Conference on Applied Statistics in Agriculture 2006 - 18th Annual Conference Proceedings

A VISUAL AID FOR STATISTICIANS AND MOLECULAR BIOLOGISTS WORKING WITH MICROARRAY EXPERIMENTS

Deborah L. Boykin

Earl W. Taliercio

Rowena Y. Kelley

W. Paul Williams

See next page for additional authors

Recommended Citation

Boykin, Deborah L.; Taliercio, Earl W.; Kelley, Rowena Y.; and Williams, W. Paul (2006). "A VISUAL AID FOR STATISTICIANS AND MOLECULAR BIOLOGISTS WORKING WITH MICROARRAY EXPERIMENTS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1119

Author Information

Deborah L. Boykin, Earl W. Taliercio, Rowena Y. Kelley, and W. Paul Williams

A Visual Aid for Statisticians and Molecular Biologists working with Microarray Experiments

Deborah L. Boykin[1], Earl W. Taliercio[2], Rowena Y. Kelley[3] and W. Paul Williams[4]

[1]USDA-ARS Mid South Area Statistics Office, Stoneville, MS 38776
[2]USDA-ARS Mid South Area Cotton Genetics Unit, Stoneville, MS 38776
[3]Department of Plant and Soil Sciences, Mississippi State University, MS 39762
[4]USDA-ARS Corn Host Plant Resistance Research Unit, Mississippi State, MS 39762

Abstract

The use of microarrays to measure the expression of large numbers of genes simultaneously is increasing in agriculture research. Statisticians are expected to help biologists analyze these large data sets to identify biologically important genes that are differentially regulated in the samples under investigation. However, molecular biologists are often unfamiliar with the statistical methods used to analyze microarrays. Presented here are methods developed to graphically represent microarray data and various types of errors commonly associated with microarrays to help visualize sources of error. Two case studies were used. In case study one, genes differentially regulated when two corn lines, one resistant and one sensitive, were treated with *Aspergillus flavus* isolate NRRL 3357 or left untreated were investigated. Analyses and images showing 3 types of variation are shown. Genes were ranked according to fold change and re-ranked after adjusting for potential sources of error. In case two, cotton genes differentially regulated in 1-day-old fiber compared to whole ovules or older fibers were investigated. Data and sources of error were imaged as described for case one and genes with significant changes in gene expression were identified.

Introduction

Microarray experiments do not determine changes in gene expression as accurately as those utilizing other technology, such as RNA blot analyses or real time PCR quantification (Chuaqui et al., 2002). The major benefit from the use of microarray experiments is the ability to rapidly screen thousands of genes simultaneously in one experiment. This is a selection method that identifies target genes and their regulatory areas on a genome scale which can be further evaluated by the more accurate methods mentioned above. Statisticians and molecular biologist both recognize the need for more accuracy in these Microarray studies. Accuracy can be improved through experimental designs and proper analysis of Microarray data.

In the past, molecular biologist performed experimentation utilizing technology that produced error-free dataset by eliminating possible sources of error. Therefore, many molecular biologists do not have a strong statistical background. With Microarray technology, data are not error-free. Statistical methods are needed to control and measure biological variation and hypothesis testing is needed to address objectives of the study. Also, complexity of biological issues involved in the experiment hinders the statistician's ability to comprehend issues related to the experimental design. Communication between scientist and statisticians is often hindered by a lack of knowledge in the other's field of expertise.

The Microarray experiment is a complex process with many error sources. A general overview of this process includes (Allison et al., 2006):

- Planning - Define objectives and experimental design
- Biological Process - Collect samples needed to test hypothesis of interest and extract RNA
- Data Generation - Target synthesis, labeling, hybridization and acquiring an image of the array
- Data Processing - Process image making calibrations to achieve best image quality and quantify fluorescent signals resulting in numerical values for spots on the arrays
- Data analysis - Data normalization, filtering and statistical testing
- Interpretation - Identify genes with statistical and biological significance and detect patterns in genes.

Much progress has been made in the last decade toward increasing precision of microarray experiments by improvements in technology and experimental design. The terms “biological rep” and “technical rep” are often used to divide the error sources into two groups. Two arrays are technical replications when they use the same target sample under the same labeling and hybridization conditions (i.e. sub sample). Two arrays are biological replications when they use different samples representing true replications of the treatment effects under study. Improvement in microarray technology in the past decade has resulted in technical variation being less of a concern compared with biological variance (Clarke et al., 2006).

It is important that the biologist understand the statistical model used to describe the data and test hypothesis. This includes the error sources being described by the model. This paper presents a visual aid for observing error as a spatial pattern on a microarray slide. With these graphs, statisticians can illustrate to a scientist how a statistical model can explain and control variability in a study and identify spatial patterns of variability within an array. By identifying spatial patterns of error, statisticians can better understand technical sources of errors in microarrays. Better communication between biologists and statisticians will result in better statistical designs and /or statistical models describing the data.

Case Study I - Aflatoxin in Corn

Aflatoxin is a naturally occurring toxic chemical by-product resulting from the growth of a fungus, *A. flavus*, on corn and other crops such as peanuts and cotton. A microarray experiment was used to identify the potential genes that controlled resistance to *A. flavus*. Differential expression of genes from tissues inoculated with *A. flavus* was measured by a ratio of gene expressions when *A. flavus* was present, versus gene expressions when *A. flavus* was not present (or by the difference of log transformed values). Comparing this differential expression between resistant and susceptible corn lines will aid in identifying genes that control resistance to aflatoxin.

The experiment consisted of a field design containing 2 varieties of corn (resistant and susceptible to aflatoxin) planted in a randomized complete block design with 3 replications. This was followed by applying an additional treatment effect in a lab using samples from the field plots. cDNA microarrays were used to

evaluate the effect *A. flavus* by dividing plant material from each field plot sample and inoculating part of each sample with Aflatoxin. The overall experimental design was a split plot with 2 varieties as the main unit treatment and 2 sub unit treatments for inoculated or un-inoculated plant material taken from field plots.

