

# An Exploration of Affymetrix Probe-Set Intensities in Spike-In Experiments

Karla V. Ballman  
Terry M. Therneau

Technical Report #74

July 2005

Copyright 2005 Mayo Foundation

# 1 Introduction

Yogi Berra once said, “You can observe a lot just by watching.” This sentiment reflects a basic tenet of exploratory data analysis: a lot can be learned by looking at the data. In general, effective data analysis proceeds iteratively; the analyst looks at the result of one step and then directs the next step accordingly. If the analyst begins without an *a priori* hypothesis, the initial look at the raw data is to determine whether there is apparent structure. If the analyst does have an *a priori* hypothesis, the initial look at the data can confirm or disprove an anticipated structure or relationship. Results of this first step then serves as the basis for ensuing analyses such a fitting a model and assessing the goodness of its fit by examining the model residuals.

Many models have been proposed for generating gene expression values from Affymetrix probesets. These range from robust averages [1] to multiplicative models [2] to linear models [3, 4] to models that incorporate binding information based on probe sequence content [5]. To determine an appropriate model for gene expression values as a function of probe intensity values, a good start is to look at data from known datasets. Specifically, what is the nature of the relationship of the amount of target RNA and observed fluorescence intensity for the three publicly available spike-in datasets using the Affymetrix (human) GeneChip platform? These are Gene Logic U95A (GL), Affymetrix U95A (Affy95), and Affymetrix U133A (Affy133)? In each dataset, a set of RNA targets was spiked-in at known concentrations. The goal of our study was to gain an understanding of the nature of the relationship between the actual amount of mRNA in a sample and the observed intensity level produced by the microarray platform as well as variation about the model. Our hypothesis was that the observed intensities of probe level data will follow a simple model for general binding experiments suggested many years ago by Finney [6].

We begin with brief descriptions of the Affymetrix spike-in datasets. This is followed with a discussion of the model proposed by Finney [6] to describe the binding behavior of most biological assays and a description of how these models were fit to the spike-in data. Next, plots of the probesets by concentration level with the fitted curves superimposed are presented. Observations based on the plots as well as the results of the fits are presented in Section 4. We close with a discussion of the implications of our observations with respect to appropriate models for gene expression intensities as a function of the probe intensities.

## 2 Descriptions of the data

### 2.1 Affymetrix U95A spike-in data

This dataset was created by Affymetrix and is publicly available at their web site [www.affymetrix.com](http://www.affymetrix.com); search on the phrase “Latin square data” to find the link to the page containing a description of the experiment and the downloadable files of data. In this experiment, mixtures of a common RNA background, in which 16 probesets were spiked in according to 14 different concentrations (0, 0.25, 0.5, 1, 2, 4, ..., 512, 1024 pM), were hybridized to a set of arrays. In most cases, each pattern of the 16 probeset concentrations

| pattern | Probeset |      |      |      |      |      |      |      |      |      |      |      |      |      |  |
|---------|----------|------|------|------|------|------|------|------|------|------|------|------|------|------|--|
|         | 1/12     | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9/16 | 10   | 11   | 13   | 14   | 15   |  |
| 1       | 0        | 0.25 | 0.5  | 1    | 2    | 4    | 8    | 16   | 32   | 64   | 128  | 512  | 1024 | 256  |  |
| 2       | 0.25     | 0.5  | 1    | 2    | 4    | 8    | 16   | 32   | 64   | 128  | 256  | 1024 | 0    | 512  |  |
| 3       | 0.5      | 1    | 2    | 4    | 8    | 16   | 32   | 64   | 128  | 256  | 512  | 0    | 0.25 | 1024 |  |
| 4       | 1        | 2    | 4    | 8    | 16   | 32   | 64   | 128  | 256  | 512  | 1024 | 0.25 | 0.5  | 0    |  |
| 5       | 2        | 4    | 8    | 16   | 32   | 64   | 128  | 256  | 512  | 1024 | 0    | 0.5  | 1    | 0.25 |  |
| 6       | 4        | 8    | 16   | 32   | 64   | 128  | 256  | 512  | 1024 | 0    | 0.25 | 1    | 2    | 0.5  |  |
| 7       | 8        | 16   | 32   | 64   | 128  | 256  | 512  | 1024 | 0    | 0.25 | 0.5  | 2    | 4    | 1    |  |
| 8       | 16       | 32   | 64   | 128  | 256  | 512  | 1024 | 0    | 0.25 | 0.5  | 1    | 4    | 8    | 2    |  |
| 9       | 32       | 64   | 128  | 256  | 512  | 1024 | 0    | 0.25 | 0.5  | 1    | 2    | 8    | 16   | 4    |  |
| 10      | 64       | 128  | 256  | 512  | 1024 | 0    | 0.25 | 0.5  | 1    | 2    | 4    | 16   | 32   | 8    |  |
| 11      | 128      | 256  | 512  | 1024 | 0    | 0.25 | 0.5  | 1    | 2    | 4    | 8    | 32   | 64   | 16   |  |
| 12      | 256      | 512  | 1024 | 0    | 0.25 | 0.5  | 1    | 2    | 4    | 8    | 16   | 64   | 128  | 32   |  |
| 13      | 512      | 1024 | 0    | 0.25 | 0.5  | 1    | 2    | 4    | 8    | 16   | 32   | 128  | 256  | 64   |  |
| 14      | 1024     | 0    | 0.25 | 0.5  | 1    | 2    | 4    | 8    | 16   | 32   | 64   | 256  | 512  | 128  |  |

Table 1: Latin square design for the Affymetrix U95A experiment. Probesets 1 through 16 are 37777\_at, 684\_at, 1597\_at, 38734\_at, 39058\_at, 36311\_at, 36889\_at, 1024\_at, 36202\_at, 36085\_at, 40322\_at, 407\_at, 1091\_at, 1708\_at, 33818\_at, 546\_at, respectively.

was replicated three times. A cyclic latin square design was used for the spike-in pattern of the target RNAs, which is displayed in Table 1. Irizarry et al. [7] provide a more detailed description of this experiment.

## 2.2 Gene Logic U95A spike-in data

The three Gene Logic experiments are nicely described in a white paper that accompanies the dataset, which can be requested from the Gene Logic website ([www.genelogic.com/media/studies](http://www.genelogic.com/media/studies)). These data are also described by Irizarry et al. [7]. The three experiments use the U95A and U95Av2 GeneChips. In all three instances, there were 11 spiked-in genes corresponding to bacterial genes normally used as (quality) controls for the array processing; the probesets for these genes each contain 20 probe pairs. One issue with this dataset is potential spatial contamination. All the spiked-in probes are in rows 13 and 14 (one row for PM, one for MM) of the  $640 \times 640$  array layout; columns 2 through 241 are BioB-5, BioB-M, BioB-3, BioC-5, BioC-3, BioDn-5, DapX-5, DapX-M, DapX-3, CreX-5, and CreX-3 in that order, 20 columns for each.

