the influence of outliers. Further, gross errors can be confused with changing genes. Good design dictates replication and factorial design across multiple chips, which can be used to detect outliers in a similar fashion to the approach for differential gene expression. Residual deviations of each replicate from the mean could be plotted against the average intensity. Robust estimates of center and scale could be used as above in formal Bonferroni-style tests for outliers.

Time or other progressions over multiple levels might be examined for linear or quadratic trends using orthogonal contrasts (8). With multiple factors, polynomial or other orthogonal contrasts can be considered for main effects and for interactions. Each contrast can be analyzed in a similar fashion to the above.

Consider r condition levels in increasing order. The importance of each polynomial component in the data can be measured by a specific trend associated with a contrast. That is, consider the average intensity $X=\sum N_k/r$ and contrasts of the form $Y=\sum c_kN_k$. For instance, with five conditions representing a linear series of glucose levels, one might investigate linear and quadratic contrasts: $Y_1=2N_5+N_4-N_2-2N_1$ and $Y_1=2N_5-N_4-2N_3\text{-}N_2+2N_1$. Again, assume that most genes are not changing, and proceed with a similar specification as for two conditions. 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.

A.6. Software Implementation

The analysis procedure is 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 "microarray" library. The function robustscale() computes the center $m(X)$ and spread $s(X)$, while the function pickgene() plots Y against X, after backtransforming to show fold changes, and picks the genes with significant differences in expression. Examples include the simulations and graphics presented here. A separate library contains the obesity and diabetes microarray data.

References

1. Klaassen, C. A. J., and Wellner, J. A. (1997) *Bernoulli,* **3**, 55-77
2. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H. Y., He, Y. D. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. (2000) *Cell*, **102**, 109-126
3. Wahba, G. (1990) *Spline Models for Observational Data* (Soc. Indust. Appl. Math., Philadelphia, PA, 1990).
4. Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997) *J. Biomed. Optics,* **2**, 364-374
5. Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2000) *Tech. Rep. 578*, Dept. Biochem., Stanford U.
6. Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2000) *Tech. Rep.*, Dept. Statist., Stanford U.
7. Kerr, M. K., Martin, M., and Churchill, G. A. (2000) *Tech. Rep.*, Jackson Laboratory
8. Lentner, M., and Bishop, T. (1993) *Experimental Design and Analysis*, 2nd ed., Valley Book Company, Blacksburg, VA

Accesion #	fold	Description	Function	average	p-value
AA033381	-5.46	No significant homologies	-	0.506	0.0028
aa190005	-5.33	No significant homologies	-	1.056	0.0015
AA198324	-5.64	No significant homologies	-	0.428	0.0016
aa267014	-7.04	No significant homologies	-	1.154	0.0000
aa544897	-10.23	No significant homologies	-	0.346	0.0000
aa596237	-4.14	No significant homologies	-	1.377	0.0095
aa616759	-5.46	No significant homologies	-	0.522	0.0027
c80103	2.74	No significant homologies	-	3.398	0.0076
AA048974	6.40	No significant homologies	-	0.565	0.0001
AA035870	6.74	No significant homologies	-	1.091	0.0000
AA103862	8.10	No significant homologies	-	0.268	0.0013
w19022	7.03	No significant homologies	-	0.694	0.0000
w45955	-7.92	No significant homologies	-	0.356	0.0001
w49331	-5.05	No significant homologies	-	0.849	0.0093
w64688	6.73	No significant homologies	-	0.943	0.0000
W48968	3.55	Sequence Withdrawn	-	2.147	0.0016
W51181	5.08	Sequence Withdrawn	-	0.475	0.0062
AA168767	7.41	Sequence Withdrawn	-	0.090	0.0072
ET63493	6.27	Mus musculus cyclin-dependent kinase inhibitor (p15INK4b)	Cell Cycle	0.434	0.0002
aa711625	2.58	Similar to Rat ERG2	Cholesterol Metabolism	23.856	0.0081
AA146437	3.18	Mus musculus cathepsin S	Cysteine Protease	3.442	0.0001
AA089333	3.66	Mus musculus cathepsin S precursor	Cysteine Protease	4.128	0.0000
W13263	4.07	Mus musculus ctsk	Cysteine Protease	2.581	0.0000
U59807	3.14	Mus musculus cystatin B (Stfb)	Cysteine Protease	5.263	0.0000
AA009095	5.46	Similar to cathepsin D	Cysteine Protease	0.392	0.0052
X54511	4.17	M.musculus Myc basic motif homologue-1	Cytoskeletal Function	2.904	0.0000
X54511	2.75	M.musculus mbh1	Cytoskeletal Function	3.625	0.0040
X64361	4.83	M.musculus vav	Cytoskeletal Function	0.535	0.0140
AA050934	-3.80	Similar to Rat kinesin-related protein 2 (KRP2)	Cytoskeletal Function	1.799	0.0039
AA062269	-5.80	Mouse protamine 1	DNA Binding Protein	0.877	0.0009
m27501	-4.77	Mus musculus protamine 2	DNA Binding Protein	0.990	0.0143
X87096	6.40	M.musculus brevican	ECM Function	0.340	0.0038
X99143	7.47	M.musculus hair keratin, mHb6	ECM Function	0.203	0.0064
x14951	3.12	CD18 antigen beta subunit, leukocyte adhesion protein (LFA-1)	ECM Function	2.272	0.0142
I13732	5.66	Mouse macrophage-specific integral membrane protein	ECM Function	0.962	0.0006
AA080172	-4.06	Mus musculus phosphoenolpyruvate carboxykinase 1	Gluconeogenesis	4.763	0.0000
W47728	6.20	Similar to Rat glycogen synthase kinase 3 alpha	Glycogen Metabolism	0.577	0.0002

