STATISTICAL ANALYSIS OF HIGH DENSITY OLIGONUCLEOTIDE ARRAYS: A *SAFER* APPROACH

Daniel Holder, Richard F. Raubertas, V. Bill Pikounis, Vladimir Svetnik, Keith Soper
Merck Research Laboratories, WP37C-305,West Point, PA 19486

Key Words: median polish, microarray, probe-specific effects

Abstract

High-density oligonucleotide arrays allow researchers to measure mRNA transcript abundance for thousands of genes on a single array. The large number of genes, multiple sources of variation, and typically small number of experimental units (EUs) combine to make analysis of data from these arrays challenging. We describe our experience in applying data analytic techniques to replicated microarray experiments. In particular, we focus on a framework that includes attention to Scale, Additive Fits to account for probe and chip biases, assessment of Experimental-unit variability, and where possible, using processes that are Resistant to extreme values (hence the SAFER acronym). Our approach aims to provide results in terms that are familiar to the analyst and easily interpreted by biologists. Moreover, by emphasizing resistant methods and not ignoring variability due to sampling, the approach aims to provide measures that speak to the repeatability of the results.

1 Introduction

The ultimate goals of gene expression experiments using microarrays and the analysis of data arising from these experiments are varied. Elucidation of biological pathways, classification of samples or compounds, and discovery of patterns of co-regulation are only a few of the many uses for this assay. However, underlying these high level questions are the more basic questions of 1) for which genes have we detected expression, and 2) for which genes has the expression level changed between experimental conditions. We are interested in addressing these questions in a statistically sound and effective manner. To do so requires an approach that allows us to effectively a) quantify transcript abundance b) assess the precision of our estimate of transcript abundance and c) assess the quality and integrity of the data. This manuscript focuses on a few important aspects of a) and b).

2 Data description

We consider data arising from high-density oligonucleotide DNA arrays. Characteristics of these arrays are described elsewhere ([1, 2]). Briefly, each gene or EST to be queried is represented by a probeset on the array. An array contains thousands of probesets. Typically, each probeset consists of 14 to 20 probe pairs, where each probe-pair is composed of a perfect match (PM) oligonucleotide probe and a mismatch (MM) oligonucleotide probe. Labeled samples representing mRNA populations are hybridized to the arrays and gene expression is quantitated by fluorescence. We generally begin analysis after the initial image processing has summarized the pixel intensities, so that we have one fluorescence value for each probe.

3 The Basic Approach

There are multiple sources of variation in these experiments, including variation due to the sampling of the EUs, preparation of the samples (including isolation of mRNA and production of cRNA), measurement error due to the hybridization of the array, fluorescence detection, and contamination (e.g. dust).

We believe that there are advantages to looking at all of the arrays in an experiment simultaneously. The signal of individual probe-pairs is consistent over the arrays, and thus can be modeled and used to facilitate quantification on an individual array. This approach was pioneered by Li and Wong([3]). Li and Wong fit a multiplicative model to fluorescence values Y=PM-MM,

$$Y = (\text{probe effect}) * (\text{chip effect}) + \text{error}$$

using least squares. To obtain resistance to extreme values Li and Wong take an iterative approach of testing for and omitting outliers. We prefer to fit a simi-

lar model using methods that are more resistant than least squares. In particular we fit the model

$$f(Y) = \text{probe effect} + \text{chip effect} + \text{error},$$

using a highly resistant polishing technique ([4, 5]). To enhance additivity the fluorescence values (PM-MM) are transformed. Li and Wong's success with a multiplicative model suggests using f(Y)=log(Y). However, since log(Y) is undefined for Y < 0 and has large derivative for small Y, we use a linear-log hybrid transformation for $f()$. This transformation has the property that it is linear over an interval [0, c], but changes like log Y for large Y. It is continuous and smooth (continuous first derivative for Y > 0) and can improve the fit of an additive model to the data (as judged by a Tukey one degree of freedom test for additivity). Results also suggest that the transformed data have variances that are fairly stable over a wide range of expression values.

For many experiments, the largest components of variability are likely to be 1) between-EU variability and 2) between preps (mRNA, cDNA) variability. Yet the emphasis in many papers describing analysis of microarray experiments has been on estimating measurement error (pixel-pixel, spot-spot, gene-gene(within an array)). To estimate between-EU variability, it is necessary to measure at least two EUs in one or more treatment groups. Moreover, it seems unlikely that between-EU variability is constant or a simple function of expression level for all of the genes. In some circumstances, for example when one of the goals of the experiment is to compare the variability of expression levels for different genes, the assumption of similar sampling variability across the EUs is inappropriate.