Experiment 1 consisted of 26 arrays. CreX-3 was at 0 concentration on all arrays and the remaining ten spiked-in genes were at 0, 0.5, 0.75, 1, 1.5, 2, and 3 pM (one array each), 5 and 100 pM (two arrays each), and 12.5, 25, 50, 75, and 150 pM (three arrays each). All arrays had a common complex RNA background derived from an acute myeloid leukemia (AML) tumor cell line. Experiment 2 used the latin square design shown in Table 2. Most rows were done in triplicate, yielding 32 arrays. The background was again an AML tumor

| pattern | BioB-5 | BioB-M | BioB-3 | BioC-5 | BioC-3 | BioDn-3 | DapX-5 | DapX-M | DapX-3 | CreX-5 | CreX-3 |
|---------|--------|--------|--------|--------|--------|---------|--------|--------|--------|--------|--------|
| 1       | 0.5    | 37.5   | 25     | 75     | 100    | 50      | 1.5    | 1      | 3      | 2      | 5      |
| 2       | 1      | 50     | 37.5   | 100    | 3      | 75      | 2      | 1.5    | 5      | 25     | 12.5   |
| 3       | 1.5    | 75     | 50     | 3      | 5      | 100     | 25     | 2      | 12.5   | 37.5   | 0.5    |
| 4       | 2      | 100    | 75     | 5      | 12.5   | 3       | 37.5   | 25     | 0.5    | 50     | 1      |
| 5       | 3      | 1.5    | 1      | 25     | 37.5   | 2       | 12.5   | 5      | 50     | 0.5    | 75     |
| 6       | 5      | 2      | 1.5    | 37.5   | 50     | 25      | 0.5    | 12.5   | 75     | 1      | 100    |
| 7       | 12.5   | 25     | 2      | 50     | 75     | 37.5    | 1      | 0.5    | 100    | 1.5    | 3      |
| 8       | 37.5   | 5      | 3      | 0.5    | 1      | 12.5    | 75     | 50     | 1.5    | 100    | 2      |
| 9       | 50     | 12.5   | 5      | 1      | 1.5    | 0.5     | 100    | 75     | 2      | 3      | 25     |
| 10      | 75     | 0.5    | 12.5   | 1.5    | 2      | 1       | 3      | 100    | 25     | 5      | 37.5   |
| 11      | 100    | 1      | 0.5    | 2      | 25     | 1.5     | 5      | 3      | 37.5   | 12.5   | 50     |

Table 2: Latin-square design for Gene Logic experiment 2.

cell line. Experiment 3 used a different latin square with concentrations of 0.5, 0.75, 1, 1.5, 2, 3, 5, 12.5, 25, 50, 75, and 100. There were three replicates of 12 patterns giving 36 arrays in total. The background material for this experiment was RNA extracted from human tonsillar tissue.

### 2.3 Affymetrix U133A spike-in data

The U133A spike-in dataset has 14 separate hybridizations of 42 spiked transcripts in a complex human background (HeLa cell line). Thirty of the spiked transcripts correspond to cDNA clones isolated from total RNAs of a lymphoblast cell line (and are not expressed in the HeLa cell line), eight of the spiked transcripts are made from artificial sequences, and the remaining four spiked transcripts are Affymetrix eukaryotic controls that are available as part of a polyA spike control kit. There were fourteen groups of three genes each and the concentration of each gene within a group was spiked at the same concentration. The fourteen groups are found in Table 3. A cyclic latin square was used for the group concentrations of 0, 1/8, 1/4, 1/2, 1, 2, 4,  $\dots$ , 256, and 512 pM. Each pattern appeared on three replicate arrays, yielding a total of 42 arrays. There were 11 probe pairs for each of 38 genes and 20 probe pairs for each of the four Affymetrix eukaryotic controls. Additional information about this experiment is available at the Affymetrix website mentioned in Section 2.1.

## 3 Analysis methods

Many years ago, Finney [6] suggested using a logistic function to fit data obtained from radioligand assays, where  $x$  is the log of the (known) dose and  $y$  the log of the observed intensity from the assay. Using a logistic function to also fit gene expression microarray data seems reasonable. The intensity values on these arrays span a wide range (roughly 10 to 46,000), so using an S-shaped curve to model the true value and observed value makes sense. The lower thresholds of the observed intensities are likely due to background binding

| Group |                 |                 |                 |
|-------|-----------------|-----------------|-----------------|
| 1     | 203508_at       | 204563_at       | 204513_s_at     |
| 2     | 204205_at       | 204959_at       | 207655_s_at     |
| 3     | 204836_at       | 205291_at       | 209795_at       |
| 4     | 207777_s_at     | 204912_at       | 205569_at       |
| 5     | 207160_at       | 205692_s_at     | 212827_at       |
| 6     | 209606_at       | 205267_at       | 204417_at       |
| 7     | 205398_s_at     | 209734_at       | 209354_at       |
| 8     | 206060_s_at     | 205790_at       | 200665_s_at     |
| 9     | 207641_at       | 207540_s_at     | 204430_s_at     |
| 10    | 203471_s_at     | 204951_at       | 207968_s_at     |
| 11    | AFFX_r2_TagA_at | AFFX_r2_TagB_at | AFFX_r2_TagC_at |
| 12    | AFFX_r2_TagD_at | AFFX_r2_TagE_at | AFFX_r2_TagF_at |
| 13    | AFFX_r2_TagG_at | AFFX_r2_TagH_at | AFFX_r2_DapX_at |
| 14    | AFFX_LysX_3_at  | AFFX_PheX_3_at  | AFFX_ThrX_3_at  |

Table 3: Affymetrix U133A spike-in experiment groups consisting of three genes each spiked at the same concentration.

or lower limits of detection of the instrumentation (the MM probes were meant to estimate this), while the upper limit may be due to biochemical or instrumentation saturation.

Functions other than the logistic can be fit to the binding data, some with superior physical model rationales, but Finney’s suggestion has at least three points in its favor. Firstly, the variation in measured data is often proportional to the mean (particularly for radioligand and other measurements based on photon counts), and so  $\log(\text{count})$  is approximately equivariant. This greatly improves the utility of the plots, and also somewhat simplifies fitting the curves since weighting is not needed. Secondly, the curves fit the actual data very well. Finally, the parameters of the fits are readily interpretable.

Since each probe has its own set of hybridization characteristics, we fit each separately. Specifically, the data are the different concentration levels for each probe ( $x$ ), which corresponds to the concentration level of the spiked transcript, and the observed intensities produced by the array ( $y$ ). Each spiked transcript consists of a probeset with 11 to 20 probes; perfect match (PM) and mismatch (MM) probes were plotted separately. The data were not normalized prior to fitting.

### 3.1 Logistic fits

The model used to fit the data, suggested for general binding data, was

$$\log(y) = a + bf(c[\log(x) - d])$$

where  $f$  is the logistic function,  $f(x) = \exp(x)/[1 + \exp(x)]$ . To fit the perfect match (PM) and mismatch (MM) probes simultaneously, we used a five-parameter version of the model

$$\log(y) = a + bf(c[\log(x) - d - e]) \quad (1)$$

where

- $f$  is the logistic function defined above,
- $a$  is the lower threshold for the probe pair (i.e. the background level),
- $b$  is the range of the curve,
- $a + b$  is the upper threshold for the probe pair (i.e. the saturation level),
- $bc/4$  is the slope of the curve at its inflection point,
- $d$  is the inflection point for the PM probes, and
- $d + e$  is the inflection point for the MM probes.

All plots and fits were done on log (base 2) transformed data. As a consequence, the shift parameter  $e$  represents the specificity of the MM probes relative to the PM probe for the target of interest. A value of  $e = 3$ , for example, would indicate that specific binding to the MM probe at a concentration of  $2^3x$  will be the same as binding to the PM probe at a concentration of  $x$ . Notice that our choice of a five parameter model to simultaneously fit the PM and MM probes forces each pair to have common lower and upper thresholds. All parameters were constrained to be positive.