U96386	-4.82	Mus musculus activin beta E subunit	Growth Factor	1.167	0.0035
u02883	-5.56	Mus musculus Balb/c mammary-derived growth inhibitor (MDGI)	Growth Factor	1.002	0.0010
AA097231	5.02	Mus musculus Rac2	GTP Binding Protein	0.628	0.0076
AA137962	7.97	similar to Human rab-14	GTP Binding Protein	1.076	0.0000
ET62522	5.20	Mus musculus defensin	Immune Function	0.632	0.0042
W41745	3.11	Mus musculus Fc receptor	Immune Function	5.090	0.0000
M21285	-2.68	Mouse stearoyl-CoA desaturase	Lipogenesis	25.716	0.0046
aa137436	2.69	Similar to stearoyl-CoA desaturase	Lipogenesis	17.940	0.0025
M21050	2.84	Mouse lysozyme M	Lysosomal Function	16.350	0.0005
X63349	-5.92	M.musculus tyrosinase-related protein-2	Metabolism	1.599	0.0000
j05663	-5.03	Mouse vas deferens androgen related protein (MVDP)	Metabolism	0.682	0.0105
U16297	-8.60	Mus musculus cytochrome B561	Metabolism	0.506	0.0000
AA023099	9.64	Mus musculus dUTPase	Metabolism	0.221	0.0002
AA111277	6.97	Mus musculus hippocalcin-like 1 (Hpcal1)	Metabolism	0.348	0.0008
AA116710	-6.73	Mus musculus serum response factor	Metabolism	0.847	0.0001
W10926	8.01	Mus musculus ubiquitin like protein	Metabolism	0.762	0.0000
c76068	2.71	Similar to Mouse mitochondrial genes coding for three transfer RNAs	Mitochondrial	11.905	0.0020
M20625	5.43	Cytochrome c	Mitochondrial	0.406	0.0033
u69135	2.51	Mus musculus UCP2	Mitochondrial	10.515	0.0161
W16250	6.20	Similar to Mus musculus ATP synthase	Mitochondrial	0.459	0.0002
J04179	8.27	Mouse chromatin nonhistone high mobility group protein	Nuclear Protein	1.138	0.0000
J04179	3.88	Mouse chromatin nonhistone high mobility group protein (HGM-I(Y))	Nuclear Protein	1.445	0.0131
m30844	-3.94	Mus musculus B2 protein	Nuclear Protein	6.960	0.0000
aa408365	-7.18	Mus musculus SRp25 nuclear protein	Nuclear Protein	0.380	0.0001
x93999	-7.64	Gal beta-1,3-GalNAc-specific GalNAc alpha-2,6-sialyltransferase gene	Protein Metabolism	0.513	0.0000
U77083	5.09	Mus musculus CD13/aminopeptidase N	Protein Metabolism	0.844	0.0059
U93862	2.54	Mus musculus ribosomal protein L41	Protein Synthesis	26.383	0.0116
x03479	4.03	Mouse mRNA fragment for serum amyloid A (SAA) 3	Secreted Protein	1.491	0.0049
X03505	8.29	Mouse serum amyloid A	Secreted Protein	0.693	0.0000
W17473	-3.64	Mus musculus angiotensinogen	Secreted Protein	5.016	0.0000
AA106347	-3.22	Mus musculus angiotensinogen precursor	Secreted Protein	2.293	0.0095
M82831	5.33	Mus musculus macrophage metalloelastase	Secreted Protein	1.878	0.0000
W85163	14.06	Mus musculus migration inhibitory factor	Secreted Protein	0.168	0.0000
ET63455	6.42	Mus musculus serum amyloid A-4 protein (Saa4)	Secreted Protein	0.310	0.0108
AA124352	4.88	Similar to Human neuromedin B	Secreted Protein	0.705	0.0117
x04673	-3.45	Mouse adipsin	Serine Protease	10.273	0.0000
W36455	-4.61	Mus musculus adipsin	Serine Protease	12.089	0.0000
AA105229	-6.85	Mus musculus hepsin	Serine Protease	0.318	0.0047

