# Joint Estimation of Calibration and Expression for High-Density Oligonucleotide Arrays

Ann L. Oberg*, Douglas W. Mahoney, Karla V. Ballman, Terry M. Therneau
Department of Health Sciences Research, Division of Biostatistics
Mayo Clinic College of Medicine
Rochester, MN 55905
U.S.A.

February 2, 2006

Abstract

Motivation: The need for normalization in microarray experiments has been well documented in the literature. Currently, most analysis methods treat normalization and analysis as a series of steps, with summarized data carried forward to the next step.

Results: We present a unified algorithm which incorporates normalization and class comparison in one analysis using probe level perfect match and mismatch data. The algorithm is based on calibration models common to most biological assays, and the resulting chip-specific parameters have a natural interpretation. We show that the algorithm fits into the statistical generalized linear models framework, describe a practical fitting strategy and

*To whom correspondence should be addressed

present results of the algorithm based on metrics used in affycomp [6, 10]. The algorithm ranks amongst the top third of the affycomp competitors, performing best in measures of bias.
Availability: R functions are available on request from the authors.
Contact: oberg.ann@mayo.edu

## Key Words

Normalization; Calibration; High-density oligonucleotide arrays; Generalized linear models

# 1 Introduction

Microarray technology provides researchers with a powerful tool to measure expression levels of thousands of genes in a specimen sample simultaneously. Due to the costly nature of these studies and sometimes sample scarcity, an experiment typically yields results for only a few samples. In addition, it is usually feasible to pursue further research regarding only a small subset of the genes on the array. Hence, it is important to make efficient use of the data in order to distinguish biological variation from random error.

Currently, most methods of class comparison involve a separate normalization step and comparison step. The purpose of the normalization step is to remove systematic sources of variability while preserving the biologic variation of interest. The simplest of the normalization methods (applied by Affymetrix MAS software [1] and others) involves setting the overall mean of each chip equal to the same number. (The microarray laboratory in our institution for instance, uses the value of 1500.) This assumes a linear relationship between the true expression level and the fluorescent intensity actually observed over the entire range of gene expression values. There is compelling evidence that this relationship is in fact nonlinear [4, 7, 11, 14], and several more sophisticated normalization methods have been proposed. The normalized probe-level data are then typically summarized into a per-gene estimate of expression. Class comparisons and other analyses are then performed on the data at this level.

We show here that the massive number of probes on a high-density oligonucleotide array chip actually allows the construction of an *indirect* calibration curve using the available data in an unsummarized form, and that these curves have several desirable features. We present an algorithm that

unifies the normalization and class comparison steps into one analysis, show how it fits into the statistical class of generalized linear models of McCullagh and Nelder [12] and describe a practical fitting strategy. Section 2 describes the data sets used for illustrative purposes. Section 3 motivates and introduces the concept of calibration and describes its application here. Section 4 describes the algorithm and its implementation and Section 5 describes the performance of the method based on the benchmark of Affymetrix GeneChip expression measures (Affycomp) of Cope et. al. [6, 10]. Conclusions and discussion are presented in Section 6.

# 2 Data

The properties of the proposed algorithm were evaluated on data from publicly available experiments. These data sets are briefly described in this section and are available from the Affymetrix web site http://www.affymetrix.com or on the affycomp web site http://affycomp.jhsph.edu.

## 2.1 Affymetrix U95A spike-in data

The Affymetrix U95A spike-in data set has sixteen genes which were spiked in varying concentrations into a pancreas background in concentrations ranging from 0 to 1024 picomolar (pM) and hybridized in a cyclic Latin Square design onto Hu95A chips. There were at least 3 chips per concentration combination. In this data set, there are only 16 genes expected to show fold changes in concentration expression. All other 12,610 genes should be identically expressed on all arrays. A more complete description can be found in [6, 10] or via a search for "Latin square data" on the Affymetrix web site.

### 2.2 Affymetrix U133A spike-in data

The Affymetrix U133A spike-in data set has forty-two genes spiked into a human HeLa cell line background in a cyclic Latin square fashion and hybridized onto Hu133A chips. The concentrations range from 0 to 512 pM and sample was hybridized to 42 chips. In this data set, there are only 42 genes expected to show fold changes in concentration expression. The other 22,258 genes should be identically expressed on all arrays. More details can be found via a search for "Latin square data" on the Affymetrix web site or in [6, 10].

## 3 Calibration

In simpler, low volume assays, it is common to normalize the data by use of a *direct calibration* curve $f$

$$y = f(x) + \epsilon$$

where $y$ is the observed data from the assay, $x$ is the true concentration, and $\epsilon$ is random error. In a 96 well ELISA assay, for instance, $f$ is directly estimated by using reference samples of known concentration in one row of wells. The fitted curve is used to recover estimates of the true values $x$ from the data $y$.

Microarray data does not typically contain the information needed for a direct calibration. In particular, it is not clear what could be used as reference targets that would fairly represent the wide variety of probes on the array. (See the Gene Logic technical note entitled "Optimization of an external standard for the normalization of Affymetrix GeneChip arrays" available from the Gene Logic web site http://www.genelogic.com for one approach

to this issue, however). Here, we focus on an *indirect* calibration function applied to an array as a whole accounting for the largest effects, with future work dedicated to accounting for probe-specific background effects.