A least squares approach was used to fit the curves to the data, which was not normalized and was not background corrected. Some of the probes, e.g. probe 10 of Figure 3, showed no binding at all across the range of concentrations. The fit of equation 1 for these *non-informative* probes is indeterminate: setting  $b = 0$ ,  $c = 0$ , or  $d > 20$  all give a horizontal line. Computer minimization routines typically have trouble with numerical situations such as this. As a solution, the fit was constrained to have  $c > 0.3$  and a penalty of  $(b - 6)^2$  was added to the sum of squares. The penalty causes the “large  $d$ ” option to be optimal for *non-informative* probes, without substantially affecting the fit for well-behaved probes. The penalty was chosen by looking at constrained fits for the well-behaved probes; the histogram of the fitted values of  $b$  was tightly centered around 6 and nearly symmetric.

## 3.2 Data plots

The observed expression values were plotted on the  $y$ -axis and the spike-in concentrations were plotted on the  $x$ -axis. A plot was made for each probe in the probeset; both the perfect match (PM) and mismatch (MM) values were plotted using different symbols. Plots for all the probe pairs within a probeset, referred to as a panel, were plotted on the same page in the order they appeared on the gene, with the first probe in the upper left hand corner and subsequent probes plotted from left to right and then down. Fitted logistic calibration curves were superimposed on the data plots. Figures 1, 2, and 3 show one panel (i.e. probeset) from each spike-in experiment, Affy95, GL, and Affy133, respectively. We have learned a considerable amount about the behavior and characteristics of Affymetrix data by looking through the complete set of plots for a spike-in experiment and strongly encourage others to look through at least one set for themselves.

Panels were generated for all spiked-in genes of each experiment. The entire sets for Affy95, GL, and Affy133 are available as supplementary appendices at the Mayo Clinic website [mayoresearch.mayo.edu/mayo/research/biostat](http://mayoresearch.mayo.edu/mayo/research/biostat) via the **Technical Reports** link. Appendix A contains the panels for the Affy95 dataset. This dataset has 59 arrays so each plot has 59 PM and 59 MM points. Each page (panel) corresponds to probeset for one spiked-in gene (16 total); each probeset contains 16 probe pairs yielding 256 total plots. Appendix B contains the panels for the three GL experiments. All three experiments have 220 plots (11 panels, 20 probe pairs each). There are 26 PM and MM points per plot in experiment 1, 32 PM and MM points per plot in experiment 2, and 36 PM and MM points per plot in experiment 3. Appendix C contains the plots for the Affy133 dataset. There are 498 plots (38 panels with 11 probe pairs and 4 panels with 20 probe pairs), with 42 PM and MM points each.

## 4 Results

An immediate observation is that the log transformation of the probe intensities stabilizes the variance. From the plots in Figures 1 - 3, the data appear to be distributed equivariantly throughout the range of the true concentrations. The residual plot in Figure 4 for the data in Figure 1 further supports this. Hence, if no background correction is performed on the probes, the log transformation adequately stabilizes the variance. On the other hand, if the probe values have been adjusted for background, then as has been noted by several others, the log transformation does not adequately stabilize the variance. Log transformed gene expression values generated by Affymetrix MAS 5.0 exhibit larger variance for lower expression levels primarily because they are a summary of  $\log_2(PM) - \log_2(MM)$  values, which is meant to adjust for non-specific binding. The problem persists even when only  $PM'$  values are used, where  $PM' = PM - c$  are PM values that have been adjusted by a constant that estimates background (a combination of spatial, instrument, and non-specific binding); evidence of this is that expression values produced by RMA [7] and GCRMA [5] using only the  $PM'$  probes also exhibit larger variance for lower expression values than for the higher expression values.

Looking at the plots of observed intensity by dose level, there are different types of behavior for the probe pairs. The observed intensity level of the probes generally increases with the dose level for both the perfect match (PM) and mismatch (MM) probes; this indicates that in general, the MM probes measure signal as well as non-specific hybridization although their measure of signal is attenuated. Most probes exhibit an S-shape increase that appears to be well approximated by the logistic curve. The majority of probe pairs are *well-behaved* meaning that the PM probes have higher observed intensities than the MM probes and both MM and PM observed intensities increase as the true concentration increases. For *well-behaved* probe pairs, the shapes of the MM curves appear similar to the shapes of the PM curve but are shifted to the right. As mentioned earlier, some probes are *non-informative*, meaning both the PM and MM curves are flat across the range of the spike-in dose levels. It is hypothesized that a *non-informative* probe does not measure the gene

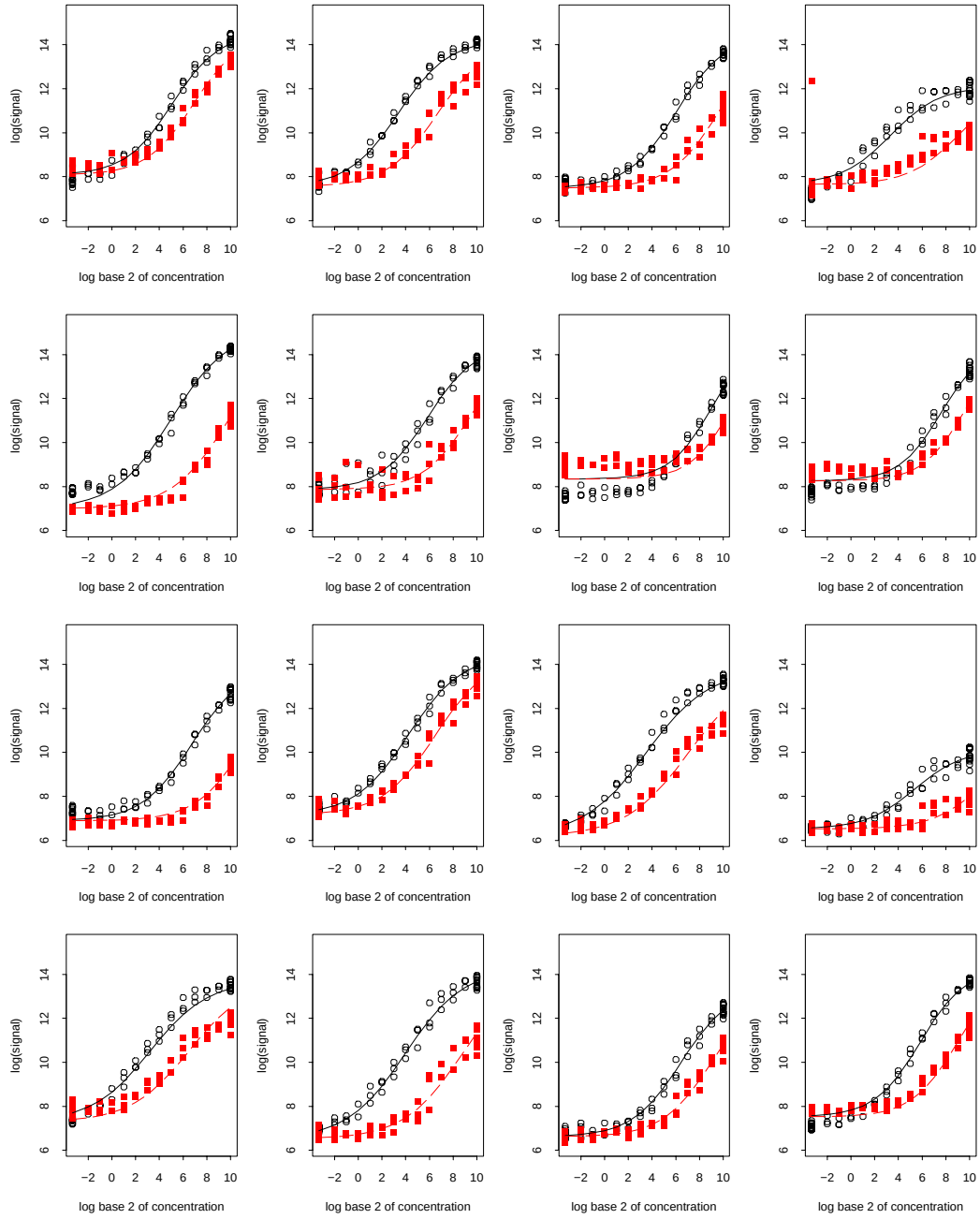


Figure 1: A panel of plots corresponding to the probeset 684\_at from Affy95. Each plot in the panel shows the observed intensity value versus the concentration of the spiked gene for a probe pair. Open circles represent the PM probe values and the solid line is the corresponding fitted logistic calibration curve. The filled squares and dashed line represent the MM probe values and fitted logistic calibration curve, respectively.



