# Rough Draft
# Not For Distribution

# Quality Control Metrics for Array Comparative Genomic Hybridization Data

Jeffrey C. Miecznikowski[ab1], Daniel P. Gaile[ab], Conroy, Jeffrey, Norma J Nowak[c]

[a]Department of Biostatistics, University at Buffalo, Buffalo, NY 14214-3000, USA
[b]Department of Biostatistics, Roswell Park Cancer Institute, New York 14263
[c]Cancer Genetics, Roswell Park Cancer Institute, New York 14263

Short title:

**Quality Control for aCGH**

Proofs to be sent to:

Jeffrey C. Miecznikowski
Department of Biostatistics
School of Public Health and Health Professions
249 Farber Hall
University at Buffalo
3435 Main Street
Buffalo NY 14214-3000, USA

[1] Corresponding Author. Department of Biostatistics, School of Public Health and Health Professions, 249 Farber Hall, University at Buffalo, 3435 Main Street, Buffalo NY 14214-3000, USA. Tel:+1(716) 829 2754. e-mail:dpgaile@buffalo.edu

## Abstract

The main focus in cDNA microarray analysis is determining which genes are differentially expressed. Scientists apply known statistical methods to model the structure of the experiment or develop new approaches for assessing statistical significance and assume that the data consist of the signal plus random noise. Here, we report the results of some exploratory analyses of such data that show the existence of sources of significant systematic variation which are not necessarily accounted for in standard analyses. Furthermore, we consider not only the variation due to the pin/print-tip as in previous work, but also the row and column location on the microarray chip, and the relative location from the well-plate. Removal of this extra variation can affect both the size of differential gene expression, and which genes are inferred to be differentially expressed. Further, we present violin plots as a measure to evaluate the quality of the probes broken down over pin, 384 well plates, plate rows and plate columns.

## Introduction

In any new technology, studies involving quality control and tolerance should be performed. Once the technology has been benchmarked as suitable, continuing checks should be performed in order to ensure that the technology remains in check. With the burgeoning field of high throughput genomics, specifically microarrays, there have been numerous studies devoted to each facet of the technology, from image processing and data acquisition to end stage processing of gene networks. The success of each data processing step is highly dependent on the steps preceding it. In order to ensure integrity in the data based decision making process, it is necessary to have a set of tools which ensures the technology is in control and producing reliable results.

By advocating the use of linear models based on the spotting process and violin plots for hybridization quality this paper demonstrates quality control metrics to ensure the integrity of the spot assay process. The rest of this paper is organized as follows: an overview of the aCGH microarray technology, a section on quality control metrics with examples and results, and a discussion of applicability of the methods.

Pin tip (sometimes called print tip) microarray technology was invented in the early 1990s (ref). The technology has grown tremendously and now there are various flavors of probes and target elements. Target elements can include genes, oligonucleotides, or bacterial artificial chromosomes and new microarrays chips can now contain on the order of a hundred thousand probes. Since the technology is maturing, the cost of analyzing a sample has been steadily decreasing so experiments are now being performed on hundreds of samples, rather than just a handful. Through this explosion of data, it can sometimes be lost that the microarray technology can be viewed as a machine. Although it does not require traditional maintenence, it does require a set of quality control standards to ensure stability. We propose a set of quality control figures to benchmark the lab's microarray facilities and ultimately pinpoint the location of errors in the process. We produce ANOVA tables that show the effects of the systematic variation. We also produce violin plots that demonstrate the quality of hybridization across several different variables in the technology. Although our results can be applied to any print tip microarray setting, our examples will focus mainly on Roswell Park Cancer Institute's (RPCI) array based Comparative Genomic Hybridization (aCGH) facility.

## Materials and Methods

Array based Comparative Genomic Hybridization (aCGH) technology is similar to cDNA arrays and is an extension from conventional CGH that is used to identify and quantify DNA copy number changes across the genome in a single experiment. The advantages of aCGH include high-resolution and high-throuput measurement capability, furthermore, more quantitative analysis of the genomic aberrations.

In aCGH technology, the array elements or targets are laid out on a glass slide and are probed with dye labeled samples. In bacterial artificial chromosome (BAC) aCGH technology the target DNA elements are cloned in a bacterial culture and then physically arrayed in a two-dimensional grid on a chemically modified glass slide.