Accordingly, we have found it useful to employ techniques to assess the sampling variability for each gene individually. In particular, since the number of EUs sampled is usually small, we have found the ANOVA method useful for estimating a variability parameter that includes sampling variability. Tests in which the observed test statistics are assessed against a distribution obtained by permuting the labels attached to the EUs are also useful (e.g., [6]).

4 Details of our Algorithm

In this section we describe some specifics of an algorithm we have employed that exemplifies our approach. First, we set notation and terminology. Our approach assumes that the experiment included $M \geq 1$ experimental conditions, referred to as treatment groups. One or more arrays are present for each group. The arrays within a group are referred to as replicates. At least one treatment group must have two or more replicate arrays.

Treatment groups are indexed by i, $i = 1, \ldots, M$. Replicate arrays within a treatment group are indexed by j, $j = 1, \ldots, n_i$, where n_i is the number of arrays for group i. Thus the pair of indices ij identifies a single array. On each array there are K probesets, indexed by k. Probes within a probeset are indexed by ℓ, $\ell = 1, \ldots, L_k$, where the number of probes per probeset, L_k, is typically 14–20. Note that k also will be used to index the gene nominally probed by probeset k.

We begin statistical analysis with the probe-level data. Probe-level data consist of PM and MM values for each probe, for each probeset, for each array. $PM_{ijk\ell}$ is the PM value for the ℓ-th probe in probeset k on the j-th array in treatment group i. $MM_{ijk\ell}$ is defined similarly.

$Y_{ijk\ell}$ will represent the intensity "signal" that is the basis for estimating gene expression levels. Typically it will be $PM_{ijk\ell} - MM_{ijk\ell}$, possibly adjusted for background or positional effects on the array, but it could also be $PM_{ijk\ell}$ alone.

In our implementation of the approach we allow for the possibility that $Y_{ijk\ell}$ might be missing for some probes on some arrays (e.g. "masked" probes).

4.1 Scaling

This step performs a simple pre-normalization to adjust for overall differences in brightness between arrays.

a) For each probe in each probeset, we calculate the median of MM across all arrays:

$$med_{k\ell} = median(MM_{11k\ell}, \ldots, MM_{Mn_Mk\ell})$$

b) For each array ij, we calculate a resistant, weighted linear regression of $(MM_{ijk\ell} - med_{k\ell})$ on $med_{k\ell}$. The weights are $1/med^2_{k\ell}$. The result is a fitted slope coefficient b_{ij} for each array. We have typically used the MM-estimator, implemented as the `lmRobMM()` function in S-PLUS ([7]).

c) Multiply all the signal values on an array by the scaling factor $SF_{ij} = 1/(1 + b_{ij})$. The scaled signal is

$$SS_{ijk\ell} = Y_{ijk\ell} SF_{ij}.$$

d) We typically plot the scaling factors SF_{ij} so that the variation between arrays can be examined.

4.2 Transformation

In this step we apply a nonlinear transformation, the "hybrid transformation" to signal values. The pur-

pose is to improve the additivity of array- and probe-specific effects (subsection 4.3), and to stabilize the variability of intensity values across a wide range of intensities.

$$\text{Transformed signal} = TS_{ijk\ell} = f(SS_{ijk\ell}),$$

where

$$f(x) = \begin{cases} a & \text{if } x < a \\ x & \text{if } a \le x \le c \\ c\ln(x/c) + c & \text{if } x > c. \end{cases}$$

The transformation is specified by two parameters a and c. We have typically used $a = 0$ and $c = 20$.

4.3 Additive fits to adjust for probe-specific effects

For each probeset k, we calculate a resistant additive fit to estimate array- and probe-specific effects on transformed signal. The array effects are adjusted for the probe effects.

a) The additive model is

$$TS_{ijk\ell} = GE_k + A_{ijk} + P_{k\ell} + \epsilon_{ijk\ell}$$

where

GE_k = Grand effect: overall signal level for probeset k across all probes and arrays