Our choice of the calibration functional form is driven by the general principle that for any biological assay that spans a wide range, an S-shaped curve between the true (unmeasured) value and the observed assayed value is almost a guarantee. A lower threshold on the observed data can be due to background binding, lower limits of detection for the instrumentation, or other causes (the mismatch (MM) probes are actually designed to estimate this), while an upper limit may be due to either biochemical or instrumentation saturation. Given the wide range of gene expression values, from around 10 to 46,000 seen on the Affymetrix platform, this S-shape relationship is likely to be true.

Furthermore, for choosing the shape of the curve, we followed the review article of Finney [8] which describes analytical approaches to the binding problem in radioligand assay. He states that for most problems, a logistic or probit function fit to the log of the true value (on the horizontal axis) versus the log of the observed value is sufficient. A further advantage of this is that the data in question is approximately equivariant on the log scale [3], making both plots and analyses more straightforward.

Hekstra [9] and Burden [5] have shown the the Langmuir isotherm

$$2^y = a + b\frac{2^x}{2^x + 2^K}$$

is an appropriate equation for the binding curves on theoretical grounds, and that it successfully fits several data sets. As elsewhere in this paper, $y$ and $x$ are on the $\log_2$ scale. The parameters $a$ and $b$ are scaling factors per chip. The per-probe constant $K$ depends on the binding efficiency of each

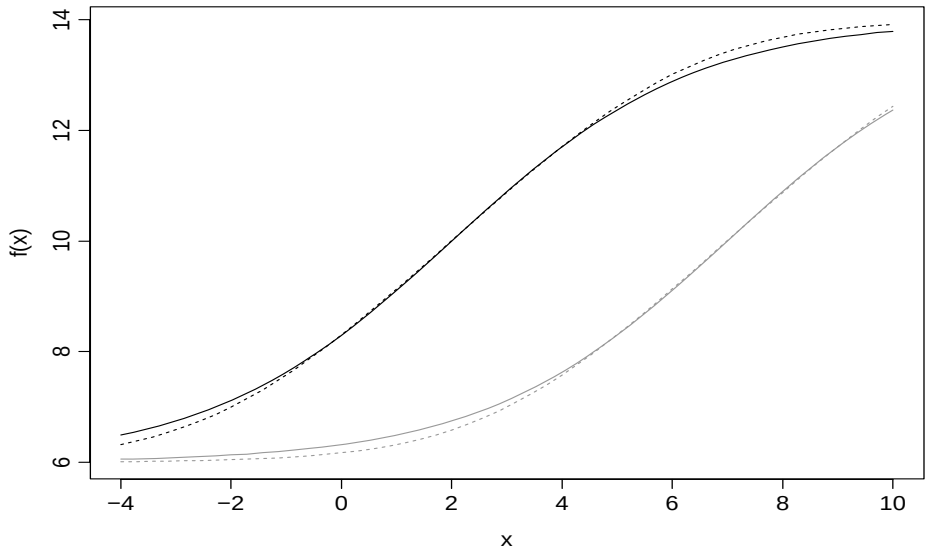


Figure 1: Two Langmuir isotherm curves (dashed lines) over the concentration and fluorescence range found in the U95 spike-in data. Matching logistic curves (solid lines) are overlayed in order to show the similarities between the two curves.

probe.

Figure 1 contains two Langmuir isotherms over the concentration and fluorescence range found in the U95 spike-in data with matching logistic curves overlayed. The Langmuir and logistic curves are nearly identical in each case, and we have chosen to retain the logistic fit for this report due to its statistical familiarity. For the Langmuir form, $a$ and $b$ control the upper and lower thresholds, and $\log_2(K)$ translates the curve left-right. For the logistic form $y = \gamma_1 + \gamma_2 \exp(\eta)/[1 + \exp(\eta)]$, $\eta = \gamma_3(x - K)$, $\gamma_1$ and $\gamma_2$ again control the upper and lower limits, the location parameter $K$ corresponds to a per probe binding constant, and $\gamma_3 = 1/2$ causes the curve to have slope