After creation of the chip, differentially labelled total genomic DNA from a "test" and a "reference" cell population are cohybridized to the BAC clones using blocking DNA (Cot-1) to suppress signals from repetitive sequences. After hybridization, a GenePix Axxon scanner generates two images of the chip at the wavelengths of light corresponding to the two dyes. The images are processed to generate a single number corresponding to each sample for each spot on the chip. For the RPCI facilities, Genepix is currently used to perform the image processing. The resulting ratio of the fluorescent intensities at a location on the chromosomes is approximately proportional to the ratio of the copy numbers of the corresponding DNA sequences in the test and reference genomes.

The RPCI arrayer which generates the aCGH chips for our data consists of 48 pins to transport the samples from 384 well plates (4 96 well plates fused together) to the microscope slide. Each of the chips used in this experiment was created from 51 plates, where each plate was used twice in the spotting process.

## The Arrayer Procedure

For our data, the 48 pins are arranged in a 12 × 4 matrix structure, approximately 4.5 mm on center, so that they transport the probes to the slide so that each pin fills one region or "grid" of the chip. The spots are approximately 80 $\mu$m in diameter, with respective centers 150? $\mu$m apart from each other to ensure no overlap between spots.

The array has the spots laid out in a 116 × 348 array of 40368 spots. More specifically, each of the grids within the array (corresponding to pin number) has dimensions 29 × 29, thus there are 841 spots per grid (pin). The chip's spot locations are labeled consecutively row-wise within each pin, first numbering within Pin 1 (1-841), followed by the spots within 2 (842-1682), etc. Thus, the spot location values range from 1 to 40368.

## Intensities

The data we analyzed is contained in a spreadsheet and gives the intensity readings from the Cy3- and Cy5-labeled probes for each spot, as produced by the image processing software. We let $P_1$ denote the intensity of the Probe 1 signal that was Cy5-labeled for a specific spot. For each spot $i$, we let $log(\frac{P_{1i}}{P_{2i}})$ denote the differential log expression between the two probes for that spot. For the Nowak array facility the sample labelled with Cy5 represents a collection of mRNA from a pool of normal subjects. Hence, in studying various cancerous tumors the standard $log(\frac{P_{1i}}{P_{2i}}) \equiv M$ values can be interpreted as the logarithm of tumor to control values ($log_2 T/C$).

We should note at this point that there are 384/48 = 8 spots per grid per plate. Since 8 × 102 = 816 and each grid has 841 spots, there are $841 - 816 = 25$ blank spots in each grid. Since each plate is used twice, each spot is replicated within a grid on the chip. Put another way, this procedure can produce the expression levels of $40,368/2 - 25 \times 48 = 18984$ BAC clones per chip arranged in a two-dimensional array on the slide that accomodates up to 40368 spots. The remaining $25 \times 48 = 1200$ spot locations remain unused and are, therefore, not considered in this analysis. (Note however, these unused spots may contain valuable information regarding the laser scanner settings).

For each spot we consider the $log_2(\frac{P_{1i}}{P_{2i}}) \equiv M$ as the differential log expression between the two probes for each spot $i$. From Sellers et al. (2004) there are three obvious sources of possible systematic variation which are a consequence of the experimental procedure and do not contribute to differential gene expression. Summarizing each effect, the first is the physical layout on the glass slide; one can imagine that there are spatial effects across the slide (caused, for example, by the way the dye-labeled material is hybridized to the slide) which would manifest as a pattern of row and/or column effects if the data were analyzed as 348 × 116 array. The second source of variation stems from the 384 well plates which are the source of the spots on the glass slide; one can imagine that there are effects which are localized to one (or more) specific plates which would appear as localized effects on the glass slide Also there are potentially effects due to the 384 well plate rows (16 rows) and 384 well plate columns (24 columns). Note that the localization is complicated because of the arrayer procedure described above; recall the complex numbering scheme. The third source is due to the pins themselves. One can easily imagine that the pins vary in size or some other property that causes the observations to vary from quadrant to quadrant on the chip. Equally, one can imagine a serial (in time) correlation among the observations caused by, for example, the pins not being adequately cleaned between successive dips into the wells on the plates. As those authors state: “This is not intended to be an exhaustive list of possible sources of systematic variation, but simply a short list of obvious possibilities.” The key point here, and in all subsequent analysis, is that we assume a random spatial distribution of the probes on the microarray chip.

Numerous papers have been devoted to the general problem of normalization in aCGH arrays, however, few studies have focused on the systematic variation present in microarrays. Through a series of linear models that address the systematic variation above, we can obtain a signature of our lab’s aCGH technology. Note that the use of these linear models is derived from Sellers et al. (2004).