A_{ijk} = Array-specific effect for array ij

$P_{k\ell}$ = Probe-specific effect for probe ℓ

$\epsilon_{ijk\ell}$ = Residual variability

b)We fit the additive model using a resistant method such as median polish (see section 6B of [8]). Median polish is an iterative method that operates on a matrix by alternately extracting row and column medians. The result can sometimes depend on whether the iteration starts with rows or with columns. We have adopted the convention that the iteration starts with extracting medians for arrays (across probes). Iteration continues until convergence or until a limit on the number of iterations is reached. We use a limit of 50 iterations. The polish produces values for GE_k, A_{ijk}, $P_{k\ell}$, and $\epsilon_{ijk\ell}$. Only GE_k and A_{ijk} are needed for subsequent steps.

c) We typically compute Tukey's test ([8]) for non-additivity of array and probe effects for each probeset. However, we do not interpret the resulting p-values literally since this test is *not* resistant to outliers or non-Gaussian noise distributions. Instead, the p-values provide an index of non-additivity that can be used as a check on values for parameters such as a and c in the previous subsection.

4.4 Normalization

Although the the arrays were scaled to each other in subsection 4.1, that procedure probably does not adequately normalize the arrays. The magnitude of array specific biases may depend on expression level, so we seek a procedure that will make the arrays comparable across the entire range of expression. We have found the following procedure useful for this purpose.

a) Identify the set of genes, $\mathcal{K}$, with the least between-treatment-group variability. If the number of treatment groups is $M = 1$, use all genes: $\mathcal{K} = \{1, 2, \ldots, K\}$. Otherwise, for the gene represented by probeset k, define:

$med_{ik} = \text{median}(A_{i1k}, \ldots, A_{in_ik})$, the median array effect associated with probeset k for all n_i arrays in treatment group i. Define the between group variability, the within group variability and their truncated ratio by,

$$\begin{aligned} BV_k &= \sum_i |med_{ik}| \\ WV_k &= \sum_i \text{median}_j(|A_{ijk} - med_{ik}|), \\ RV_k &= BV_k / \max(WV_k, 0.01), \end{aligned}$$

respectively.

Partition the genes into four bins based on quartiles of GE_k, $k = 1, \ldots, K$. Within each bin b, choose the fraction *frac* of genes with the smallest between-within ratios; call this set of genes $\mathcal{K}_b$. Then $\mathcal{K}$ is the union of $\mathcal{K}_1$, $\mathcal{K}_2$, $\mathcal{K}_3$, and $\mathcal{K}_4$. For example if there are $K = 4000$ genes, each bin will have 1000 genes. If *frac* $= 0.50$ then $\mathcal{K}_b$ will consist of the 500 genes in bin b with the smallest values of RV_k, and $\mathcal{K}$ will consist of the $4 \times 500 = 2000$ genes in $\mathcal{K}_1$ through $\mathcal{K}_4$. (In practice the number of genes per bin might not be exactly $K/4$ and the number of selected genes per bin might not be exactly (*frac*)$\times K/4$ because of rounding and because of tied values of GE_k or RV_k.) Note that *frac* is a parameter of the algorithm. It should be set to a value no higher than the proportion of genes not expected to differ between treatment groups. We currently use *frac* $= .50$.

Define $GE_{(-)}$ to be the minimum value of GE_k over all genes in $\mathcal{K}$, and define $GE_{(+)}$ to be the maximum.

b) For each array ij, fit a smooth, nonparametric regression curve to A_{ijk} versus GE_k, using only the genes k in $\mathcal{K}$. The curve should be fit by an outlier-resistant method such as *loess*. Loess is described in sections 8.1.2 and 8.4.2 of [9]. Adjustable parameters for the loess smoother are its *span* (called α in [9] and *degree* (called λ). We have used *span*$= 0.5$ and *degree*$= 1$. We also use the S-Plus default of 4 iterations for the robustness part of the fitting procedure.

The result is a separate fitted curve $g_{ij}(GE)$ for each array ij.

c) Use function $g_{ij}()$ to produce normalized array effects for each ij:

$$NA_{ijk} = A_{ijk} - g_{ij}(GE'_k),$$

where

$$GE'_k = \begin{cases} GE_{(-)} & \text{if } GE_k < GE_{(-)} \\ GE_k & \text{if } GE_{(-)} \le GE_k \le GE_{(+)} \\ GE_{(+)} & \text{if } GE_k > GE_{(+)}. \end{cases}$$

GE_k is truncated to the interval from $GE_{(-)}$ to $GE_{(+)}$ to avoid extrapolating $g_{ij}()$ beyond the range of values used to estimate it.

d) We graph the normalization curves $g_{ij}(GE)$ as a function of GE, so that the magnitude and pattern of normalization adjustments for each array can be examined.