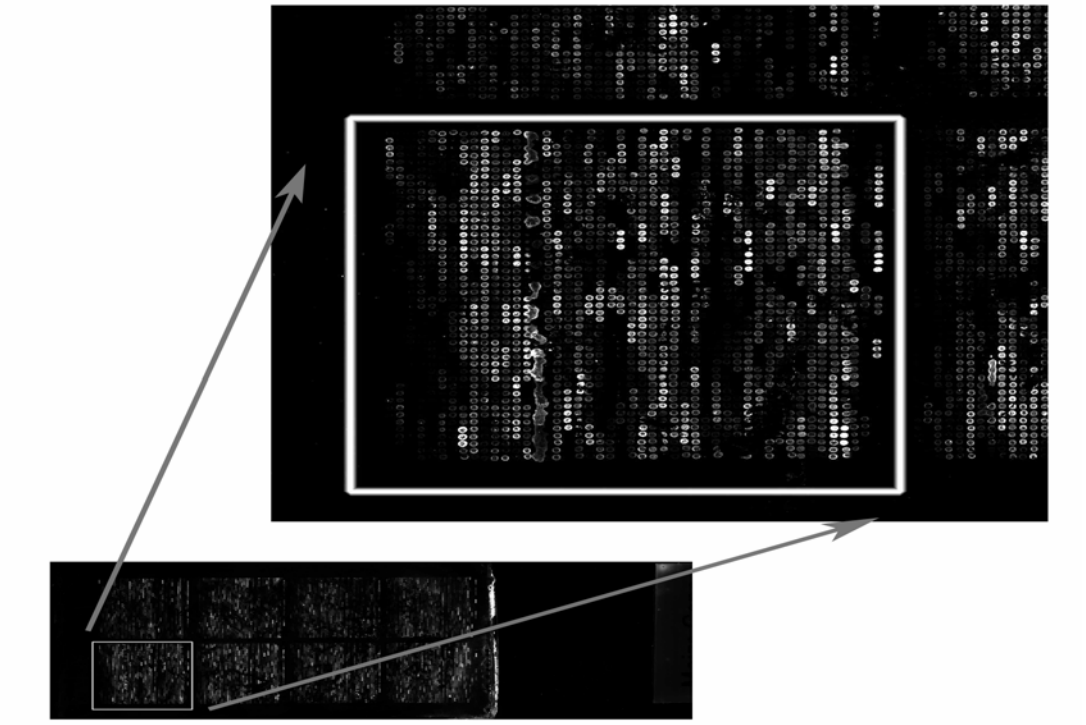

Figure 1 - Scanned image of microarray slide with zoom on block of spots in bottom left corner

Spotted cDNA microarrays consist of samples of DNA clones with known sequence content, spotted and immobilized on to a glass slide (the microarray). Microarray slides used in this study were Unigene 1-1.05 cDNA containing 5065 EST contigs (represented as spots on the microarray) from libraries derived from immature leaf, endosperm, immature ear, and roots of corn, representing approximately 4000 genes (Pontius et al.,2003). Pools of mRNA from the cell populations for each sample were purified, reverse-transcribed into cDNA, and labeled with one of two fluorescent dyes, which are referred to as “red” and “green” (Kerr et al.,2000). Material inoculated with Aflatoxin was labeled with one color and the un-inoculated material was labeled with the other color. Two pools of differentially labeled cDNA from the field plot samples were combined and applied to a microarray. A scanner measured laser induced fluorescents for each dye. Two measurements of fluorescent intensities, one for each sub-unit treatment, were taken from each spot on the microarrays. These intensities represent expressions of the gene represented by a spot for a given treatment condition.

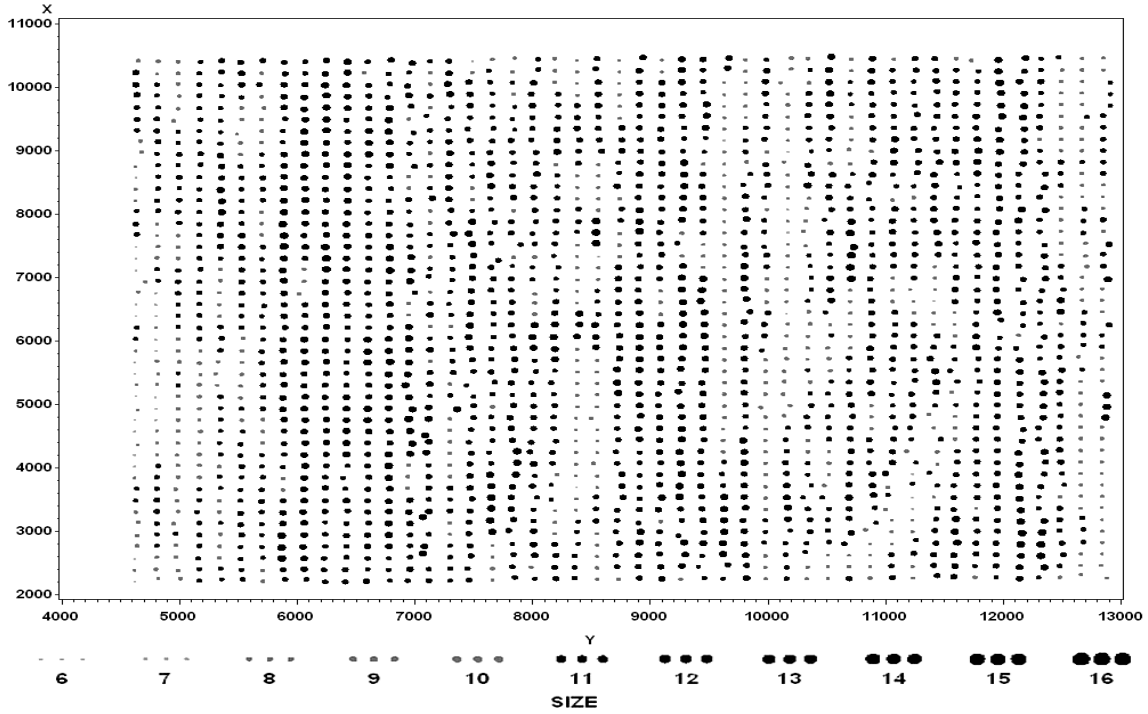

Figure 2 – Digital representation of microarray block shown inf Figure 1

The microarray was divided into 8 blocks (2x4) as shown in figure 1, this blocking has no statistical purpose but facilitates in gene (spot) identification. Two measurements of florescent intensities were taken from each spot. Figure 1 shows the actual intensity of the florescent dye. Figure 2 shows a digital representation of measured values of intensities for this same block, and figure 3 is a digital representation of the entire microarray slide. The scale used in figures 1b and 1c was calculated by taking the base 2 log of the measured intensity, rounding this value to the nearest whole number (ranging from 2 to 16) and plotting different size circles for each response. If genes were randomly placed on the microarray slide, then one should not see patterns of lighter or darker areas.

Each Microarray was a “biological replication” of the treatments under study with inoculated material tagged with red dye and un-inoculated material tagged with green dye. In order to measure bias in sub-unit treatments caused by the dye, an additional microarray was used for rep 2 for each variety and the dyes used for each sub unit treatment were swapped. This additional slide is referred to as a dye swap (Rosenzweig *et al*., 2004) and is considered a “technical replication”. Each gene was spotted three times in three adjacent spots on the microarray. This is also considered another level of “technical replication”.