In analyzing the relationship between spots on the microarray chip we consider ANOVA models corresponding to each of the three possible sources of variation described above. We consider models relating the differential log expression with each of the following factor combinations: pin number, plate number, and plate row and column locations. The effects from all of these factors among each of the three initial models was demonstrated to be significant; As advocated in the Sellers et al. (2004), we do not include a variable for the time order in order to avoid the of risk overfitting our data by including such a large number of degrees of freedom.

Of interest is the effect of each of the remaining factors on the data as a complete model. In order to determine the relative effect, we must proceed with caution. This is due to the collinearity that exists between the factors. By computing each relative effect, we can account for it in the normalization schemes. However, we can also use these linear models as a fingerprint for the lab. By examining the significance of the coefficients we can gain an understanding of the operating characteristics on the lab. By knowing the characteristics of the lab when the process is stable, we can quickly and easily pinpoint where the technology faulters.

## Building a complete model

The motivation for model building is presented in **??**. In addition to using these models as a way of removing the systematic variation, we will also use the diagnostics from these models as a fingerprint for the system. It is also important to study the coefficients from these linear models since they can provide metrics, themselves, as to the status of the imaging technology.

From the ANOVA tables, it turns out that when we build a complete model including the pins, plates, plate row and plate column effect that each of these effects are significant. These effects, as being significant are accounted for in the normalization process presented in Miecznikowski et al. (2006). Besides determining the significance of these effects, it is illuminating to examine the coefficients for these variables. Figure 1 shows two sets of figures corresponding to two batches of samples. In Figure 1 (a), we have 4 subgraphs showing the coefficients for each pin, plate, plate row, and plate column. The data for Figure 1 consisted of 24 samples run for a dye swap experiment. If we study the figure corresponding to the plate effect, we see that the trend increases as the plate number increases. This effect indicates that, with respect to plate 1, the higher number plates have smaller $M$ values. Likewise, we see a similar effect with the pin effect. This increasing nature of the pin effect indicates a spatial trend across the samples in this batch. These effects are often common in aCGH experiments and can be accounted for in normalization models (Miecznikowski et al., 2006).

From Figure 1 we also see a marked shift in the plate row coefficients. This effect was not expected and further work is required to understand why there is a clear breakpoint after row 12. Nevertheless this effect can be accounted for in reference Tech report.

Figure 1 (b) shows another set of coefficient figures. In this case, the data corresponded to a batch of 15 samples in the new19k directory. The major difference between this batch and the batch of samples used for Figure 1 (a) is that the BAC cultures were regenerated for this set of experiments. When comparing Figure 1 (a) with Figure 1 (b) we see that the plate coefficients are markedly different between the two batches of experiments. Regrowing the BAC cultures has produced a different set of coefficients for this experiment. Specifically, plates 14, 28, and 36, seem to have the smallest $M$ values. We see that the pin effect is very similar between the two batches. Basically, this indicates that the spatial trend is still present from batch to batch. Again, we see that there is break point in the plate row coefficients after row 12. From this set of experiments, it is clear that within a batch these effects are all significant and must be accounted for in a normalization scheme. Further, these models can change significantly from batch to batch and should be rerun for each batch. These models can even be broken down and run on each sample. By running the models on each chip we can gain insight on the specific operating characteristics for each chip.

By running set of linear models, we can get an understanding for how the various systematic effects influence the output $M$ values from an aCGH experiment. Other summary measures from these experiments can be presented in a statistical summary format giving a measure of the quality control for each batch of samples. From examining the coefficients from a linear model on a pin, plate row, and plate column level, we can explore other types of measures that can be broken down in a similar format. By using violin plots, we can explore pin, plate, plate row, and plate column information over a collection of samples. Specifically we will use violin plots to examine the question of spot quality for each assayed BAC. Spot quality is determined according to the strength of the signal obtained from the laser scanner at that spot location.

Figure 2 shows a "violin plot" for the two batches discussed for Figure 1. Violin plots have numerous references in current statistical literature as a way of combining the information available from local density estimates with the basic summary statistics inherent in standard box plots. By combining the box plot and the density trace on a single plot, comparing the distributions of several variables via violins plots is a great tool for aCGH microarrays (Hintze and Nelson, 1998).