4.5 Estimate expression levels

The final estimated expression level for gene k on array ij is

$$\begin{aligned} X_{ijk} &= GE_k + NA_{ijk} \qquad \text{(transformed scale)} \\ X'_{ijk} &= f^{-1}(X_{ijk}) \qquad \text{(original intensity scale)}, \end{aligned}$$

where f^{-1} is the back transformation defined by

$$f^{-1}(x) = \begin{cases} a & \text{if } x < a \\ x & \text{if } a \le x \le c \\ c\exp(\frac{x-c}{c}) & \text{if } x > c. \end{cases}$$

a and c are the parameters of the hybrid transformation defined in subsection 4.2 above.

4.6 Summarize expression levels by treatment group and gene

Recall that n_i is number of replicate arrays for treatment group i, $i = 1, \ldots, M$. Define n_{ik} to be the number of replicate arrays in group i for which we have data for gene k; i.e., the number of arrays j for which X_{ijk} is not missing. The mean and standard error of expression level for group i (transformed scale) are:

$$\begin{aligned} \bar{X}_{ik} &= \frac{1}{n_{ik}} \sum_j X_{ijk} \\ s^2_{ik} &= \frac{1}{n_{ik}-1} \sum_j (X_{ijk} - \bar{X}_{ik})^2 \\ SE_{ik} &= \sqrt{s^2_{ik}/n_{ik}} \end{aligned}$$

When the degrees of freedom for estimating the standard error for each group is small, we often replace s^2_{ik} with the pooled within-group variance estimate,

$$s^2_{pk} = \frac{\sum_i (n_{ik}-1)s^2_{ik}}{\sum_i (n_{ik}-1)}.$$

The approximate mean and standard error of expression level on the original intensity scale are:

$$\bar{X}'_{ik} = f^{-1}(\bar{X}_{ik})$$

$$SE'_{ik} = \begin{cases} 0 & \text{if } \bar{X}'_{ik} < a \\ SE_{ik} & \text{if } a \le \bar{X}'_{ik} \le c \\ (\bar{X}'_{ik}/c)SE_{ik} & \text{if } \bar{X}'_{ik} > c. \end{cases}$$

4.7 Examine the relation of variability to expression level

a) Compute between-group and pooled within-group variances of expression level on the transformed scale. These are similar to BV_k and WV_k calculated in subsection 4.4, but use means and variances rather than medians and median absolute deviations. The formulas for the grand mean (GM), between group mean square (BG), and pooled within-group variance (s^2_{pk}) are

$$GM_k = \sum_i \sum_j X_{ijk} / \sum_i n_{ik}$$

$$BG_k = \frac{1}{M-1} \sum_i n_{ik} (\bar{X}_{ik} - GM_k)^2$$

b) As in subsection 4.4, partition the genes into four bins based on quartiles of GE_k (not GM_k). Within each bin select the fraction *frac* of genes with smallest between-group variability BG_k. Call the collection of selected genes from all bins $\mathcal{K}$. Fit a loess curve to $\sqrt{s_{pk}}$ (i.e., the fourth root of s^2_{pk}) versus GE_k for the genes in $\mathcal{K}$. (The reason for fitting the fourth root rather than s^2_{pk} or s_{pk} is that the distribution of the former is less skewed.) We have used loess parameters *span* = 0.33 and *degree* = 1. Let $h(GE)$ be the fitted curve.

c) Plot $\sqrt{s_{pk}}$ against GE_k for the genes in $\mathcal{K}$ and overlay the curve $h(GE)$. This graph shows whether and how the replicate variability of gene expression values depends on expression level.

4.8 Compare expression levels between groups

Equal variances of expression level (on the transformed scale) are assumed for all groups for a given

gene. The pooled estimate of the variance for gene k is s^2_{pk} from subsection 4.6.

a) A p-value for comparison of treatment groups i_1 and i_2 for gene k is obtained from a two-sided, two-sample t-test.

$$t = \frac{\bar{X}_{i_1 k} - \bar{X}_{i_2 k}}{s_{pk}\sqrt{\frac{1}{n_{i_1 k}} + \frac{1}{n_{i_2 k}}}}$$

Refer t to a t-distribution with $\sum_i (n_{ik} - 1)$ degrees of freedom to determine the p-value. Note that the test statistic t is calculated using estimates of means and variances on the hybrid transformed scale.