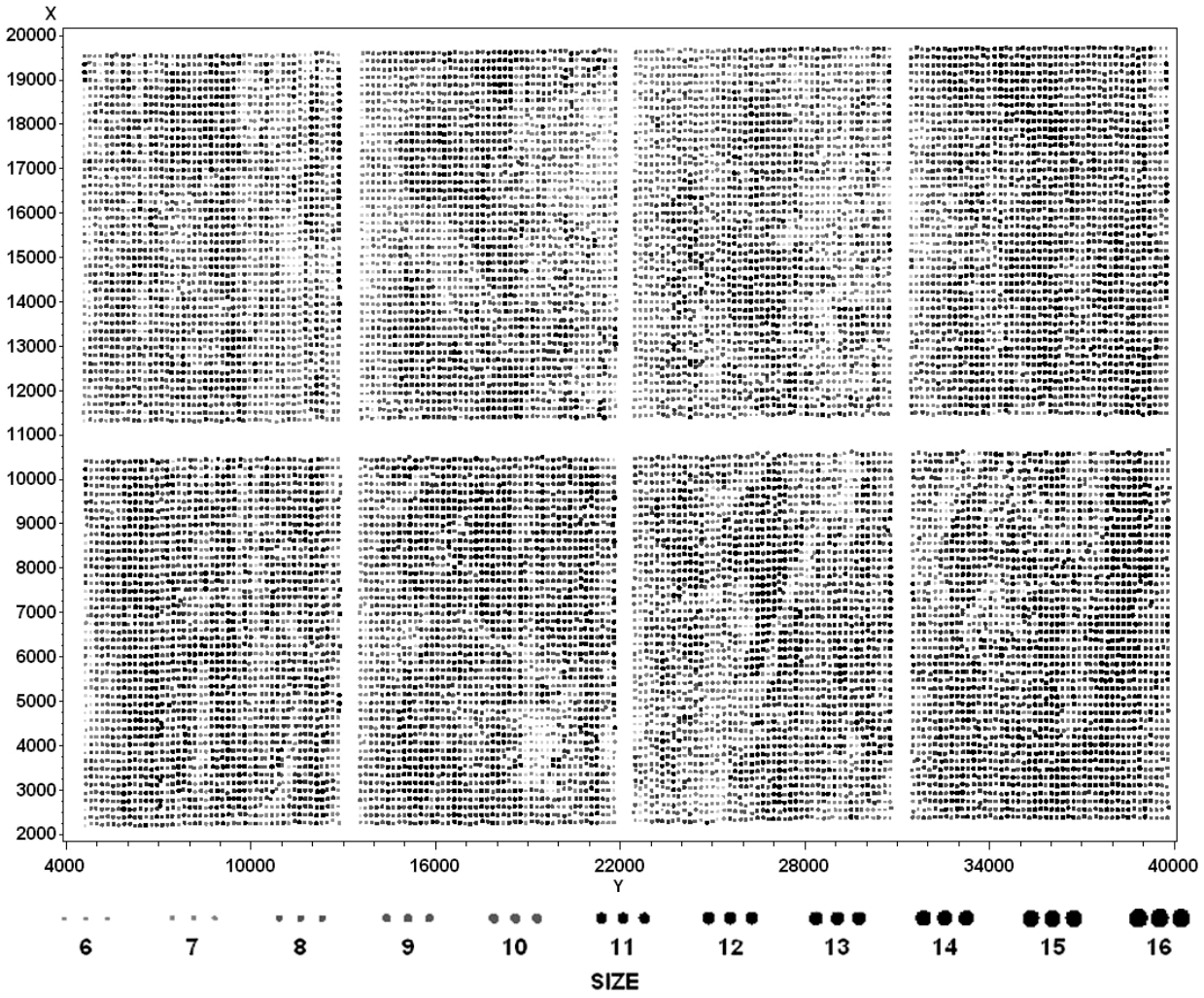

Figure 3 – Digital representation of entrie microarray slide

Separate analysis of variance was performed on data for each gene. DNA microarrays are designed to compare relative transcript abundance (measured by fluorescent intensities) between different samples and should not be used to compare transcript abundance between different genes within the sample (Hekstra et al., 2003). The objective of the study is to identifying genes that control resistance to aflatoxin. The fixed effects in the analysis of variance are variety, inoculation, and variety X inoculation. Intensity differences expressed by genes when exposed to aflatoxin ($\log_2$ intensity for inoculated – $\log_2$ intensity for un-inoculated) is referred to as a differential expression and will be a ratio if converted back to original scale. The fixed effect of variety x inoculation will compare the differential expression between resistant and susceptible corn varieties and will be the main hypothesis of interest.

The biggest challenge is to construct a statistical model to control and describe the various sources of variability encountered in the study and to adjust for bias due to dye in the sub unit treatments. The goal of this paper is to use the graphs to show how the statistical models can control and measure variability.

Therefore, the models used for the paper serve this purpose and are not necessarily meant to be the best model to address the objectives of the study.

Analysis of variance (ANVOA) was performed using Proc Mixed (SAS, 2004). There was no preliminary "data cleansing" or normalization (Quackenbush, 2002). Analysis was performed on $\log_2$ intensity measurements for each gene. Data for 1 gene consisted of 3 spots array x 4 arrays per variety x 2 varieties with 2 measured intensity values (inoculated and un-inoculated) per spot for a total of 48 observations per gene or EST contigs. There were over 5000 EST's and therefore there were over 5000 analysis. ANOVA should included fixed effects: genotype, inoculation, genotype*inoculation and dye; and random effects: rep, rep*genotype, spot (rep genotype) and rep*inoculation (genotype). The residual for this analysis was spot*inoculation (rep genotype), which is sub sampling error for the three spots per slide. A fixed effect for dye was included since rep 2 required an additional microarray that swapped dyes used for inoculation treatments. Including this term in the analysis made an adjustment for a dye effect between inoculation treatments.

To help scientists better understand the concepts of multiple sources of variability and the use of a statistical model to control variability, the same type of graph used in figure 3 was also used to display residuals from analysis of variance. Residuals were created with three different analyses in order to display three types of errors. (1) Simple within array variability - Residuals for within array variability were obtained without adjusting for an overall effect of spot [i.e. spot (rep genotype) was not included as a random effect]. (2) Adjusted within array variability - Residuals for within array variability were obtained after adjusting for spot. (3) Residuals for between array variability were obtained from an analysis on data after averaging over subsamples. See Appendix for the SAS code used for analysis.

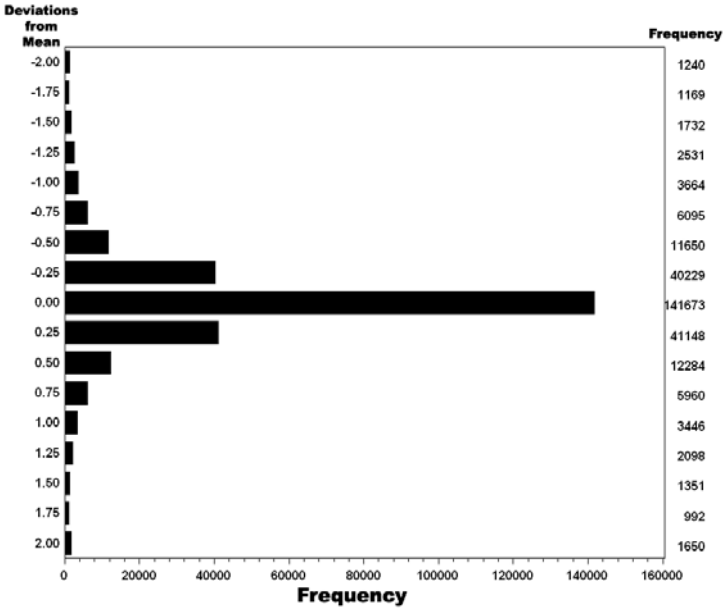

Figure 4 - Histogram of residuals measuring simple spot to spot variability within an array.