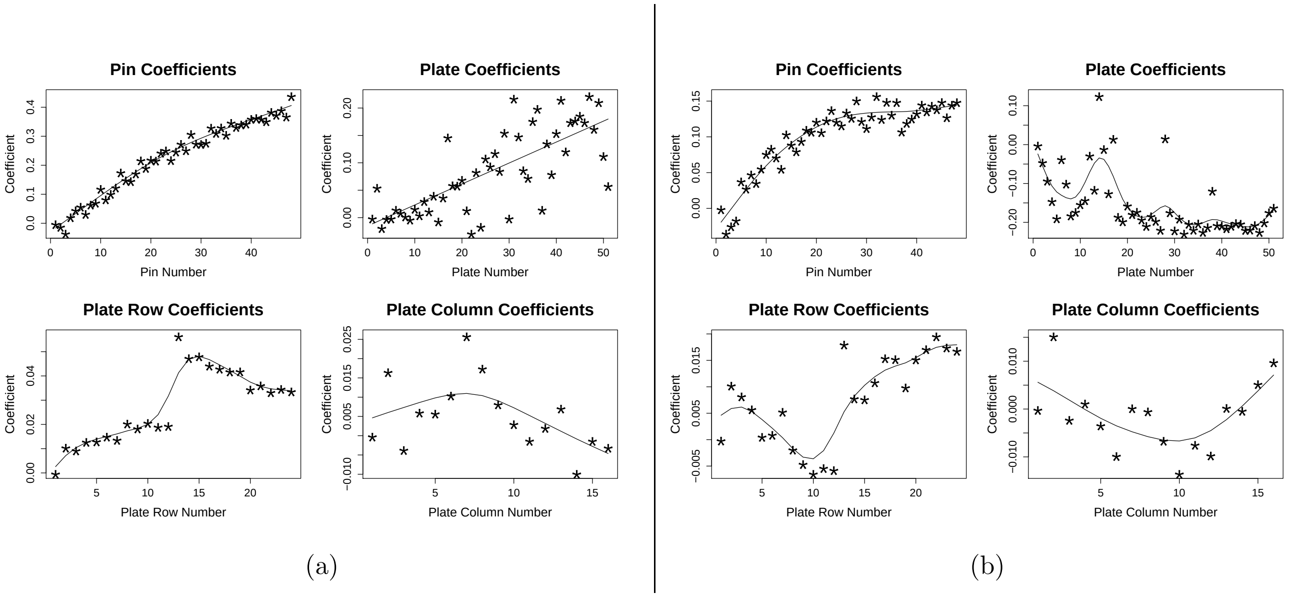
For our aCGH lab, there are potentially three flags used to determine poor quality spots. The BAC spot can be flagged for having a signal-to-noise value. The signal-to-noise value is determined by taking the mean value of the pixels in the signal and dividing them by the standard deviation of the background. If this value is too low ($< 2.5$) then the spot is of poor quality. The spot can be manually flagged by the user to determine poor quality or the spot can be flagged of poor quality because of a dim signal in one of the channels. Dimmness is determined by having a mean signal value under a prescribed cutoff in one of the channels. With this set of flags used for poor spot quality, the distribution of the percentage of poor quality spots was broken down over several different discrete variables in the spotting process.

Figure 2 (a) shows the percentage of BACs from each pin which get flagged with a quality control problem on at least one of the samples in this study (15 samples). This batch of samples is the same batch used for Figure 1 (a). Namely for each pin on each sample, we took the percentage of spots from that pin that had a quality control problem as determined by the image processing software. We then computed that percentage for each sample in the batch, and examined the distribution of percentages for that specific pin via a violin plot. Similarly, Figure 2 (b) shows the percentage of BACs from each plate which get flagged with a quality control problem on at least one sample (24 samples). Figure 2 (b) confirms our suspicions in Figure 1 (b). From Figure 1 (b), our conclusion was that the higher numbered plates consisted of low (or dim) $M$ values, since the coefficients for those plates were large. By examining the violin plots for the plates in Figure 2 (b) we see that a large number of these higher numbered plates ($> 40$) consisted of a large number of spots that were flagged with quality control problems due to dimness.

We can also produce violin plots for the other batch of experiments from the newly generated BAC cultures. From Figure 2 (d) we see that only a few of the plates seem to consistently have quality control problems. This is an agreement with Figure 1 (b) where we see large coefficients for plate 14, 28 and 36. Subsequently, these plates have been removed from the process and the BAC materials for those plates has been regenerated.