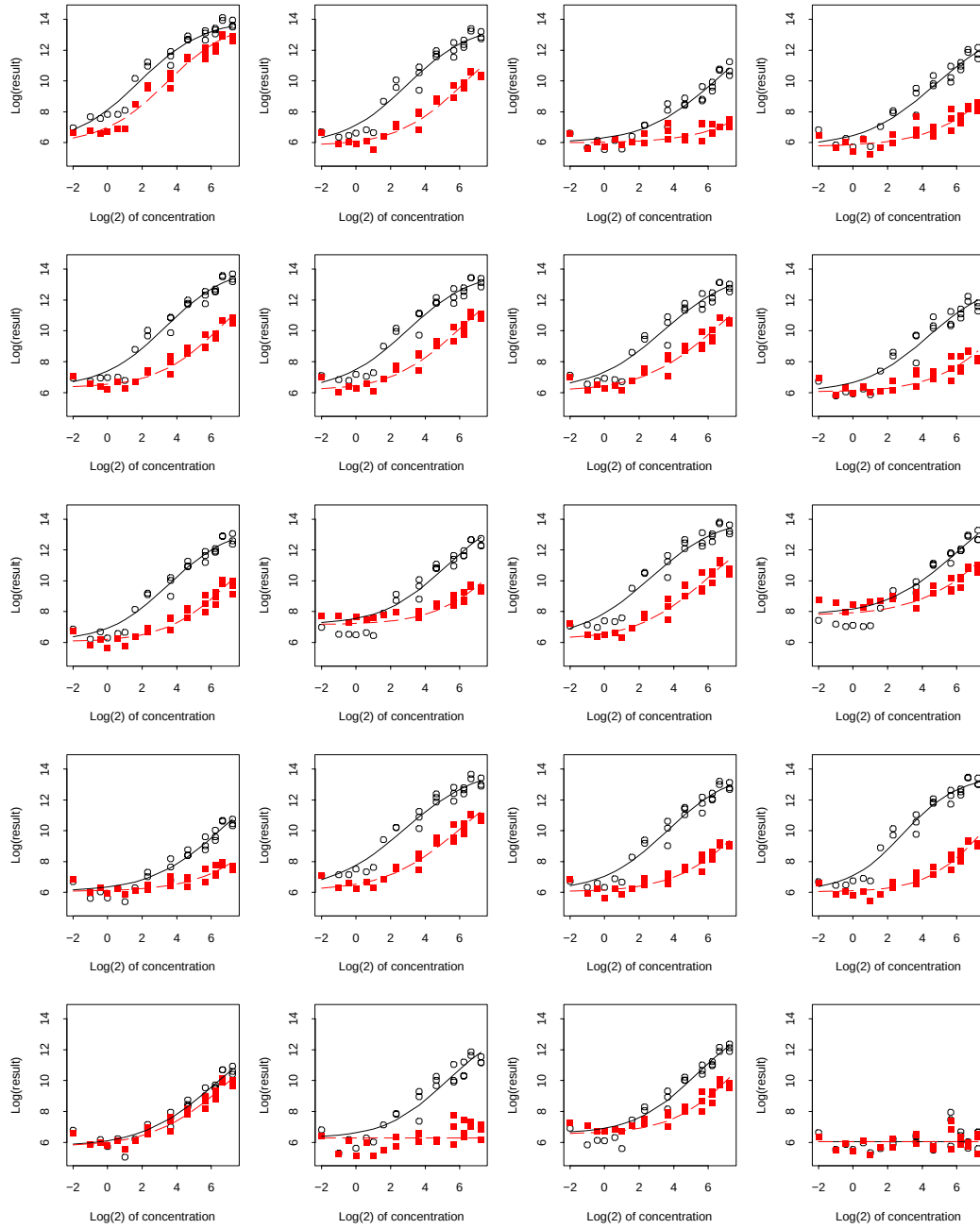


Figure 2: A panel of plots corresponding to the probeset CreX-5 from GL experiment 1. Each plot in the panel shows the observed intensity value versus the concentration of the spiked gene for a probe pair. Open circles represent the PM probe values and the solid line is the corresponding fitted logistic calibration curve. The filled squares and dashed line represent the MM probe values and fitted logistic calibration curve, respectively.