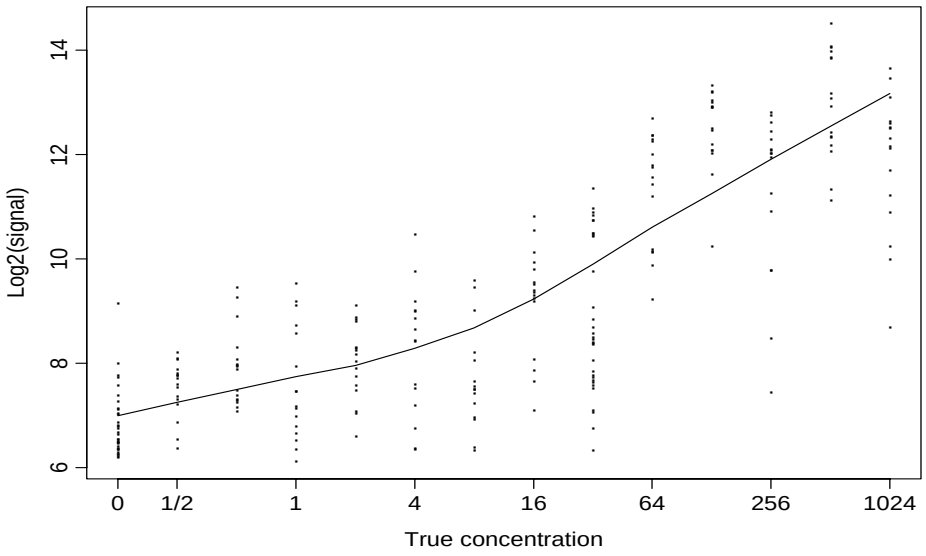


Figure 2: True concentration in pM versus observed intensity for chip 1 of the Affymetrix U95 spike-in experiment, along with a fitted lowess curve. Genes spiked in at concentration 0 are plotted at 1/2 of the next largest concentration, but labeled as 0 concentration.

1 at the midpoint; this is what is shown in the figure.

The importance of a per-probe binding constant is shown in Figures 2 and 3, both of which show the data for chip 1 of the U95 experiment. Figure 2 shows the raw fluorescence data versus the spiked in values, equivalent to assuming $K = 0$ for all probes, along with a lowess smooth; Wu and Irizarry [15] pursue this in more detail. In contrast, Figure 3 shows the result with probe-specific values for $K$; based on a fit to all 59 chips and 256 PM/MM probe pairs of the U95 spike-in with $x=$ the true spike-in amount. The fit is much tighter, and in particular the logistic (Langmuir) form of the true calibration curve is much more clear.

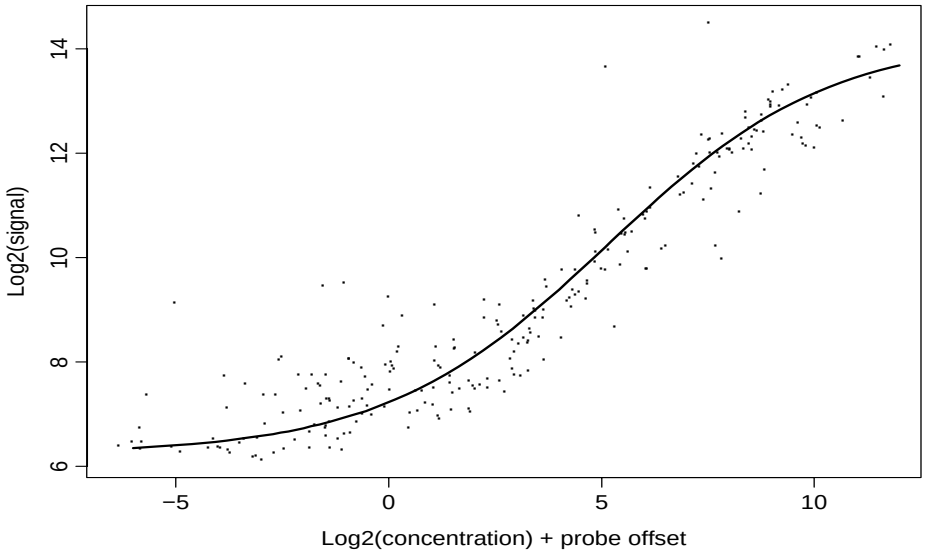


Figure 3: True PM concentration plus probe effect (i.e., $p_{ij}$) in pM versus observed intensity for chip 1 of the Affymetrix U95 spike-in experiment, along with a fitted logistic curve allowing for a probe effect in the model. Genes spiked in at concentration 0 are plotted at 1/2 of the next largest concentration, but labeled as 0 concentration.

# 4 Implementation

## 4.1 An Algorithm

A goal of many microarray experiments is to evaluate differences in gene expression between experimental groups. The normalization function is necessary, but is not of interest in and of itself. Our basic calibration model is

$$y_{ijk} = f_k(\eta_{ijk}) + \epsilon$$

where $y_{ijk}$ is the observed intensity value on the $\log_2$ scale for the $j$th probe associated with the $i$th gene on the $k$th chip, $f_k$ is the chip specific calibration function for the $k$th chip, and $\eta_{ijk} = g_{ik} + p_{ij}$ where $g_{ik}$ is the expression of the $i$th gene on the $k$th chip and $p_{ij}$ is the binding efficiency of the $j$th probe associated with the $i$th gene. The primary parameters of interest are the gene effects $g_{ik}$, secondarily the calibration functions $f_k$ themselves, and the probe effects $p_{ij}$ are ancillary.

Note that if the chip functions $f_k$ were known, then this is the estimation of a generalized linear model [12] with $f^{-1}$ as the link function and $\eta_{ijk}$ as the linear predictor. This opens up a whole set of well developed tools for estimation and hypothesis testing, e.g., contrast statements.

It is not feasible to fit all the parameters at once, however, and we use a simple iterative algorithm where each step is described in more detail in subsequent sections:

0. Create initial estimates $\hat{f}_k$ of $f_k$.

1. Solve for $p_{ij}$ and $g_{ik}$, given $\hat{f}_k$ and $y_{ijk}$, creating $\hat{p}_{ij}$ and $\hat{g}_{ik}$.
   Notice that this step can be done separately for each gene $i$.