## Discussion

We focused on $M$ to build our linear models and ANOVA tables. As expected based on the work in Sellers et al. (2004), the pins, plates, plate row and plate column are all significant factors in predicting the $M$ values. For a specific experiment consisting of multiple chips, we can produce figures showing the coefficients for each variable. These figures can act as fingerprint for the technology and specifically for that experiment. From experiment to experiment the values of these coefficients should be stable, changes in these coefficients indicates that the microarray technology has changed characteristics, possibly due to the process being out of control. As noted, there were remarkable differences in the plate coefficients between the two experiments. This coincides with the target elements being regenerated for the second experiment. This notion of examining variables on a



(a)

(b)

Figure 1: **Linear Models as Quality Control**: (a) The set of linear coefficients for an older BAC experiment involving 24 number of chips. A loess smoothed line is shown for each set of coeffiecients. Note that for the plate row coefficients there is a significant break after plate row 12. Similarly there is a steady linear gradient for the pin coefficients. In (b) an experiment consisting of 15 samples on new chips where each spot probe has been reconsitituted. The coefficients due to pin, plate, plate row and plate column are much less in magnitude than in (a). Correspondingly, the quality of the signal for the experiments in (b) is superior to those in experiment (a).

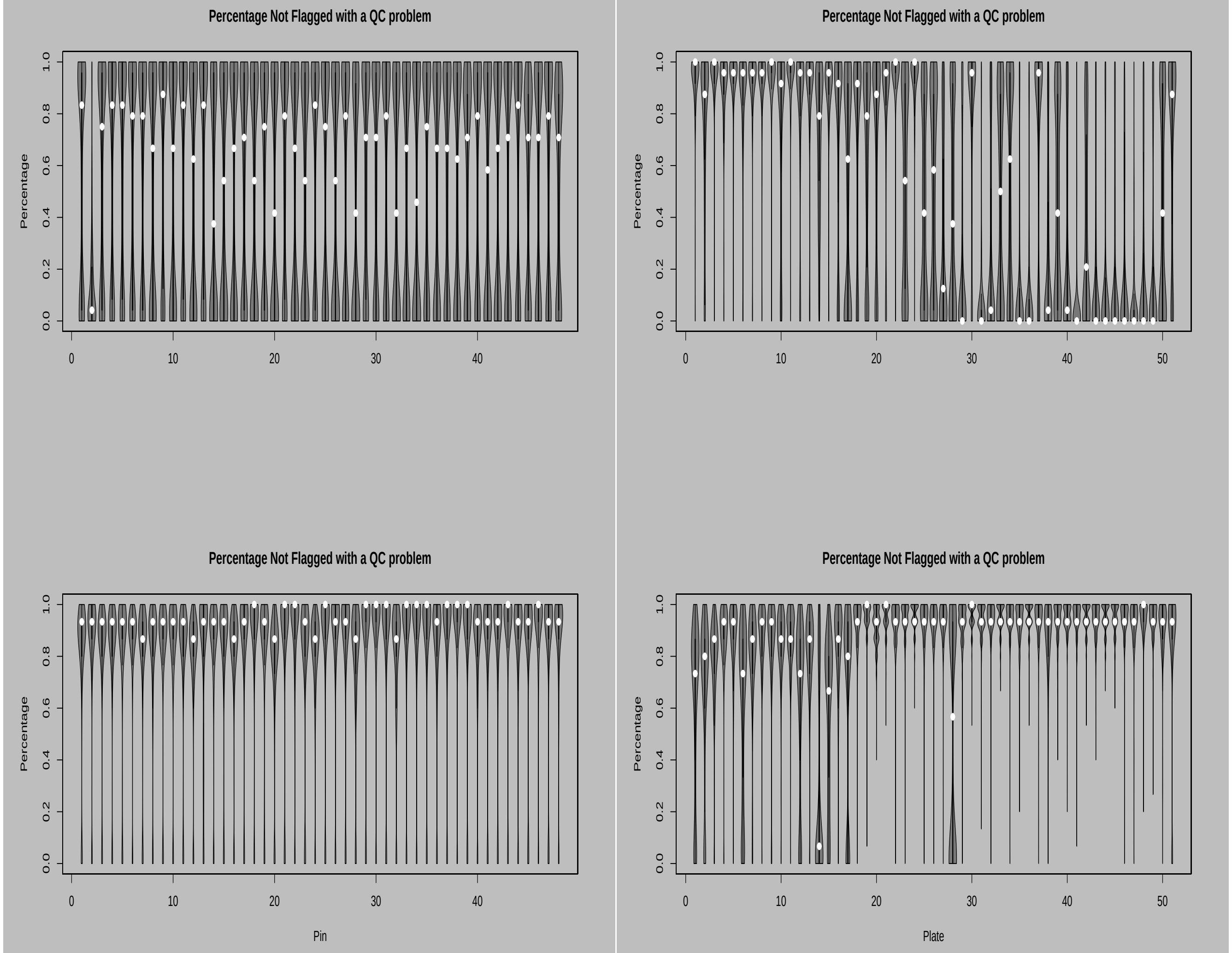
pin, plate, plate row and plate column level can be extended with violin plots regarding the quality control for each pin. From Figure 2 showing the percentage of spots without a quality control problem we can quickly and easily pinpoint the regions where the assay is consistently failing quality control measures. The concept of violin plots for quality control, can further be extended to other commonly reported spot variables such as background mean, background s.d., and other outlier flags.