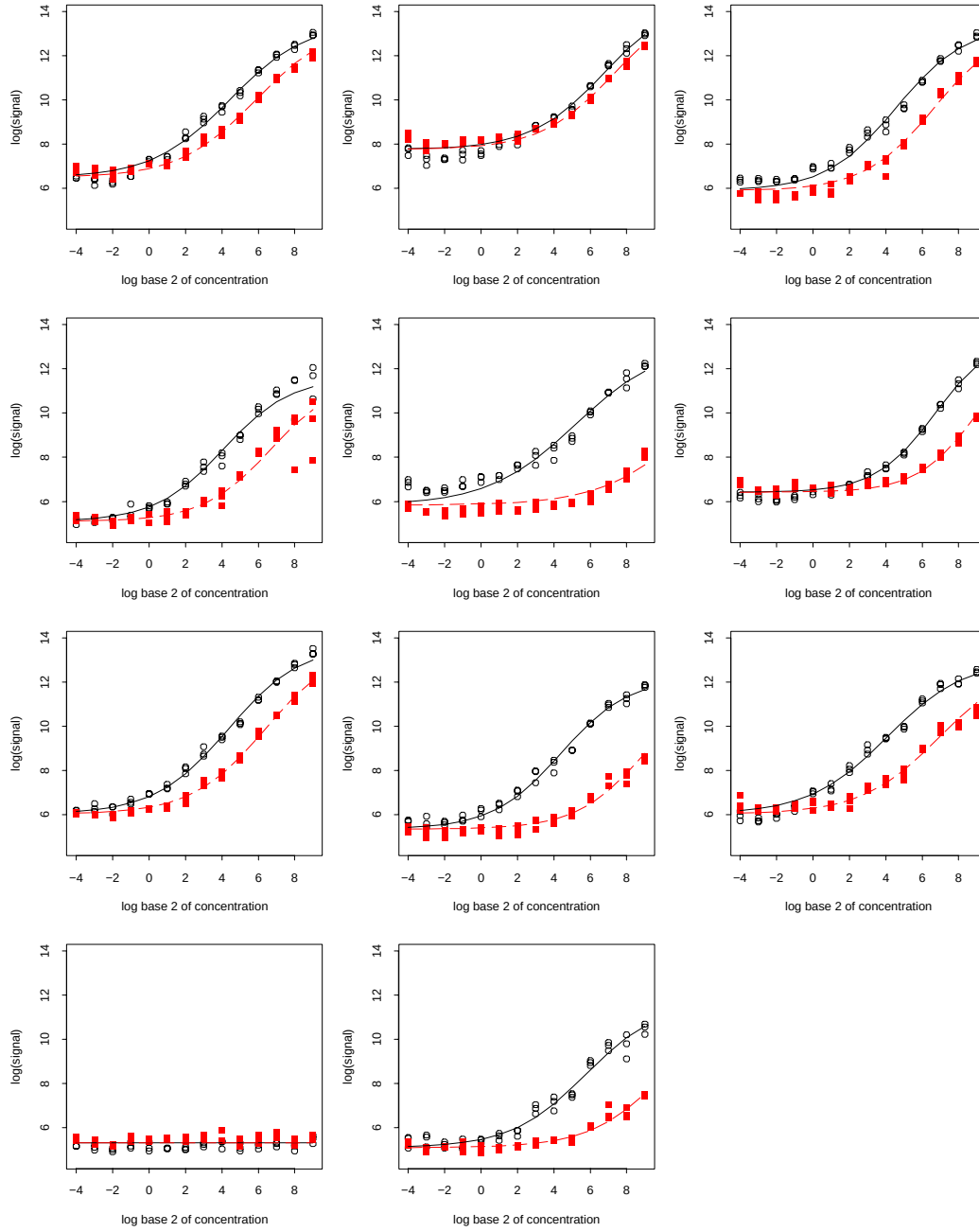


Figure 3: A panel of plots corresponding to the probeset 206060\_s\_at from Affy133. Each plot in the panel shows the observed intensity value versus the concentration of the spiked gene for a probe pair. Open circles represent the PM probe values and the solid line is the corresponding fitted logistic calibration curve. The filled squares and dashed line represent the MM probe values and fitted logistic calibration curve, respectively.

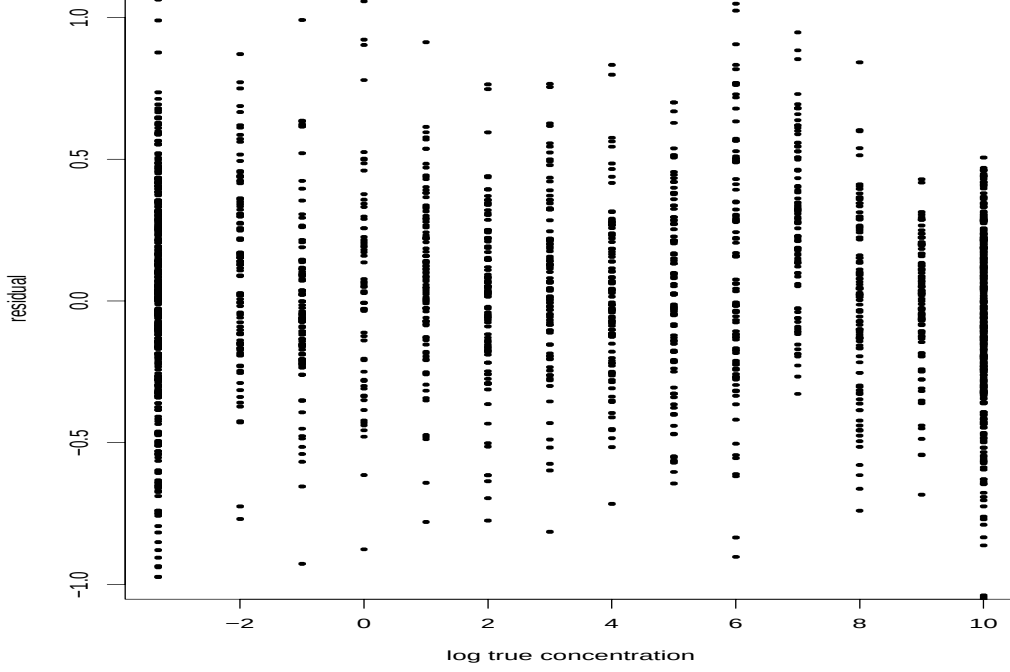


Figure 4: The residual plot from the logistic fits displayed in Figure 1.

of interest perhaps because the 25mer on the array is incorrect either due to a mistake in selecting the probe or in manufacturing the array, or because post-transcriptional processing in the cell that cut that part of the sequence out of the mRNA. Regardless, the probe appears to measure only non-specific hybridization or it just has terrible sensitivity. Another pattern observed is probes with MM background greater than PM background; these are labeled *background aberrant*. A plausible explanation for this pattern is cross-binding; the swap of base 13 actually makes the MM a good match for some other RNA species. A final probe type is labeled the *Affy assumption*. This is where the PM probe observed intensity increases and the observed intensity level of the MM probe remains flat as the spiked-in dose increases (which is an implicit assumption in the MAS 5 formulas). Table 4 summarizes the results of subjectively classifying the probe pairs into one of the four classes: *well-behaved*, *non-informative*, *background aberrant*, and *Affy assumption*. The classification was made by looking at each plot and judging which of the four patterns best describes its behavior. As a result, the numbers in Table 4 would likely differ, hopefully only slightly, for another observer.

Some plots display behaviors not captured by the four classes described above. Two such behaviors are plots with outliers and plots with PM background greater than MM background. Outliers were subjectively determined as points that are far from the fitted logistic curve, an example is the fourth probe in Figure 1. As is the case here, most outliers did not greatly affect the overall logistic fit. Each experiment has some plots (i.e. probes)

| probe pair type     | Affy U95A<br>(n=256) | GL1<br>(n=200) | GL2<br>(n=220) | GL3<br>(n=220) | Affy U133A<br>(n=498) |
|---------------------|----------------------|----------------|----------------|----------------|-----------------------|
| well-behaved        | 204 (80%)            | 169 (85%)      | 154 (70%)      | 133 (60%)      | 408 (82%)             |
| non-informative     | 18 (7%)              | 2 (1%)         | 8 (4%)         | 14 (6%)        | 5 (1%)                |
| background aberrant | 18 (7%)              | 22 (11%)       | 10 (5%)        | 6 (3%)         | 60 (12%)              |
| Affy assumption     | 19 (6%)              | 7 (4%)         | 48 (22%)       | 67 (30%)       | 25 (5%)               |

Table 4: Summary of the number of probes in each of the four different probe type classifications: *well-behaved*, *non-informative*, *background aberrant*, and *Affy assumption*.

with outliers; there are approximately 11 plots (4%) in Affy95 data; approximately 21 plots (11%), 138 plots (63%), and 56 plots (25%) in experiments 1, 2, and 3, respectively, of the GL data; and approximately 15 plots (3%) in the Affy133 data. Each dataset also has plots for which the PM background level is greater than MM background, violating the assumption that PM and MM background levels are the same. (This assumption is also violated by the *background aberrant* probes.) Based on a subjective assessment of the plots, the percent of plots (i.e. probes) for which PM background is higher than MM background ranges from a minimum of 3% (GL experiments 1 and 3) to a maximum of 14% (GL experiment 2). A possible explanation for this behavior is cross-binding of the PM probe with some other RNA species.

Overall, it appears as though the GL experiments are the noisiest, with the largest number of probe plots containing outliers. In addition, the first and second GL experiments appear to have a smaller proportion of *well-behaved* probes; interestingly though, these same two experiments have a larger fraction of *Affy assumption* probes. The GeneChip U95A probes appear to have a greater fraction of *non-informative* probes compared to the GeneChip U133A probes. Finally, it is apparent that the Affymetrix assumption that MM probes measure only non-specific hybridization is rarely correct, as has been noted by other investigators [4]. In the results that follow, the most detail is provided for the Affy95 experiment. Commonalities and deviations of results for the other spike-in experiments are noted.