2. Solve for $f_k$, given $\hat{p}_{ij}$, $\hat{g}_{ik}$, and $y_{ijk}$, creating an updated $\hat{f}_k$.
   Notice that this step can be done separately for each chip $k$.

3. Iterate steps 1 and 2 a fixed number of times, or until convergence.

### 4.2 Calibration Function, $f_k$

We parameterize the logistic calibration function for the $k$th chip as

$$f_k(\eta_{ijk}) = \gamma_{1k} + \gamma_{2k} \frac{e^{\gamma_{3k}(\eta_{ijk} - \gamma_{4k})}}{1 + e^{\gamma_{3k}(\eta_{ijk} - \gamma_{4k})}}.$$

The parameters are the lower threshold $\gamma_{1k}$, the upper threshold $\gamma_{1k} + \gamma_{2k}$, the inflection point $\gamma_{4k}$ and the slope $\gamma_{3k}$. The chip inflection parameter $\gamma_{4k}$ is completely confounded with the per-chip gene expression $g_{ik}$. This confounding corresponds to an obvious physical interdigitation: if each signal on chip 2 were 20% larger than the corresponding signal on chip 1, it is not possible to say, without outside information or assumptions, whether this is a chip effect (such as a different sample handling method or different scanner) or that all gene products on the second chip are actually more highly expressed (such as in a dilution experiment). The first corresponds to $\gamma_{41} > \gamma_{42}$ and $g_{i1} = g_{i2}$ for all $i$, the second to $\gamma_{41} = \gamma_{42}$ and $g_{i1} > g_{i2}$ for all $i$.

For computation, we will assume without loss of generality that $\gamma_{4k} = 0$ for all $k$, i.e., set the chip effect to 0. After the fit, this can be adjusted based on chosen constraints, for instance rescaling so that all of the gene effects $g_{ik}$ have mean 0 within a chip will credit all systematic variations to chip effects rather than to the experimental unit that was hybridized to that chip. This is likely what will be done most often in practice.

The other non-identifiability is between gene effects and probes within gene: $\eta_{ijk} = g_{ik} + p_{ij} = (g_{ik} + c) + (p_{ij} - c)$ for any constant $c$. We have

chosen to make the perfect match (PM) probe effects within each gene sum to zero. Any comparison of the absolute values of $g_{ik}$ between genes would require the assumption that the average binding efficiency of their two probe sets is identical.

The MM data are used in addition to the PM data in this algorithm. A wide range of $p_{ij}$s facilitates the fitting process, and the MM data help to ensure a wide range. Hence, the information added by the MM data is key to the success of this algorithm. In fact, the algorithm has substantially more difficulty recovering the true calibration curve without the MM data when evaluated in simulations (data not shown).

### 4.3 Initial estimates

As an initial estimate of the lower threshold of the calibration function, we use a percentile of the MM data. If the MM probes measure only background binding, as was purposed in their design, then we would expect an average MM probe to be close to the true lower threshold. In actuality, some of them will have activity [3], so a percentile between the 20th and 30th is used. For the upper threshold, we use the 99th percentile of the PM values since the 100th percentile represents saturated housekeeping genes.

An initial value for the slope $\gamma_{3k}$ can be set to 1 since this only controls the scale of the primary parameters $g_{ik}$ and $p_{ij}$. As stated above, the offset parameter $\gamma_{4k}$ is set to 0.

### 4.4 Estimation of gene and probe effects given $f_k$

Since this estimation will be done on a very large number of genes, the calculations should be as fast as possible. Calls to a general non-linear

estimation would be slow. The iterative weighted least squares (IWLS) approach pioneered in generalized linear models [12] allows fast and easy estimation of the logistic parameters. For a given gene $i$, we have

$$\begin{aligned} y_{ijk} &= f_k(\eta_{ijk}) \\ &= f_k(g_{ik} + p_{ij}) \\ &= f_k(X\beta)\,. \end{aligned} \tag{1}$$

Suppressing the subscript $i$, we see that the design matrix $X$ is the same as that for a classic two-way balanced analysis of variance. The index $j$ ranges from 1 to twice the number of probe pairs for an Affymetrix array (both PM and MM data are used), and $k$ from 1 to the number of chips. Given starting estimates for the lower and upper threshold parameters $\gamma_{1k}$ and $\gamma_{1k}+\gamma_{2k}$, the IWLS update is the solution to a weighted regression with a working dependent variable $z$ and case weights $w_{ijk}$, where

$$\begin{aligned} z_{ijk} &= \hat{\eta}_{ijk} + [y_{ijk} - f_k(\hat{\eta}_{ijk})]/d_{ijk} \\ d_{ijk} &= \frac{\partial f_k(\eta_{ijk})}{\partial \eta_{ijk}} \\ w_{ijk} &= d_{ijk}^2/V(\hat{\eta}_{ijk}) \end{aligned}$$

where the derivative in $d_{ijk}$ is evaluated at $\hat{\eta}_{ijk}$ and $V(\hat{\eta}_{ijk})$ is the variance of the $y_{ijk}$ vector evaluated at the current value of the linear predictor $\hat{\eta}_{ijk} = x_{ijk}\hat{\beta}$.