In the first case, residuals represented simple variation of 3 spots for the same inoculation treatment. The residuals from all analysis for each gene were pooled and Figure 4 shows a histogram of these residuals. Figure 3 is a graph showing the intensity values relative to where they occur on an array. This same style of graph is used to show where largest within slide variability occurs on an array. Residuals for the non-inoculated treatment of one microarray are shown in figure 5. Larger gray spots represent large negative residuals and larger black spots represent large positive residuals. A large residual based on simple within slide variability occurs with the 3 spots for the same gene on the same array does not agree. From the graph shown in Figure 5, there are

obvious areas containing large residuals. Patterns found in residuals were usually an indication of variability in microarray technology. By identifying the sources of variability found in these graphs, statisticians can better understand the microarray process and possible sources of variability that might be controlled by experimental design.

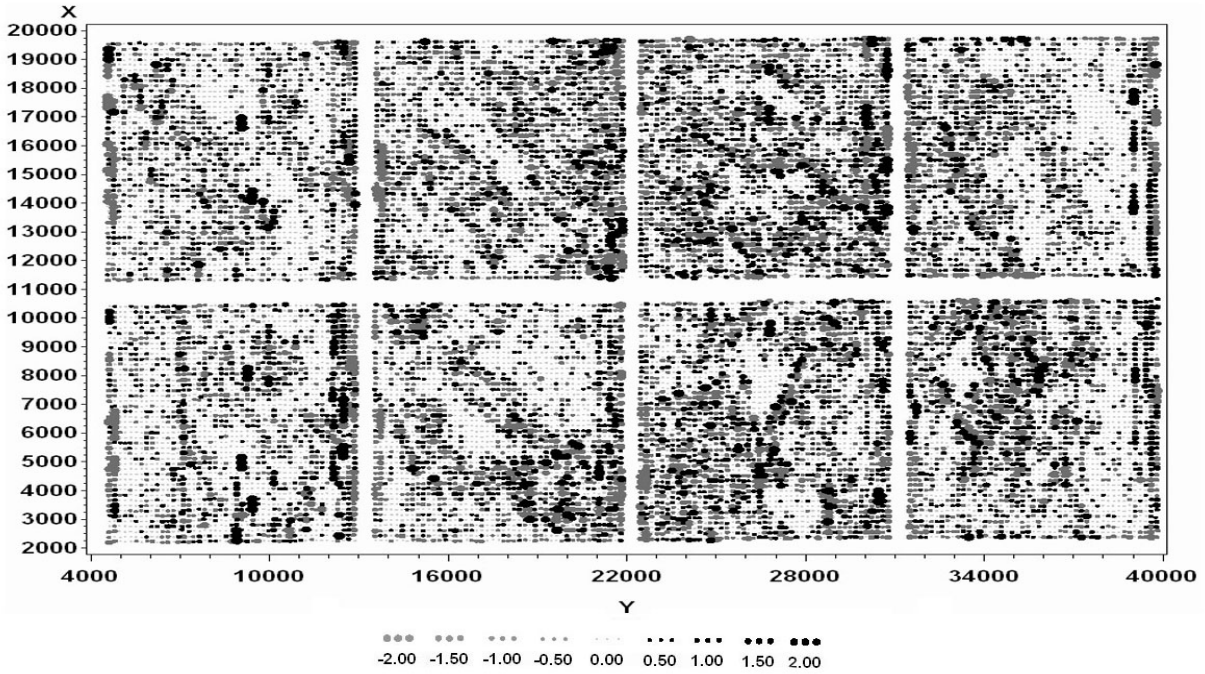

Figure 5 – Above is a digital representation of residuals representing simple within array variability

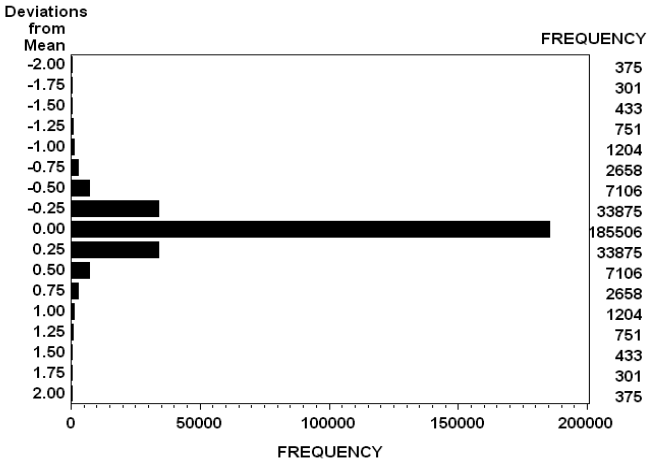

Figure 6 - Histogram of residuals measuring adjusted spot to spot variability within an array.

Simple within array variability shown in figures 4 and 5 ignores the fact that subunit treatments are paired, i.e. each spot contains an intensity reading for inoculated and non-inoculated treatments. Therefore, real subsampling error should be adjusted for an overall spot or block effect by including spot (rep genotype) as a random effect. Residuals from the second analysis were adjusted for this spot effect. This reduction in subsampling error may effectively remove error that scanning related software attempted to remove when using "data cleansing" methods such as background subtraction (Yang *et al*., 2002). Figures 6 and 7 show within array variability for the same slide shown in figures 4 and 5, after adjusting for spot as a block effect. Improvement in precision by accounting for a

block experimental design effect is shown by comparing histograms in figures 4 and 6 and spatial variability within the microarray in figures 5 and 7. This simple comparison should help show scientist advantage of using a statistical model to control variability.