## 4.1 Goodness of fit

### 4.1.1 Affymetrix U95A experiment

The panels of plots and fits for the Affy95 experiment are found in supplementary Appendix A. The fits were extremely good for all but a few *background aberrant* probes. A stemplot of the standardized residuals for the 256 fits (not shown) had two clear parts: a fairly symmetric central portion with a median of 0.4, and a long outlier tail made up of 20 of the 256 fits. Of the 20 fits with the largest SSE, 13 were *background aberrant* probes, one had a PM background that was considerably larger than the MM background, and four had outliers.

### 4.1.2 GeneLogic U95A experiments

The panels for the three GL experiments are found in Supplementary Appendix B. The fits for experiment 1 were very good. An interesting artifact was array 23, which was responsible for a large outlier MM value in the sequence of plots starting with BioB-3 probes 19 and 20 through probe 15 of BioC-5. Given that these probes were contiguous on the array, this suggests a scratch or some other physical problem. There were roughly 22 *background aberrant* probes, but the overall magnitude of this problem on the goodness of the fits was much less than in the other spike-in experiments. There were only two *non-informative* probes. Since these genes were used as controls for all GeneChip U95A arrays, it is perhaps not surprising that the choice of probes produced such a large fraction of informative probes.

It is particularly interesting to compare experiments 2 and 3, which have very similar designs but a different human RNA background into which the set of genes are spiked. It is immediately apparent that the spread of the data about the fitted curves for experiment 2 was far larger than for the other spike-in experiments. An exploration of the outliers revealed that over half of them belonged to only four arrays: 1, 4, 6, and 9, which correspond to the first, second, third, and fourth expression patterns, respectively. This raised the possibility of an error in the concentration data. On examination, we found that swapping the patterns for arrays 1 and 6 and for arrays 4 and 9 yielded a large improvement in the fit. That is, assume that array 1 used the third pattern (rather than the first), array 6 the first pattern (rather than the third), etc. Given this, for arrays 1, 4, 6, and 9, we assumed a clerical error had been made. The plots in Appendix B are the plots after the correction. For some genes such as BioB-3, this switch did not alter the concentration levels very much and the plots were little changed; for others, such as Cre-X, the improvement was substantial.

### 4.1.3 Affymetrix U133A experiment

Again, the logistic fits were quite good for the majority of the plots. A stemplot of the standardized residuals of the 498 fits consisted of a symmetric kernel (with an upper value of approximately 0.6) with a long right tail that accounted for about 10% of the probes. The largest errors were again for *background aberrant* probes, where the baseline binding for PM and MM were not the same; this accounted for all the errors greater than 0.6 with six exceptions, four of which were plots with outliers.

## 4.2 Parameters

### 4.2.1 Affy U95A experiment

A summary of the parameter estimates of equation (1) corresponding to the fitted curves is presented in Table 5. The slope of the logistic curve at its inflection point is  $bc/4$ ; these values are listed in the table along with the values for the other parameters. The slope of the curves at their inflection point was nearly constant; the 10th and 90th percentiles differed from the median by less than 25% with a median value of 0.71.

The multiplier  $c$  was even more stable than the slope, and a refit of the data with this parameter fixed at the median value ( $c = 0.47$ ) did not substantially change the fit (not

| Percentile | Lower     |       |      |       | MM         |        |
|------------|-----------|-------|------|-------|------------|--------|
|            | Threshold | Range | $c$  | Slope | Inflection | Offset |
| min        | 6.3       | 0.1   | 0.30 | 0.02  | 1.1        | 0.0    |
| 10%        | 6.5       | 5.4   | 0.37 | 0.53  | 3.3        | 1.7    |
| 25%        | 6.6       | 5.8   | 0.43 | 0.62  | 4.1        | 2.4    |
| 50%        | 7.1       | 6.0   | 0.47 | 0.71  | 5.4        | 3.4    |
| 75%        | 7.9       | 6.4   | 0.52 | 0.79  | 7.3        | 4.6    |
| 90%        | 8.9       | 6.7   | 0.57 | 0.86  | 9.9        | 6.1    |
| max        | 13.4      | 7.5   | 0.77 | 1.05  | 224.7      | 251.2  |

Table 5: Summary of the values of the parameters of the logistic fits for the Affy95 experiment.

shown). It should be noted that setting the slope equal to 0.71 (i.e.  $bc/4 = 0.71$ ) produced noticeably inferior fits (not shown). The range of the curves was relatively stable around a central value of 6.0. This was a bit surprising, since we initially expected that it would be the upper threshold that was constant. Most investigators believe that RNA concentrations above 256 are unlikely to occur in microarray experiments conducted as part of medical research. Thus the upper bend in curves from real data would not be observed, and the upper threshold may need to be set equal to a constant whose value is based on prior knowledge.

The parameters with the most variation from fit to fit were the intercept  $a$  and the PM inflection point  $d$ . The extreme values for the inflection points all belonged to *non-informative* probes where the parameter can only be determined to be “very large”. The maximum sensitivity for most of the probes occurred between the concentrations of  $2^4$  and  $2^7$ . On average, the MM probes had a specific binding affinity of approximately 1/10 of the corresponding PM probe for the target.

The variation in baseline (background) binding may be due to multiple causes. Some possible explanations include: (1) non-zero expression for some genes in the global RNA background into which the genes were spiked (although one would expect the probes to be chosen so this is not so), (2) cross-binding of some expressed RNA to the probes, or (3) differences in background binding affinities due to the nucleotide base content of the 25mer probe. Naef and Magnasco [8] suggest that probe binding affinities are largely dependent on the position of the specific nucleotide within the 25mer probe. They found that position-dependent affinities result in substantially better fits than fits based on the frequencies of each base in the probe (e.g. models based on the number of G and C bases). We determined approximate values of the position-dependent binding affinity values for each base (A, C, G, T) from Figure 3 in the paper by Naef and colleagues [8]. These values were used to compute a binding affinity value for each MM probe of the spiked-in probesets in the Affy133 experiment. The correlation between these computed binding affinities and the estimated background levels of the probes from the logistic fits (i.e. estimated value of  $a$ ) was quite good,  $r = 0.74$ . Finer models have been proposed that account for nearest-neighbor interactions along the 25mer, such as the stacking energy, positional-dependent-

nearest-neighbor model [9]. Models using nucleotide base content information could be employed to estimate the background binding affinities for a probe, i.e. the value of  $a$  for a probe-specific logistic calibration curve.

#### 4.2.2 GeneLogic experiments

Table 6 shows the values of the parameters of the fits for three GL spike-in experiments. There was quite good consistency of the parameters across fits and fits across the experiments; however, the third experiment appeared to have lower overall intensity, as reflected in the lower threshold values. This shows the need for normalization if the data were to be compared across experiments. Other results of the fits on the GL spike-in data were consistent with those observed for the Affy95 spike-in data.

A major concern in gene expression experiments is the quality of the mRNA. In particular, mRNA is relatively unstable and degrades quickly if the biospecimens are not handled and stored adequately. If the mRNA becomes sufficiently degraded, the average expression level of the genes will equal background (non-specific binding). mRNA degradation can be crudely measured in terms of location of the inflection point for the PM probes. Specifically, as RNA degrades, we get the appearance of a decrease in the total amount (i.e. concentration), which affects the location of the inflection point (i.e.  $d$ ). Since RNA degradation typically starts from the 5' end of the molecule, we would expect the inflection point location for probes near this end to be systematically larger than the probes near the 3' end. The greater the amount of RNA degradation, the more pronounced this trend. Figure 5 shows the location of the point of inflection as a function of the probe location within the probeset for each of the spiked-in genes from the GL experiment 2. The lowess smoother shows an upward trend in the inflection point location as we move from the 3' end to the 5' end. Such a trend would be more pronounced in experiments done under less controlled conditions than this. Of all the experiments, GL experiment 2 exhibited the most potential RNA degradation.