## Figures

## Supplemental

ANOVA tables from the linear models. Note the models are run on experiments consisting of sets of samples. So a model is built from several chips in two separate cases. The model for case 1 is built from the old19K BAC arrays, the model for case 2 is built from the new19K BAC arrays.

Although this again demonstrates the significance of each of the factors in their effect relating to differential expression levels, this table is still not exact because the degrees of freedom under such a formulation are not



(c)

(d)

Figure 2: **Violin Plots as Quality Control**: Violin plots showing quality control for each 384 well plate used to create the BAC array. There are 51 plates used in the creation of each chip. Each spot is evaluated according to the hybridization quality and assigned a flag. A flag of poor spot quality can be assessed manually, or automatically if a spot has a poor signal to noise value or has a signal below a preassigned cut-off. Often this poor quality is indicated by dim flourescence in one of the channels due to degradation of the target material. In (a) there were 24 samples run and for each sample, the percentage of spots with a quality control flag was recorded. The violin plot in (a) shows the distribution of this percentage over the 24 samples. Note that the median is shown in white. There are several plates, namely the plates past Plate 24 are all suspect in terms of quality of plate wells. (b) represents the new growth of the BAC arrays. There were 15 samples run in this experiment. The BAC cultures have been regenerated and so each plate represents newer BACs. Clearly the problems due to degrading BACs in (a) has been resolved with the new BAC cultures. The two poor quality plates 24 and 48 have subsequently been identified and replaced.

Table 1: Approximate ANOVA table representing effect of all factors on log-T/C expressions. These terms are added sequentially (first to last).Note this ANOVA table corresponds to the new19K experiment.

| | Df | Sum of Sq | Mean Sq | F Value | Pr(F) |
|---|---|---|---|---|---|
| pin | 47 | 1454 | 31 | 304.96 | 2.2e-16 |
| plate | 50 | 3607 | 72 | 819.8 | 2.2e-16 |
| plate row | 23 | 41 | 2 | 20.216 | 2.2e-16 |
| plate col | 15 | 30 | 2 | 22.720 | 2.2e-16 |
| rep | 1 | 10 | 10 | 116.53 | 2.2e-16 |
| Residuals | 605384 | 57721.52 | .0953 | | |
| Grand Tot | 605520 | 62863.52 | | | |

the true degrees of freedom for this model (this is due to the interdependencies of the factors). However, given the procedure by which the model was established, the degrees of freedom listed in Table 1 represent upper bounds on the true degrees of freedom. As a result, the F-statistics will only increase, thus demonstrating an even greater significance of the factor effects. The pin number plate row, plate column, and plate number factor mean squares still demonstrate significance in the model.

## Conclusion

It is important to have adequate quality control measures in place to ensure microarray technology remains within tolerance. This paper demonstrates an example where the quality control samples had degraded in several of the plates and by regenerating the plates we were able to restore quality in each of the scanning channels for the probes. This problem was also evident by the size of the coefficients for linear models built to examine the systematic variation in spotted glass slide microarray technology. By implementing tools that work on a pin, plate, plate row, and plate column level we can pinpoint exactly where errors in the technology may be occurring. Further, by building linear models based on the systematic variation in the spot arraying technology we can account for a significant proportion of the variation in each array.

## References

Hintze, J. L. and Nelson, R. D. (1998), "Violin Plots: A Box Plot Density Trace Synergism," *The American Statistician*, 52, 181–184.

Miecznikowski, J., Gaile, D., Conroy, J., and Nowak, N. (2006), "aCGH Data Normalization Algorithm," *Technical report XXX, University at Buffalo, Department of Biostatistics*, 52, 181–184.

Sellers, K., Miecznikowski, J., and Eddy, W. (2004), "Removal of Systematic Variation in Genetic Microarray Data," *Technical report 779, Carnegie Mellon University.*