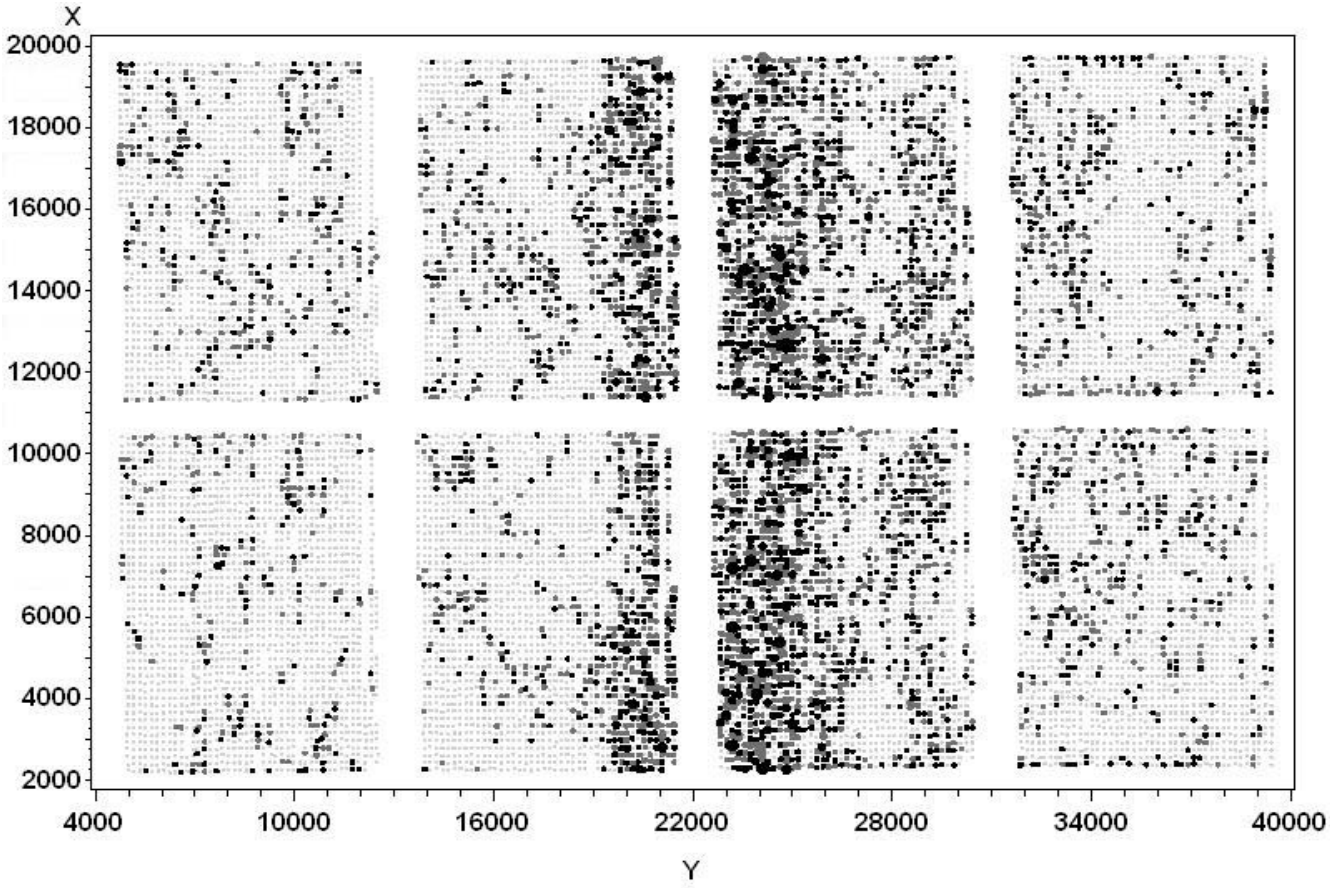

Figure 7 – Above is a digital representation of residuals representing adjusted within array

For the third analysis, residuals representing between slide variability are shown in figures 8 and 9. These residuals indicate the overall error involved in testing for an inoculation X genotype interaction and were obtained from analysis of data averaged over the three sub samples. The histogram shown in figure 8 can be compared to figure 6 to illustrate between slide variability is greater than within slide variability. Also, note that spatial variability shown in figure 9 contains the same residual for all three-sub samples of the same gene.

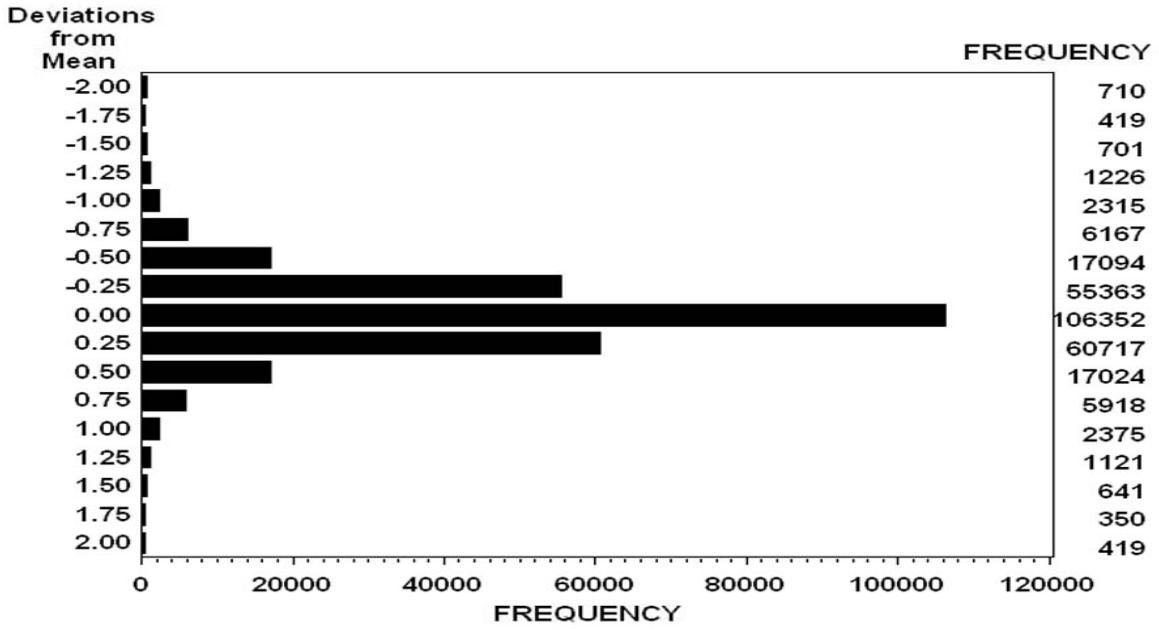

Figure 8 - Histogram of residuals measuring between slide variability.