Since we are using the $\log_2$(intensity) data *without* background subtraction, the variance is approximately constant, i.e., $V(\hat{\eta}_{ijk}) = V$ (this is evident in Figure 3). Ballman et. al. [3] examined plots for all $1,508$ probes for the spiked in genes of the U95 and U133 spike in experiments and reached the same conclusion. This provides a key simplification in the estimation

procedure. Background subtraction makes no essential difference in the formulation of Equation (1), as it corresponds more or less to subtracting a constant from $\gamma_1$; but it would add a major complication to the variance function $V(\hat{\eta})$ in that it is no longer constant.

With $V$ a constant and using the logistic formulation for $f_k$, the update formula simplifies to a regression of $y_{ijk} + d_{ijk}\eta_{ijk} - f_k(\eta_{ijk})$ on $DX$, where $D$ is a diagonal matrix with elements $d_{ijk}$. The weights $d_{ijk}$ serve to down-weight points in the far left or right tails of the function. Notice that if we were to assume $d = 1$, then this becomes a bias correction algorithm similar to fastlo [2], but with a more complex linear model (both gene and probe effects rather than the simple probe mean) and a logistic mean function rather than a nonparametric smooth.

# 5 Results

The calibration algorithm was applied to the data sets described in Section 2. The resulting gene estimates (i.e., the $g_{ik}$), centered and scaled to attribute the systematic effects to the chip, were submitted to affycomp [6, 10]. The full report can be found on the affycomp web site http://affycomp.jhsph.edu/ under method name `chipcal4` on the new assessment link.

Figure 4 shows the estimated calibration curve for chip 1 of the U95 data from the proposed method, along with the best possible estimate created from a fit where the true concentrations are known (i.e. the model fit to the spike in genes only). The two curves are surprisingly similar.

At the time of this writing, competition results were available on the new assessment link of the affycomp web site for 48 methods in the U95 table and

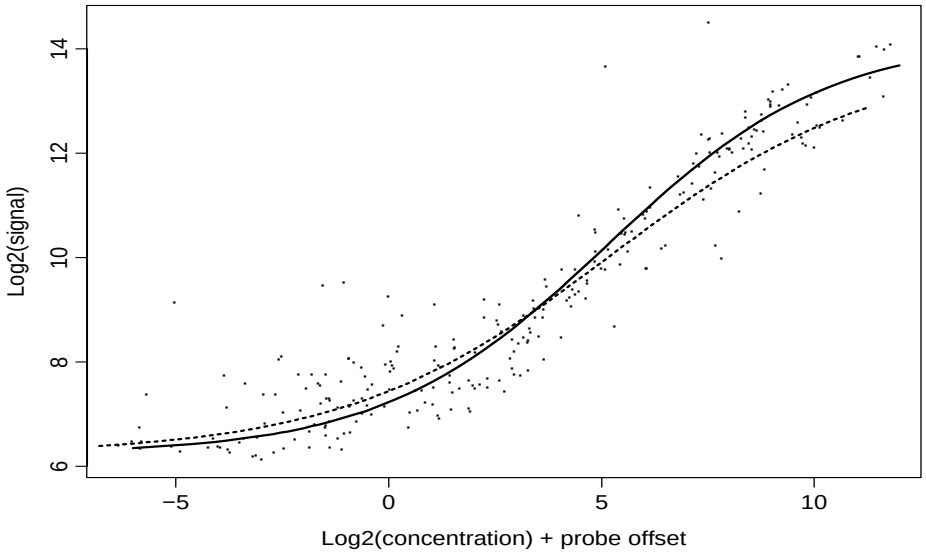


Figure 4: Estimated true calibration curve for chip 1 of the U95 experiment (solid line), from a model where the true concentration is known (i.e. fit to the spike in genes only) and the probe specific binding constants (left/right offsets) are estimated. The dashed line is the estimated binding curve from the proposed algorithm, where true concentrations are not known.

42 unique methods for the U133 table. In the 14-dimensional score displayed for all submitted methods our calibration method has average rank (based on absolute distance from the ideal value) of 21 for the U95 and 19 for the U133 result tables, receiving the highest ranks in the items measuring bias and the lowest ranks in the items measuring variance. When the methods are ranked based on their average score ranking, this algorithm places 13th and 14th for the U95 and U133 data, respectively.

A plot of observed expression intensity versus true expression intensity is displayed in figure 4a of the reports. The slopes from a regression line fitted to these data (observed versus nominal expression values) are 0.94 and 0.95 for the U95 and U133 data, respectively. A value of 1 would indicate no bias. While these slopes show that there is bias towards zero, this algorithm ranks 4th and 6th on this measure in the U95 and U133 data sets, respectively, indicating that this method has less bias than most. It is our thought that the excellent performance in this area may be due to the fact that we are, at least approximately, capturing the correct functional form of the calibration curve directly in the model.

The Achilles heel of the current algorithm lies in variance. Figure 2b for both the U95 and U133 reports indicate that variability for the non-spiked in genes increases at both the lowest and highest expression values, with the median standard deviation ranking in the top third for both data sets. The $R^2$ values from the regression lines fit to the observed versus nominal expression values of 0.70 and 0.82 for the U95 and U133 fits rank near the bottom of the list. When signal detection is based on the criteria of a fold change greater than 2, the algorithm performs reasonably well for medium and high fold changes with ROC AUC values of 0.80 to 0.77 for the U95 data and 0.76 and 0.96 for the U133 data. However, for low expression genes,

the AUC is below 0.5 at 0.36 and 0.44 in the U95 and U133 experiments, respectively, indicating a large number of false positives. Further exploration of the false positives, however, revealed that a large fraction of them are due to a single aberrant probe which had an overly large influence on the two-way ANOVA of Section 4.4 and that they are located on the horizontally flat portions of the logistic curve.