b) Calculate an expression ratio and log-ratio to compare groups on the original intensity scale. To avoid very large or infinite log-ratios (positive or negative), set a threshold d on the original intensity scale. Replace expression levels less than d by d when calculating ratios and log-ratios:

$$R_{i_1 i_2 k} = \frac{\max(\bar{X}'_{i_1 k}, d)}{\max(\bar{X}'_{i_2 k}, d)}$$

$$\text{log-ratio} = LR_{i_1 i_2 k} = \log_{10} R_{i_1 i_2 k}.$$

We have used $d = 1$ for this parameter of the algorithm.

Approximate standard errors for the log-ratios can be calculated using the following formulae.

$$LR_{i_1 i_2 k} = \log_{10}[\max(\bar{X}'_{i_1 k}, d)] - \log_{10}[\max(\bar{X}'_{i_2 k}, d)]$$

and $SE(LR_{i_1 i_2 k}) \approx$

$$\sqrt{\text{Var}\{\log_{10}[\max(\bar{X}'_{i_1 k}, d)]\} + \text{Var}\{\log_{10}[\max(\bar{X}'_{i_2 k}, d)]\}}.$$

If $d \le c$, the parameter for the hybrid transformation, then $\text{Var}\{\log_{10}[\max(\bar{X}'_{ik}, d)]\} \approx$

$$\begin{cases} 0 & \text{if } \bar{X}'_{ik} < \max(a, d) \\ \frac{1}{(\ln 10)^2} \frac{1}{\bar{X}'^2_{ik}} \frac{s^2_{pk}}{n_{ik}} & \text{if } \max(a, d) \le \bar{X}'_{ik} \le c \\ \frac{1}{(\ln 10)^2} \frac{1}{c^2} \frac{s^2_{pk}}{n_{ik}} & \text{if } \bar{X}'_{ik} > c. \end{cases}$$

Otherwise $\text{Var}\{\log_{10}[\max(\bar{X}'_{ik}, d)]\} \approx$

$$\begin{cases} 0 & \text{if } \bar{X}'_{ik} < c\ln(d/c) + c \\ \frac{1}{(\ln 10)^2} \frac{1}{c^2} \frac{s^2_{pk}}{n_{ik}} & \text{otherwise.} \end{cases}$$

5 Discussion and Conclusions

Microarrays hold great promise for helping researchers understand complex patterns of gene expression, but in many ways they are not different from other assays. Statistically sound methods for quantification of assay results are necessary. We have described an approach that emphasizes looking at all of the arrays in an experiment simultaneously. This allows resistant estimation of probe-specific effects. Moreover, we suggest that the assessment of repeatability should not ignore sources of variability that are likely to be substantial, in particular between-EU variability. We have given some details of our implementation of this approach.

References

[1] Lockhart, D., Dong, H. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology*, 1996, **14**, pp. 1675–1680.

[2] Wodicka, L., Dong, H., Mittmann, M., Ho, M.H., Lockhart, D.J. *Saccharomyces cerevisiae*, *Nature Biotechnology*, 1997, **15**, pp. 1359–1366.

[3] Li, Cheng, Wong, W.H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proceedings of the National Academy of Science*, 2001, **98**,1 ,pp. 31–36.

[4] Tukey, J.W. *Exploratory Data Analysis*, Addison-Wesley Publishing Company, Reading, Mass., 1977, p. 363–378.

[5] Emerson, John D. and Hoaglin, David C. Analysis of two-way tables by medians. In *Understanding Robust and Exploratory Data Analysis*, David C. Hoaglin, Frederick Mosteller, John W. Tukey, eds. John Wiley & Sons, New York, 1983, pp. 166–210.

[6] Tusher, V., Tibshirani, R,. Chu, G., Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Science*, 2001, **98**,9 ,pp. 5116–5121.

[7] Insightful, Inc. (1988–2001) *SPLUS 6 Software and documentation*, Seattle, WA: Insightful.

[8] Tukey, J.W. One degree of freedom test for non-additivity, *Biometrics*, **5**, 1949, pp. 232–242.

[9] Cleveland, William S.; Grosse, Eric; Shyu, William M. Local regression models. In *Statistical Models in S*, John M. Chambers and Trevor J. Hastie, eds. Wadsworth & Brooks/Cole, Pacific Grove, CA, 1992, pp. 309–376.