Adaptive Gene Picking for Microarray Expression Data Analysis

Brian S. Yandell, Yi Lin, Hong Lan and Alan D. Attie
Univeristy of Wisconsin–Madison

October 18, 2010

1 Overview

The following is adapted from Lin et al. (2002). References can be found there or at `http://www.stat.wisc.edu/~yandell/statgen`.

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. Although 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.

1.1 Transformation to Approximate Normality

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 (see Lockhart et al. 1996; Li and Wong 2001; Schadt et al. 2001; Irizarry et al. 2002; Li and Wong 2002). 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 (Li and Wong 2001; Schadt et al. 2001; Li and Wong 2002) to reduce measurement variability. More recently, Irizarry et al. (2003) and Zhang et al. (2003) have proposed normalization *without* using MM measurements.

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 (see Kerr and Churchill 2001). However, negative adjusted values can arise from low expression levels swamped by background noise. Some authors have proposed adding a small value before taking the 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 lognormal assumption. For further discussion on normalization, see Dudoit and Yang (2002) and Colantuoni et al. (2002).

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 lognormal, then this normal-scores transformation is indistinguishable from a log transformation after rescaling. We have found that log-transformed data appear roughly normal in the middle of the distribution, while the normal scores are normal throughout.

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 lognormal, 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

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

where $\text{rank}(\Delta)$ 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, with normal scores equally spaced in terms of probability mass. 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.

1.2 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 of 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 (Roberts et al. 2000; Dudoit and Yang 2002; Irizarry et al. 2002; Lee and O'Connell 2002; Colantuoni et al. 2002; Wu et al. 2002). 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 antilog or exp scale. Thus, the a axis is centered on 1 and suggests a fold change in intensity, while the d axis suggests a 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 (Kerr et al. 2001; Wu et al. 2002). 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, $\epsilon_{ijk} \sim \Phi(0, \sigma_j^2)$ 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 (see 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 nonzero, 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

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

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 (Lee et al. 2000; Kerr et al. 2001; Newton et al. 2001). This ANOVA framework allows isolation of the $(cg)_{ij}$ differential expression from the g_j gene

effect by contrasting conditions. Suppose that 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

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

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.

Although 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 fashion similar to that above. Alternatively, one can examine factors with multiple levels, say three genotypes, by an appropriate ANOVA evaluation (Lee et al. 2000).

1.3 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 nonchanging genes. Although 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 (Hughes et al. 2000; Roberts et al. 2000; Baldi and Long 2001; Kerr et al. 2001; Long et al. 2001; Newton 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. 2000). We noticed that empirically the variance across nonchanging 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 that of 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 that 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 expressions 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. For example, with 12,000 genes, the 30 contrasts d for each slice are sorted; the average of ordered values 15 and 16 is the median,

5

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 nonparametric smoothing methods would work as well. The smoothing parameter is tuned automatically by generalized cross-validation (Wahba 1990). The antilog 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 $d_j = \sum w_i \bar{x}_{ij}.\sqrt{K}$ 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 $\hat{\sigma}_j^2 = \sum_k (x_{ijk} - \bar{x}_{ij}.)^2 / \nu_1$, $\nu_1 = I(K-1)$, and ν_0 is the empirical Bayes estimate (see Lönnstedt and Speed 2001) 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.

1.4 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 (Kerr et al. 2001; Newton 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.

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 I 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 0.05 to 0.20 change in p-value). This test can be made one-sided if preferred.

Apparently less conservative multiple-comparison adjustments to p-values are proposed in Yang et al. (2000). 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. For an alternative interpretation of p-values in terms of false discovery rates, see Storey and Tibshirani (2003).

It may be appropriate to examine a histogram of standardized differences T using these critical values as guidelines rather than as strict rules. The density f of all the scores is a mixture of the densities for nonchanging 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 fashion similar to the approach for differential gene expression. Residual deviations of each replicate from the condition by 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.

2   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 that 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 scatterplot 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, the 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, and then smoothes 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 $\pi_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.

3   References

Lin Y, Nadler ST, Lan H, Attie AD, Yandell BS (2002) Adaptive gene picking with microarray data: detecting important low abundance signals. in *The*

Analysis of Gene Expression Data: Methods and Software, ed by G Parmigiani, ES Garrett, RA Irizarry, SL Zeger. Springer-Verlag, ch. 13.