Going beyond the Affycomp results, we explored the behavior of contrasts for both the spike in and null genes, using the six chips in experiments 1 and 2 of the U133 experiment. For each gene $i$, the fit gives 6 gene effect estimates $\hat{g}_{ik}$ for which $k = 1, 2, 3$ represent the first spike-in pattern and $k = 4, 5, 6$ the second. The variance matrix and residual standard error of the fit can be used to test whether $c'g = 0$ for these six chips where $c = (1, 1, 1, -1, -1, -1, 0, \ldots, 0)$ and $g = (g_{i1}, g_{i2}, g_{i3}, g_{i4}, g_{i5}, g_{i6}, \ldots, g_{i42})$, a contrast which incorporates both gene and probe variability. We compare these to results from RMA which are based on only the 6 summary expression values. Similarly, we explored the behavior of these contrasts in the U95 data utilizing the two spike in combinations having twelve replicates each.

The two chosen spike in combinations in both data sets are such that all but two of the spike in genes have a nominal fold change of 2, with the other two spike in genes having fold changes of 0 versus 0.125 or 512 versus 0 for the U133 data and 0 versus 1024 or 0.25 versus 0 for the U95 data. The remaining genes on the chip have a nominal fold change of zero. Contrast estimates and confidence intervals for the genes having fold change of two from both data sets are displayed in Figure 5 for both our algorithm and RMA. Ideally the contrast estimates would be centered at one (the plots are on the $\log_2$ scale) and the confidence intervals would not include

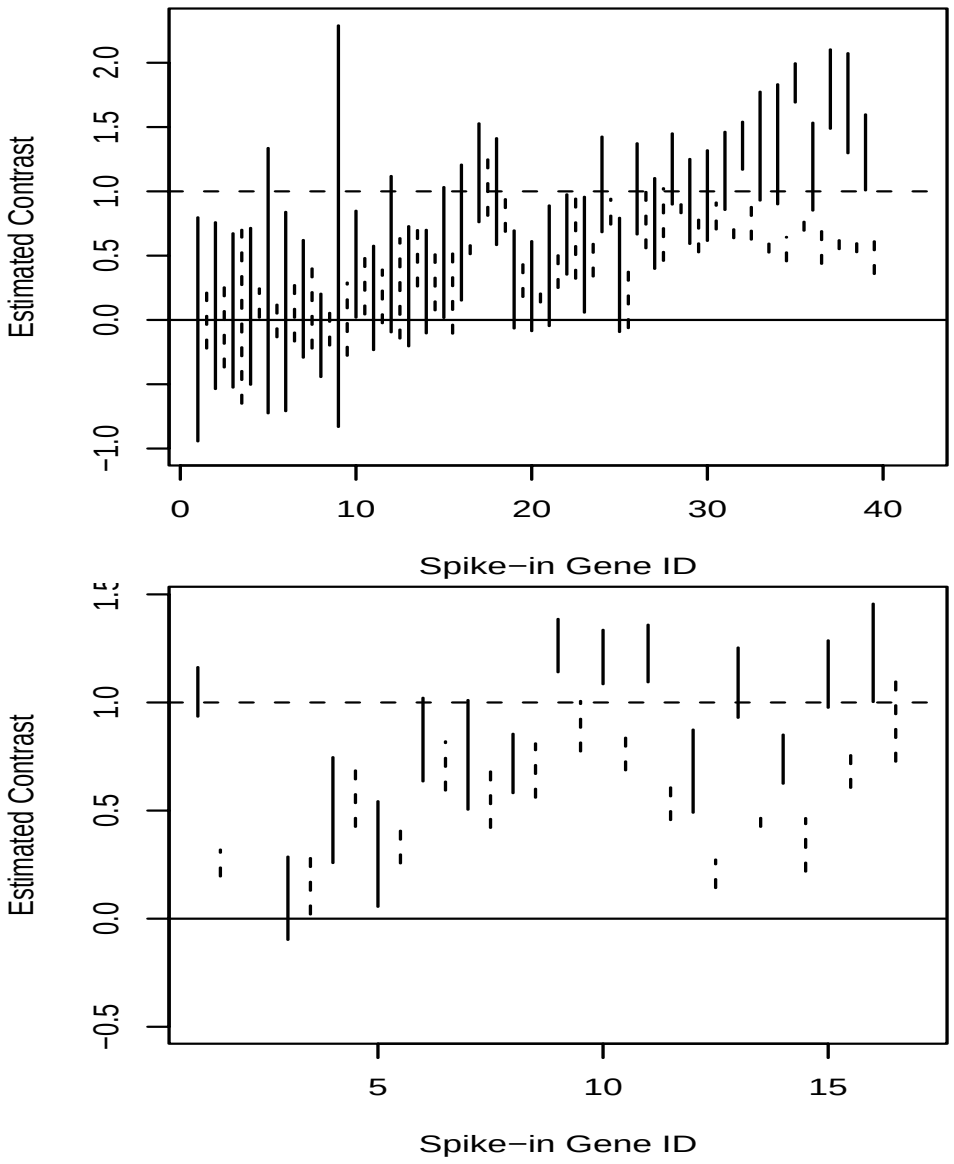


Figure 5: Contrast estimates with associated confidence intervals for the comparison of spike in genes between experiments 1 and 2 of the U133 experiment (top panel) and between the two spike in combinations having twelve replicates each in the U95 data (bottom panel). All have expected fold change of 2. Results are shown for this algorithm (solid lines) and RMA (dashed lines), and are sorted in order of total nominal abundance with lowest total nominal abundance on the left.