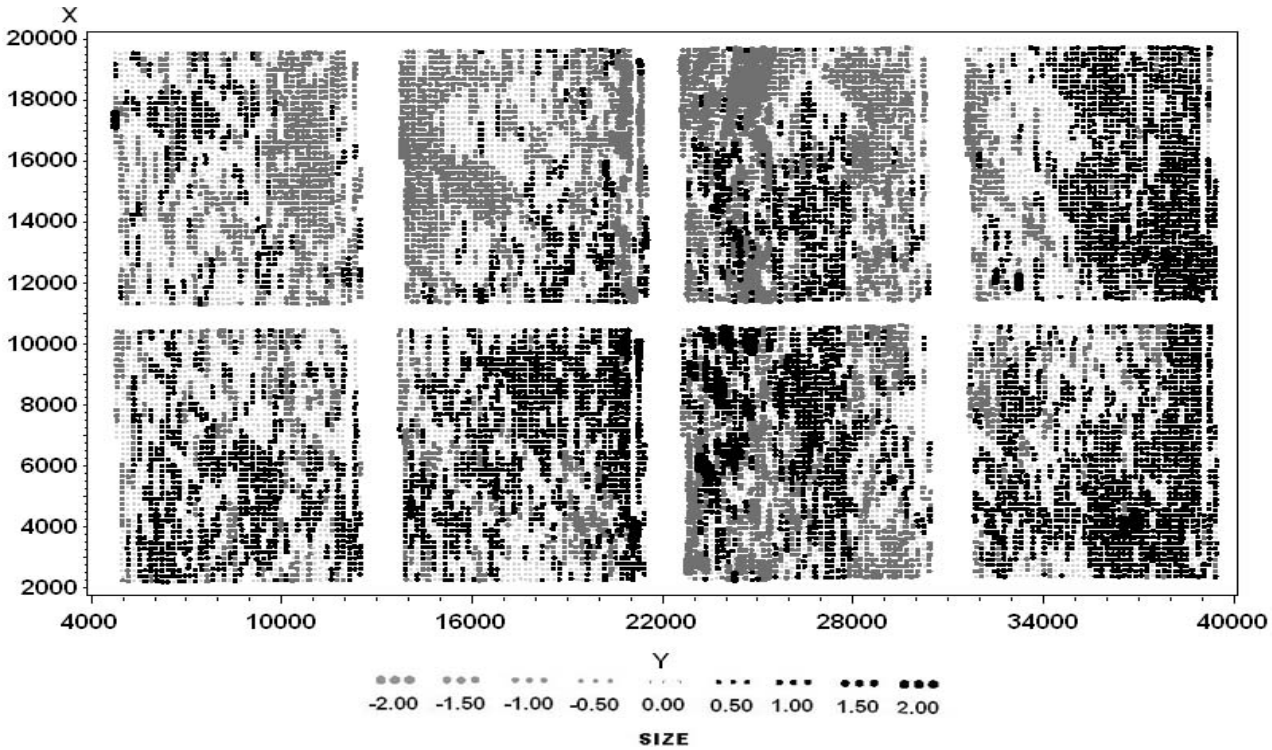

Figure 9 - Above is a digital representation of residuals based on between array variability.

Case Study II - Age of Cotton Fiber

The purpose of this experiment was to identify cotton genes that were differentially expressed in 1-day-old fiber (treatment 2) compared to ovules (treatment 1) and 10-day-old fiber (treatment 3). Experimental design consisted of three treatments taken two at a time. Treatments 1 and 2 were used for three microarray slides and treatments 2 and 3 were used for three slides. One slide of each pair was a dye swap. All microarrays were "biological" replications including the dye swap microarray.

The microarray slides used in the study were purchased from Agilent Palo Alto, CA, and had over 10,000 genes (Agilent http://www.chem.agilent.com/). The data from analyses of the cotton microarrays have been deposited in GEO under accession number GSE6855. This accession numbers describes the platform of the microarrays including the number of genes represented, how each gene is replicated and the specific sequences representing each gene. This accession number also gives detailed information on the hybridization and links the data for each hybridization to the platform.

Each gene had 2 to 5 spots per slide and each spot represented a different oligo (40-mer) that putatively represented the gene. Spots were not divided into blocks as in Case Study I. The microarray shown in figure 10a represents a grid of 206 x 430 spots. This microarray slide was much denser than the Case Study I slide. Increased density made it more difficult to view the entire slide with enough precision to detect spatial patterns. A blow up of the lower right hand corner, shown in figure 10b, illustrates an offset between rows and columns and a pattern of 5 bright spots in the corner. This is one of many control areas on this microarray. Figure 10c represented the digital measurements of the expression values shown in figure 10b. Figure 10d shows digital values for the entire slide expressed as rounded values for log2 (gene expression) for one of the treatments on the slide. The log2 values varied from 7 to 13.

Analysis of variance was performed on the data for each gene, no preliminary "data cleansing" or normalization was used. The analysis performed for each gene included fixed effects for treatment and dye and random effects for array, array*treatment and oligo (array treatment). The residual represented subsampling error for multiple oligos per slide for each gene. Residuals were created with three different analyses in order to display three types of errors. (1) Residuals for simple within array variability were obtained without adjusting for a spot effect [i.e. oligo (array treatment) was not included as a random effect]. (2) Residuals for within array variability were obtained after adjusting for the block type effect of spot. (3) Residuals for between array variability were obtained from an analysis on data after averaging oligos within a slide. See Appendix for the SAS code used for analysis.

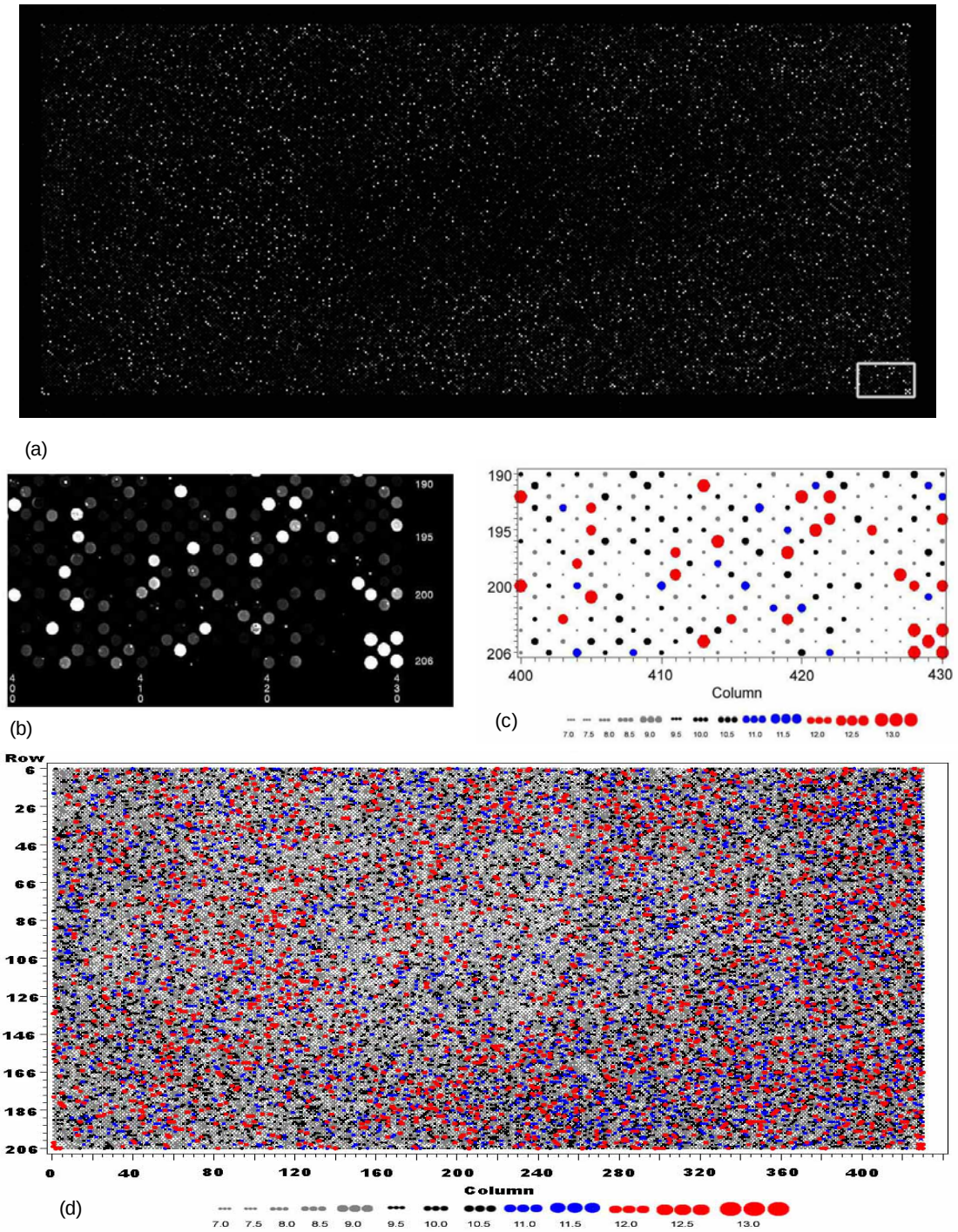

Figure 10 – (a) Scanned image of microarray slide (b) Zoom in on bottom right corner (c) Digital representation of bottom right corner (d) Digital representation of whole slide.