#### 4.2.3 Affymetrix U133A experiment

A summary of the parameters from the fits is given in Table 7. Like the results for the U95A arrays, the parameter  $c$  was nearly constant; the 10th and 90th percentiles differed from the average by only 20%. Refits of the data with this parameter fixed at 0.52 yielded essentially the same fit (data not shown). The range of the curves was likewise very stable around its central value of 7.2. This was a somewhat larger range than the U95A arrays; however, the fact that the upper threshold was rarely observed in this data (due to the lower maximal concentration) means that this parameter was not well identified. The quartiles for the PM/MM offset parameter gave a fold-difference range of  $2^{1.9} = 3.8$  to  $2^{3.5} = 11.2$  for the relative specificity of the PM probe. The range of these numbers was similar to those for the U95A data, though somewhat smaller. Other results of the fits on the Affy133 spike-in data were consistent with those observed for the Affy95 spike-in data.

|     | Experiment 1       |       |      |            |              |
|-----|--------------------|-------|------|------------|--------------|
|     | Lower<br>Threshold | Range | $c$  | Inflection | MM<br>Offset |
| min | 5.4                | 6.5   | 0.36 | 0.6        | 0.2          |
| 10% | 5.7                | 7.5   | 0.45 | 2.4        | 1.5          |
| 25% | 5.9                | 7.8   | 0.41 | 3.2        | 2.0          |
| 50% | 6.2                | 8.0   | 0.51 | 4.8        | 2.6          |
| 75% | 6.9                | 8.0   | 0.54 | 6.1        | 3.4          |
| 90% | 7.8                | 8.1   | 0.56 | 8.1        | 4.1          |
| max | 10.8               | 8.3   | 0.65 | 39.9       | 15.0         |
|     | Experiment 2       |       |      |            |              |
|     | Lower<br>Threshold | Range | $c$  | Inflection | MM<br>Offset |
| min | 5.4                | 6.1   | 0.30 | 1.3        | 0.0          |
| 10% | 5.9                | 7.0   | 0.35 | 2.5        | 1.5          |
| 25% | 6.1                | 7.5   | 0.42 | 3.4        | 2.3          |
| 50% | 6.5                | 7.8   | 0.48 | 4.9        | 3.1          |
| 75% | 7.1                | 7.9   | 0.51 | 6.2        | 4.3          |
| 90% | 8.2                | 8.0   | 0.56 | 8.0        | 5.8          |
| max | 10.4               | 8.3   | 0.69 | > 50       | > 50         |
|     | Experiment 3       |       |      |            |              |
|     | Lower<br>Threshold | Range | $c$  | Inflection | MM<br>Offset |
| min | 4.2                | 6.1   | 0.36 | 1.9        | 0.0          |
| 10% | 5.0                | 7.3   | 0.45 | 3.0        | 1.4          |
| 25% | 5.3                | 7.6   | 0.41 | 3.7        | 1.9          |
| 50% | 5.9                | 7.9   | 0.51 | 5.1        | 2.7          |
| 75% | 6.6                | 8.1   | 0.54 | 6.3        | 3.8          |
| 90% | 7.9                | 8.4   | 0.56 | 7.8        | 7.3          |
| max | 10.5               | 8.8   | 0.65 | 39.4       | > 50         |

Table 6: Summary of the fits for the GL spike-in data.



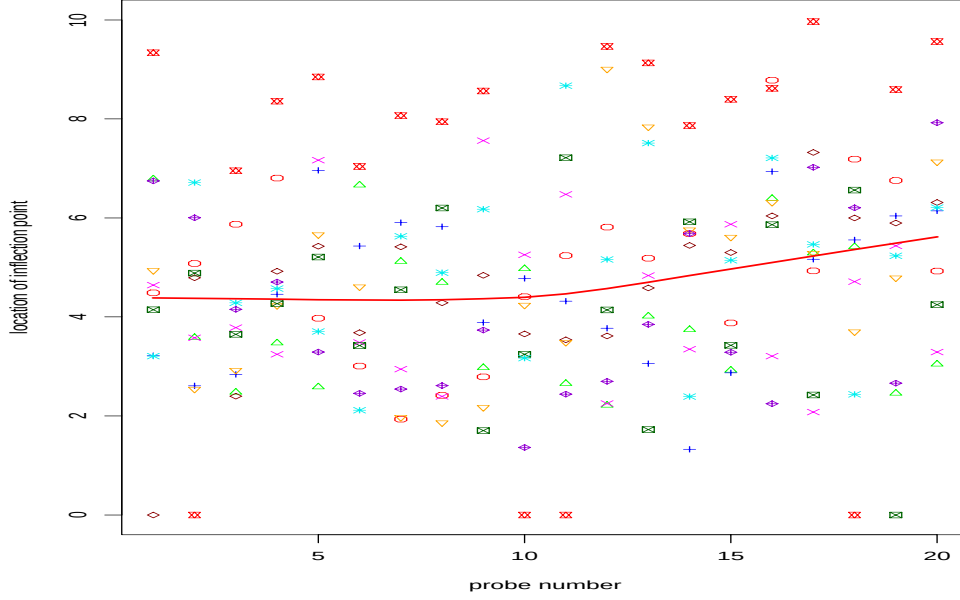


Figure 5: The location of the point of inflection as a function of probe location within a probeset. Probe 1 is closest to the 3' end and probe 20 closest to the 5' end. Points with the same shape (color) belong to the same probeset (11 total). The solid line is a lowess smooth of the location of the inflection point across all the spiked-in probesets.

|     | Lower     |       |      |       | MM         |        |
|-----|-----------|-------|------|-------|------------|--------|
|     | Threshold | Range | $c$  | Slope | Inflection | Offset |
| min | 4.8       | 5.5   | 0.30 | 0.49  | 2.3        | 0.0    |
| 10% | 5.3       | 6.8   | 0.43 | 0.76  | 3.8        | 1.3    |
| 25% | 5.7       | 7.0   | 0.47 | 0.86  | 4.5        | 1.9    |
| 50% | 6.2       | 7.2   | 0.52 | 0.95  | 5.4        | 2.6    |
| 75% | 6.9       | 7.5   | 0.57 | 1.02  | 6.4        | 3.5    |
| 90% | 7.7       | 7.7   | 0.62 | 1.11  | 7.5        | 5.1    |
| max | 11.0      | 8.1   | 1.07 | 1.88  | 273.7      | > 999  |

Table 7: Summary of the fits for the Affy133 spike-in data.

|                                                  |        |       |
|--------------------------------------------------|--------|-------|
| total sums of squares                            | 108555 |       |
| Effect                                           | RSS    | $R^2$ |
| gene concentration (linear fit)                  | 29303  | 73.0% |
| gene concentration (logistic fit)                | 22422  | 79.3% |
| gene concentration + probe effect                | 2349   | 97.8% |
| gene concentration + probe effect + array effect | 2105   | 98.1% |

Table 8: Residual sums of squares and  $R^2$  values for different models.

### 4.3 Array effects

It is widely accepted that microarrays, including Affymetrix GeneChips, require normalization prior to comparisons of gene expression levels across arrays. Based on the normalization work done with Affymetrix GeneChips [4, 10, 11], it is expected that these arrays have intensity dependent biases. A relevant question is, how large are array effects compared to (1) potential changes in gene expression levels among groups of biospecimens, and (2) the probe effects? To explore this, we determined how much these different components (i.e. gene concentration, probe, and array) contribute to the total variation of the observed intensity values as measured by the residual sums of squares (RSS). Our analysis used the PM data of the spiked-in probesets for the Affy133 experiment. The total sums of squares (TSS) of the observed PM intensities was 108555.