zero. Both algorithms have greater difficulty in detecting true changes in the lower abundance range than in the high abundance range. Performance in the mid abundance range is better in the U95 than in the U133 data. Twenty-five of the 42 fold changes in the U133 data are significant at the 5% level with this algorithm while RMA declares 30 significant (24 of them are in common). There are eleven genes which neither algorithm declares significantly differentially expressed. Fifteen of the sixteen fold changes in the U95 data are significant at the 5% level with this algorithm while RMA declares 16 significant. This algorithm declares 6.3% and 19.4% of the null genes significantly differentially expressed at the 5% level in the U133 and U95 data, respectively, while RMA declares 5.7% and 4.3% significant. Scatter plots of the distributions of the null gene test statistics for this algorithm versus the RMA algorithm are displayed in Figure 6. These results indicate that both algorithms do a reasonable job of controlling type I error in the U133 data. The current algorithm is not controlling this well in the U95 data. Inspection revealed that, as for signal detection via fold change, a single probe outlier was typically the cause of the false positives, indicating that robust estimation methods will likely improve the control of type I error. As would be expected, the statistical criterion for significance performs much better than the fold change criterion which does not account for variability present in the data.

## 6 Discussion

We have presented an algorithm for normalization and analysis of high-density oligonucleotide microarray data. Most microarray normalization routines proposed thus far are based on ad hoc methods. The current algo-