Figure 11 contains graphs of different sources of variability involved in this study. Subsampling error involved differences between the multiple spots or oligos on an array for the same gene. Two treatments were applied to the same spot, (spot being a block effect), which was similar to the first case study. Figures 11a and 11d contain results from the first analysis showing within array variability unadjusted for a spot effect. Figures 11b and 11e illustrate how much lower subsampling errors became after adjusting for overall differences between spots. Finally, figures 11c and 11f illustrate the overall error (between array variability) involved in testing treatment effects.

Unlike Case Study I (figures 6 and 8), when histograms in figures 11c and 11b are compared, one can see that total variance was similar to subsampling variance. In this case study, subsampling error represented more of the total variability involved in comparing treatments. This may have been caused by differences in technology or methodology, such as the subsamples in Case Study II being randomly placed on the array rather than adjacent as they were in Case Study I.

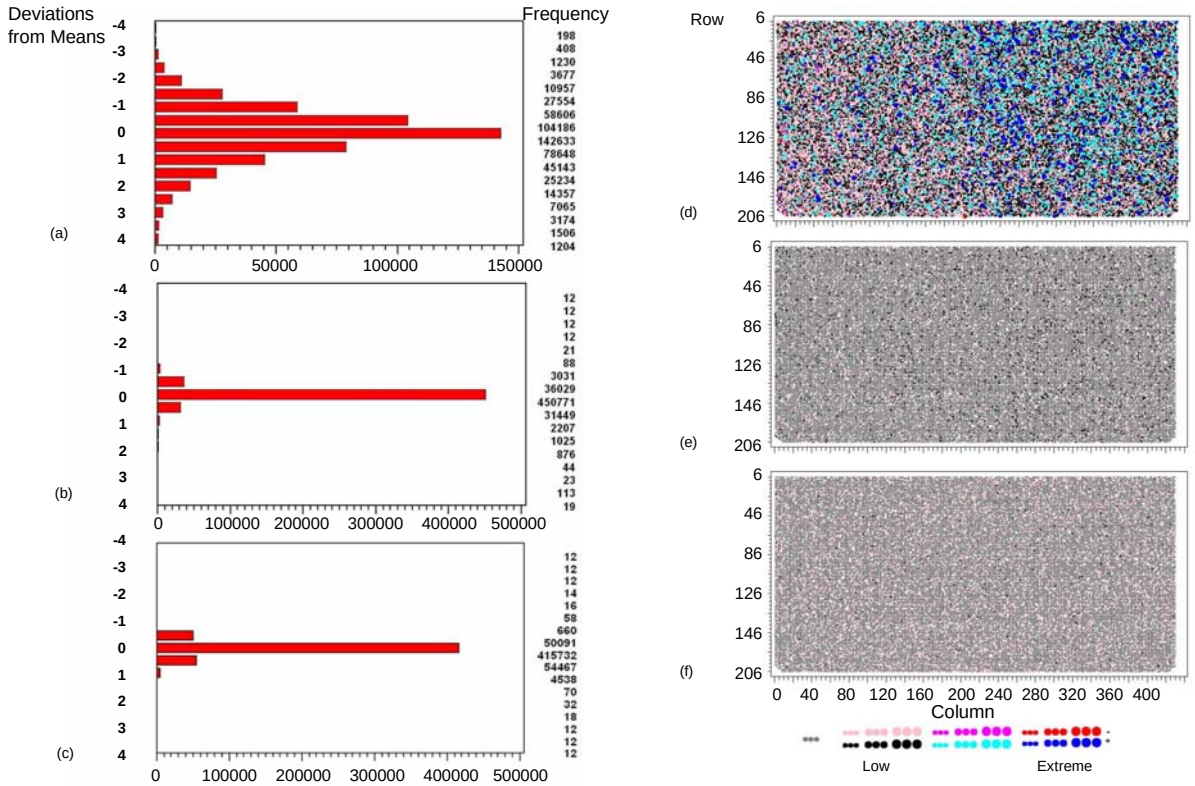

Figure 11 – (a) & (d) Residuals for simple within array variability; (b) & (e) Residuals for adjusted within array variability; (c) & (f) Residuals for between slide variability.

Results of Analysis

The goals of these analyses were to control error, increase precision, and address questions of interest. Statisticians, with an understanding of the technical process of microarrays, are working on experimental designs and statistical models to control error and avoid bias when constructing tests to address these questions.

In case study 1: What genes behave differently between the Resistant and Susceptible genotypes when exposed to aflatoxin?

Ho: Log_2 (Resistant Inoculated) - Log_2 (Resistant Non-inoculated) = Log_2 (Susceptible Inoculated) - Log_2 (Susceptible Non-inoculated)
or,

Ho: RI/RU = SI/SU

F-test for genotype x inoculation treatment performed on the log2 transformed intensity values will address this question. Since this f-test was performed on each gene, there were over 4000 f-tests for this study. To summarize these results, genes were ranked based on significance of the F-test. In addition to this ranking, it was important to factor in the biological significance and measure false discovery rates; however these two issues will not be addressed in this paper.

It is obvious from figures 6 and 7 that a spatial pattern was in the residual that was stronger on some arrays than others. This indicated hot spots of errors on the microarray slides. The goal of this study was to pick a candidate set of genes that affect resistance to *A. flavus* and then conduct more accurate testing on the candidate gene set. Therefore, a type I error (incorrectly choosing a gene) was not as important as making a type II error (not choosing a gene for future studies). If a hot spot caused a large enough error or bias in the genes response, then it likely would be a Type II error.

To avoid Type II errors, the spot having the largest “within arraye variability” was identified for each gene in the first analysis and then omitted from the data before being reanalyzed. Genes were again ranked based on the f-test from this analysis. Table 1 compares the ranking results for both analyses where the rows divide the data based on the first analysis using all the data and the columns divide the data based on the second analysis omitting an outlier spot for each gene. In column 1/row 1, we found 6 of the 10 highest ranked genes in the second analysis were ranked in the top 100 from the first analysis. In column 1/last row, we also found 1 gene of the 10 highest ranked genes in the second analysis was ranked 2000 or more in the first analysis. Here is an example where 1 spot resulted in this gene being overlooked. After omitting this “outlier” spot, the gene would definitely be considered a candidate for further testing. A similar procedure was performed by omitting a whole array for each gene based on the residual describing between-slide-variability. A comparison of the results of this third analysis with the original analysis is shown in Table2.