ET62360	6.69	Calcium-dependent phospholipase A2 precursor	Signal Transduction	0.417	0.0001
X72862	-8.95	M.musculus beta-3-adrenergic receptor	Signal Transduction	1.575	0.0000
x93328	4.54	M.musculus F4/80	Signal Transduction	1.777	0.0001
x72862	-3.27	M.musculus beta-3-adrenergic receptor	Signal Transduction	5.781	0.0000
X65026	6.56	M.musculus GTP-binding protein	Signal Transduction	0.810	0.0001
m31810	-5.36	Mouse 2',3'-cyclic-nucleotide 3'-phosphodiesterase	Signal Transduction	0.512	0.0037
D14883	-4.56	Mouse C33/R2/IA4	Signal Transduction	1.134	0.0119
J02935	-5.12	Mouse cAMP-dependent protein kinase type II regulatory subunit	Signal Transduction	0.399	0.0148
m19681	4.93	Mouse platelet-derived growth factor-inducible protein	Signal Transduction	1.106	0.0027
AA168061	6.20	Mus musculus adenylate kinase	Signal Transduction	0.603	0.0002
u77460	9.93	Mus musculus anaphylatoxin C3a receptor	Signal Transduction	1.191	0.0000
ab009287	2.96	Mus musculus Macrosialin	Signal Transduction	3.565	0.0006
U19799	4.86	Mus musculus IkB-beta	Signal Transduction	0.451	0.0125
w14147	4.77	Mus musculus TOM1	Signal Transduction	0.549	0.0166
AA154294	13.08	Mus musculus non-receptor protein tyrosine phosphatase	Signal Transduction	0.306	0.0000
U09507	8.45	Mus musculus p21 (Waf1)	Signal Transduction	0.724	0.0000
w12140	4.83	Similar to Human putative receptor protein (PMI)	Signal Transduction	1.167	0.0024
AA138292	6.57	Similar to Rat S6 kinase	Signal Transduction	0.419	0.0001
AA114591	6.27	Similar to Rat type III adenylyl cyclase	Signal Transduction	0.615	0.0002
X12521	-8.98	Mouse transition protein 1 TP1	Telomeric	0.751	0.0000
AA144629	-5.87	Mus musculus transition protein 1	Telomeric	0.710	0.0008
x66224	12.09	M.musculus retinoid X receptor-beta	Transcription Factor	0.256	0.0000
m32370	4.46	Mouse transcription factor PU.1	Transcription Factor	1.405	0.0013
w48392	7.23	similar Id4 helix-loop-helix protein	Transcription Factor	0.573	0.0000
aa259645	5.99	Mus musculus BING4	Unknown	0.737	0.0004
d10911	3.51	Mus musculus DNA for MS2 protein	Unknown	2.580	0.0002
aa608251	-25.21	Dp1l1 mRNA for polyposis locus protein 1-like 1 (TB2 protein-like 1)	Unknown	0.535	0.0000
Z31278	5.46	Mus musculus T-ZAP	Unknown	0.653	0.0019