Table 8 shows the RSS for different models and the corresponding  $R^2$  values, where  $R^2 = 1 - RSS/TSS$ . Clearly the largest reduction in the sums of squares was made by accounting for the different concentrations of the spiked-in probesets (gene concentration). About 73% of the total variation in the observed PM intensities was explained with a linear model of the concentrations of the spiked-in probesets. When a logistic model was used instead of a line, the  $R^2$  value increased to approximately 79%, a modest gain of about 6%. Both these models assumed that the individual probes have the same (common) relationship between the known concentration and observed intensity.

If we allow each probe to have its own (logistic) model for the relationship between the known concentration and observed intensity, the resulting reduction in RSS (increase in  $R^2$ ) is considerable. The  $R^2$  value for a model using individual logistic fits for each probe, i.e. the fits shown in Figures 1 - 3, was 97.8%. The contribution of the probe-level information was smaller than that of the gene concentration level in this set of experiments; however, it was still a sizable effect.

Finally, the reduction in the RSS after adjusting for the individual array effects (using a nonlinear normalization, *fastlo* [11]) is quite small—the  $R^2$  value increases to 98.1% (a 0.3% increase over a model for gene concentration and individual logistic probe effects). In this particular set of experiments, there was very little systematic variation from array to array. It is likely that less controlled experiments, such as those typically done in a clinical/biological research project, would exhibit a greater level of array-to-array variation, although it probably still would be a relatively small portion of the total variation of the observed intensity values. It is possible to obtain a precise estimate of this relatively small

piece of the total variation via normalization because these procedures tend to use all the probes on the array (roughly 500,000 for the U133A GeneChip), and not just the probes in the probesets of interest.

## 5 Discussion

This simple look at the data has been revealing. The relationship between truth, the amount of RNA in a sample, and the observed probe intensity is well-described by a logistic curve; a phenomenon that has been observed in other binding processes. It is plausible that the flat, non-zero left tail results from non-specific binding and/or scanner effect (i.e. reflected light), when the target gene is absent or expressed at very low levels. The asymptotic behavior of the upper right tail is likely the result of a relatively small number of binding sites compared to the amount of RNA in the sample or possibly (but less likely) because the probe intensity is above the upper range of the scanner, i.e. saturation (chemical and/or scanner). In general, when there is a sufficient amount of RNA present at levels well below saturation, the relationship between the observed intensity and actual amount of RNA in the sample is essentially linear, with a slope less than one; a plausible explanation for the attenuation in binding efficiency could be that the array hybridization conditions (i.e. temperature, duration, etc.) were not optimal for the specific probe.

These data support the observation that the MM probes appear to measure the expression level of the target gene as well as non-specific binding; in a large fraction of probe pairs, the observed MM intensity increased as the corresponding RNA concentration increased. In general, the relationship between truth and observed MM intensity is well described by a logistic curve with shape identical to that of the corresponding PM calibration curve. Specifically, the calibration curves for the PM and MM probes within a probe pair have similar background levels, similar saturation levels, and similar inflection points; however, the MM curves tend to be shifted to the right of the matched PM curves. This implies that the MM probe requires a greater amount of a gene-specific RNA to yield the same observed intensity as its PM probe, or equivalently, that for a given amount of gene-specific RNA (above background level), the observed intensity of the MM probe will be lower than the observed intensity of its paired PM probe. In other words, the binding affinity of the MM probes appear to be attenuated compared to the PM probes, the consequence of the manufactured mismatch of the base in the 13th position of the 25mer in the MM probe.

Logistic curves were fit to each PM-MM probe pair by assuming the curves had identical shapes with the MM curve shifted to the right of the PM curve. The fits of these calibration curves were extremely good for over 90% of the probe pairs. Cases of poor fit generally occurred when the best fitting logistic curves for the PM and MM probe, within a probe pair, did not have the same shape due to different background levels such as for *background aberrant* probes.

Other interesting aspects of the data were revealed by the plots and fitted curves. The background levels varied from probe pair to probe pair, even for probe pairs in the same probeset. Not surprisingly, the PM and MM probes within a probe pair had essentially the

same background level; probably because they only differed in the 13th base of the 25mer. It is likely that probes with considerable overlap with respect to the make-up of their 25mer sequence are more likely to have similar background binding, as supported by work of individuals investigating binding affinity dependence on the specific content of the 25mer. Another aspect of the data was that differences in the RNA concentration explained most of the variation in the observed probe intensities. Individual probe effects also explained a sizable amount of the variability whereas the array effects were quite small in comparison to the concentration levels and probes. Although the variability among arrays is likely to be larger for less controlled experiments, i.e. those using (non-optimally) collected human tissues, differences in RNA levels and probe effects will likely still be considerably larger. Normalization is nonetheless beneficial since it is possible to precisely estimate the small systematic array effects due to the large amount of probes on the arrays.

Our goal was to look at the data from the spike-in experiments to understand the nature of the binding or calibration curves for this technology—i.e. to understand the relationship between truth (the actual amount of RNA in the sample) and the intensity values produced by the microarray platform. Knowledge obtained from this exercise can be used to assess different models and analytical procedures developed for Affymetrix GeneChip data. For example, subtracting the value of the MM probe from the PM probe as a means of adjusting for non-specific binding seems questionable since there is evidence that MM probes measure signal. Our take-home message is that the behavior of probe binding affinities is well described by an S-shaped curve, such as a logistic curve. Furthermore, for a large majority of probe pairs, the calibration curves for the PM and MM probes in a probe pair have the same shape (similar background levels, saturation levels, and points of inflection) with the MM curve shifted to the right of the PM curve. The background level varies considerably among probe pairs and appears to depend on the composition of the 25mer sequence of the probe. Much work has been devoted to modeling the effects of RNA concentration levels and normalization of arrays and less in terms of accounting for individual probe effects, although a sizable amount of variability of observed intensity levels can be explained by these effects. Our current work is focused on developing a model that appropriately captures characteristics of the data; specifically the S-shape nature of the probe calibration curves, the need for different calibration curves for each probe pair, and the relationship between the PM and MM probes. We urge others to look at this data, available in the supplementary appendices (website given in the text), and draw their own conclusions with respect to different available models and analytical techniques or as a guide in the development of new techniques.

## References

- [1] Affymetrix. *Microarray Suite User Guide, Version 5*. Affymetrix, Inc., 2001. <http://www.affymetrix.com/support/technical/manuals.affx>.
- [2] C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci.*, 98:31–36, 2001.

- [3] T-M. Chu, B. Weir, and R. Wolfinger. A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences*, 176:35–51, 2002.
- [4] R. Irizarry, B. Hobbs, F. Collins, Y. Beazer-Barclay, K. Anntonellis, U. Scherf, and T. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.
- [5] Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez Murillo, and F. Spencer. A model based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99:909–917, 2004.
- [6] D. J. Finney. Radioligand assay. *Biometrics*, 32:721–740, 1976.
- [7] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31:e15, 2003.
- [8] F. Naef and M. O. Magnasco. Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Physical Review E*, 68:011906–1–011906–4, 2003.
- [9] L. Zhang, M.F. Miles, and K. D. Aldape. A model of molecular interactions on short oligonucleotide microarrays: Implications for probe design and data analysis. *Journal of Molecular Biology*, 21:818–821, 2003.
- [10] B. Bolstad, R. Irizarry, M. Astrand, and T. Speed. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, 19(2):185–193, 2003.
- [11] K. V. Ballman, D. E. Grill, A. L. Oberg, and T. M. Therneau. Faster cyclic loess: Normalizing rna arrays via linear models. *Bioinformatics*, 20:2778–2786, 2004.