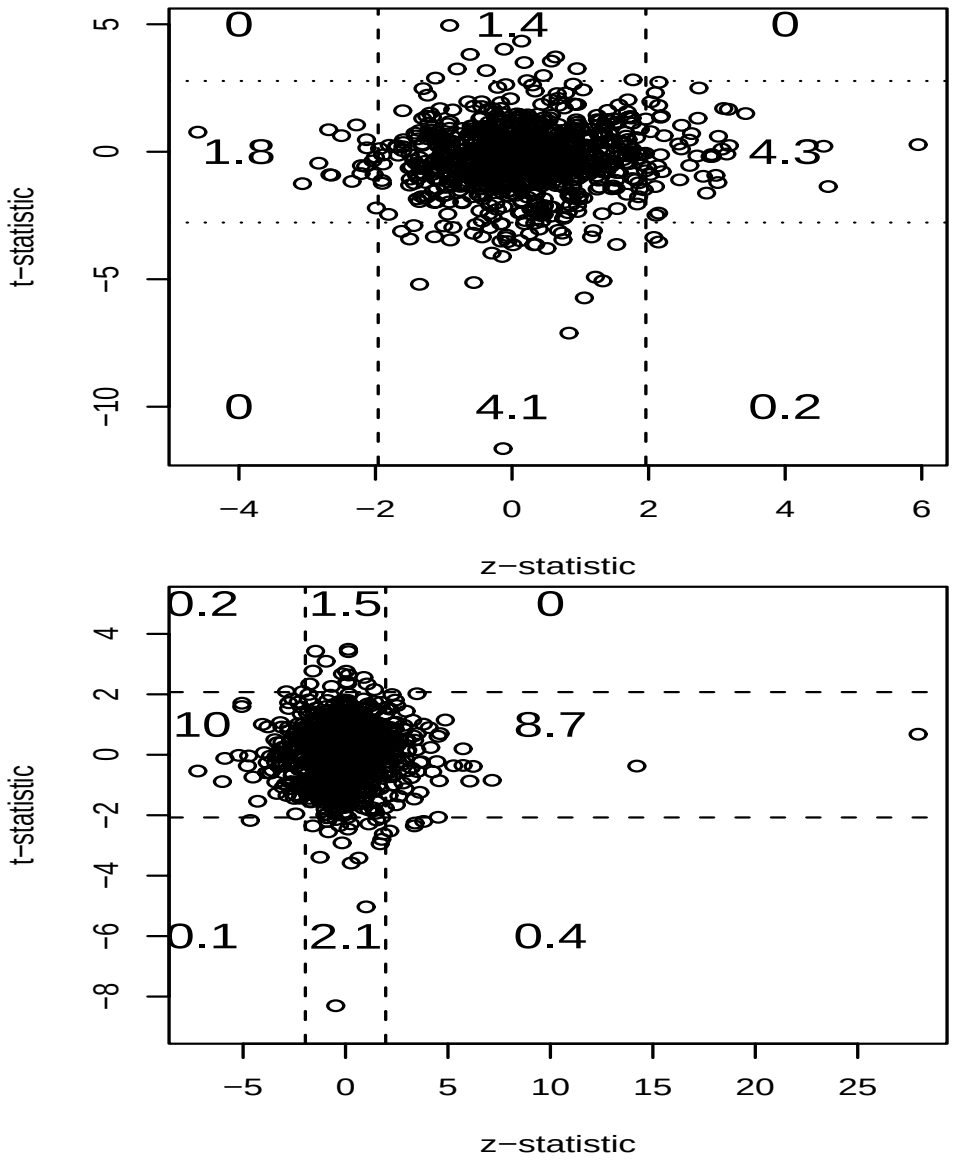


Figure 6: Distribution of null gene test statistics for the calibration algorithm (x-axis) versus RMA (y-axis) for the U133 data (top panel) and the U95 data (bottom panel). Dashed lines indicate significance at the 5% level for a z-statistic for calibration and for a $t$ distribution with 4 degrees of freedom for the U133 data or 22 degrees of freedom for the U95 data for RMA. Numbers indicate the false positive rate (%) within each region of the plot.

rithm, cast into the generalized linear model framework, brings a statistical framework to bear on the normalization problem and paves the way for a large set of well established methods to be easily applied in this situation. In addition, the generalized linear model framework allows one to conduct hypothesis tests based on variability at the probe level, accounting for all sources of variation present in an experiment.

The algorithm incorporates biological principles as well as efficient classical statistical techniques. It uses probe level PM and MM data as the input and results in a chip specific normalization or indirect calibration function. The logistic function used to model this is based on calibration models common to most biological assays and the parameters have a natural interpretation. Parameter estimates from these calibration functions could eventually be used for quality control purposes once their behavior is better understood. The algorithm is simple to implement using standard software. R functions that implement this algorithm are available from the authors.

The algorithm estimates the true calibration functions well. Results from the affycomp competition show that on average, this algorithm is amongst the top third of affycomp competitors based on average rank in the 14-dimensional score and performs well with respect to bias. There is room for improvement with respect to variance and signal detection. Including the MM probes in the modeling process is likely adding variability to the estimates, in a sense acting like background subtraction. However, it is well known that MM probes measure signal, and, without the MMs the algorithm has mediocre performance. In addition, knowledge of probe-specific non-specific binding (such as GC content) has not yet been incorporated into this algorithm. Reports have shown that incorporating this knowledge into normalization routines improves the variance at the lower expression

levels [13, 16]. Work is in progress to incorporate probe specific background binding information into the algorithm.

Regarding signal detection, the affycomp results indicate that this algorithm produces a large number of false positives based on the criteria of fold change greater than two, and the statistical contrast results demonstrate that type I error is not well controlled in the U95 data. Inspection of the null (non-spiked in) genes (i.e., those with expected fold change of zero) which were given large fold changes by the algorithm reveals that these particular genes are those that fall on the horizontally flat portions of the logistic curve and the large fold changes are due to one or two outliers. This suggests that using a robust methods would improve the algorithm with respect to this metric.

## Acknowledgment

Partial funding for A.L.O. was provided by the Fraternal Order of Eagles Cancer Research Fund.

## References

[1] Affymetrix. Microarray suite user guide, version 5. Technical Report http://www.affymetrix.com/support/technical/manuals.affx, Affymetrix, Inc., 2001.

[2] K. V. Ballman, D. E. Grill, A. L. Oberg, and T. M. Therneau. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*, 20(16):2778–2786, 2004.

[3] K. V. Ballman and T. M. Therneau. An exploration of affymetrix probe-set intensities in spike-in experiments. Technical Report No. 74, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, 2005.

[4] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[5] C. J. Burden, Y. E. Pittelkow, and S. R. Wilson. Statistical analysis of adsorption models for oligonucleotide microarrays. *Statistical Applications in Genetic and Molecular Biology*, 3:No. 1, Article 35, 2004.

[6] L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, and T. P. Speed. A benchmark for affymetrix genechip expression measures. *Bioinformatics*, 20:323–331, 2004.

[7] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying genes with differential expression in replicated cdna microarray experiments. *Statistica Sinica*, 12(1):111–139, 2002.

[8] D. J. Finney. Radioligand assay. *Biometrics*, 32:721–740, 1976.

[9] D. Hekstra, A. R. Taussig, M. Magnasco, and F. Naef. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Research*, 31:1962–1968, 2003.

[10] R. A. Irizarry, Z. Wu, and H. A. Jaffee. Comparison of affymetrix genechip expression measures. Technical Report Working Paper 86, Johns Hopkins University, Department of Biostatistics Working Papers, 2005.

[11] T. B. Kepler, L. Crosby, and K. T. Morgan. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology*, 3(7)::research0037.1–0037.12, 2002.

[12] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1983.

[13] F. Naef and M. O. Magnasco. Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(1 Pt 1):011906:011906–1 – 011906–4, 2003.

[14] F. Naef, N. D. Socci, and M. Magnasco. A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations. *Bioinformatics*, 19:178–184, 2003.

[15] Z. Wu and R. Irizarry. Stochastic models inspired by hybridization theory for short oligonucleotide arrays [extended abstract]. San Diego, CA, March 27-31, 2004. RECOMB, Copyright 2004 ACM 1-58113-755-9/04/0003.

[16] Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *JASA*, 99:909–917, 2004.