Table 1 - Rank Genes Based on Significance of Genotype X Inoculation Interaction. Rows Show Ranking from analysis using all the data with row 1 being the top 100. Columns show ranking from analysis after omitting 1 observation per gene based on within slide variability and column 1 represent the 10 highest ranked genes.

Original	New Rank (Omit within Slide Variability Outlier)						
Rank	0 - 10	11 - 100	101- 200	201- 300	301- 400	401- 500	501- 600
0 -100	6	58	16	9	3	.	3
101 -200	1	15	33	25	12	2	3
201 -300	.	7	21	18	18	7	8
301 -400	.	3	11	15	23	17	11
401 -500	1	2	5	13	8	22	14
501 -600	.	1	.	5	5	17	18
600 -700	1	.	4	.	6	9	12
700 -800	.	2	.	2	7	3	2
800 -900	.	.	2	1	4	4	9
900-1000	.	1	1	3	.	5	3
1000-1100	.	.	.	1	3	3	1
1100-1200	.	.	.	1	2	.	1
1200-1300	.	.	.	.	2	.	1
1300-1400	.	.	2	1	1	1	2
1400-1500	.	1	1	1	.	.	.
1500-1600	.	.	1	.	1	3	1
1600-1700	.	.	.	.	.	.	.
1700-1800	.	.	.	.	1	1	.
1800-1900	.	.	.	1	.	1	2
1900-2000	1	.	3	4	4	5	9
Total	10	90	100	100	100	100	100

Table 2 - Rank Genes Based on Significance of Genotype X Inoculation Interaction. Rows Show Ranking from analysis using all the data with row 1 being the top 100. Columns show ranking from analysis after omitting 1 slide per gene based on between slide variability and column 1 represent the 10 highest ranked genes.

Original	New Rank(Omit within Slide Variability Outlier)						
Rank	0-10	11-100	101-200	201-300	301-400	401-500	501-600
0 -100	7	40	26	12	7	4	3
101 -200	1	9	12	14	18	9	6
201 -300	.	5	14	8	10	10	3
301 -400	1	3	8	9	9	11	2
401 -500	1	2	5	9	6	9	6
501 -600	.	1	5	6	8	7	5
600 -700	.	3	5	6	5	2	5
700 -800	.	3	2	6	2	5	4
800 -900	.	4	2	6	1	.	3
900-1000	.	1	2	5	5	5	3
>1000	0	19	19	19	29	38	60

In case study 2: Identify genes that were differentially expressed in 1-day-old fiber compared to ovules and older fiber. There were two hypothesis of interest:

Ho: Log_2 (1 day fiber) = Log2 (ovules) Ho: Log_2 (1 day fiber) = Log2 (10 day fiber)
or,

Ho: 1 day fiber/ovules = 1 Ho: 1 day fiber/10 day fiber = 1

F-test for Treatment 1 vs. Treatment 2 and Treatment 2 vs. Treatment 3, performed on the log2 transformed intensity values, addressed these hypotheses. Since this f-test was performed on each gene, there were over 10,000 f-tests for this study. To summarize these results, genes were ranked based on significance of the F-test. As mentioned above, it was also important to factor in biological significance and measure false discovery rates. Tables 3 and 4 summarize results of this f-test and attempts to incorporate information about biological significance. Up-regulated (a gene producing more mRNA in one treatment and compared to the other treatment) indicated biological significance because the Ratio of 1day fiber to ovules was greater than 2. Down regulated indicated the reverse ratio was greater than 2. A ratio of 2 was arbitrarily considered a threshold of biological importance. The scientist, in charge of this study, when summarizing results of the f-test, constructed Table 3. The scientist chose to show biological significance in Table 3 as the main number and in parenthesis the number that was biologically and statistically significance. After consultation with the statistician Table 4 was prepared with the number of genes significantly altered in expression ($P<0.05$) in row 1. Of these statistically significant genes, rows 2 and 3 indicate the number that exceeded the 2-fold threshold considered biologically important. This change in perspective considered genes with consistent smaller fold changes and omitted inconsistent larger fold-changes.

Table 3- Results of Case Study II indicating number of genes with fold change greater than 2, which is the biologically important threshold.

	1 day fiber vs. ovules	1 day fiber vs. 10 day fiber	Both
up-regulated	471 (368)	176 (164)	65 (54)
down-regulated	244 (268)	448 (376)	44 (36)
() indicates number significant at P<0.05			

Table 4- Results of Case Study II indicating number of statistically significant genes (P<0.05).

	1 day fiber vs. ovules	1 day fiber vs. 10 day fiber	Both
Total	1877	1491	512
up-regulated	392	182	54
down-regulated	248	390	36

Conclusion

There are many sources of variability in studies involving microarrays. In the first case study, the error involved in the technology was large. It is interesting to note that when the "spot" effect was included in the model a spatial pattern emerged in the residuals (compare figures 5 and 7). This indicates that once error is somewhat under control, one can visualize other potential sources of variability. The spatial pattern shown if Figure 7 indicates something other than random variability and if this cause can be identified, then perhaps something can be added to the statistical model to better control or measure this error. Other patterns emerged in the “between array” graph shown in Figure 9. In this graph there is a technical problem that was corrected after seeing the graph.

The technical and biological error is controlled more in the second case study. Also, there is no obvious spatial pattern in residuals shown in figure 11. The spots on the microarrays used in the 2nd study are a denser than case study 1. The density of spots on an array continues to increase as the technology improves. It will become more difficult to visualize the spatial variability of an entire microarray as this density increases.

These graphs served two purposes. The molecular biologist involved in these case studies obtained a better understanding of the statistical models used to describe error and also found and corrected for sources of variability that were previously overlooked. The statistician involved in these case studies gained a better under standing of the sources of technical error as the biologist interpreted the graphs. Statisticians and molecular biologist working together can the improve power of microarray studies by identifying and reducing sources of variability.

References

Allison, D.B., Page, G.P., Beasley, T.M., and Edwards, J.W. (2006), DNA Microarrays and Related Genomics Techniques Design, Analysis, and Interpretation of Experiments, Chapman & Hall/CRC Boca Raton, FL

Bolstad, B.M., Irizarry RA, Astrand, M, and Speed, TP (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance *Bioinformatics*. 19 (2):185-193.

Chuaqui, R.F., et al. (2002) Post-analysis follow-up and validation of microarray experiments. *Nat Genet*, 2002. 32 Suppl: p. 509-14.

Clarke, J.D., and Zhu, T. (2006), Reference Microarray analysis of the transriptome as a stepping stone towards understanding biological systems Practical consideration and perspectives, The Plant Journal(45,630-650).

Hekstra, D., Taussig, A.R., Magnasco, M. and Naef, F. (2003) Absolute mRNA concentrations from sequence-specific alibration of oligonucleotide arrays. Nucleic Acids Res. 31, 1962-1968.

Keer, K.M., Martin, M., and Churchill, G.A.,(2000), Analysis of Variance for Gene Expression Microarray Data, Journal of Computational Biology 7(6): 819-837

Pontius, J. U., Wagner, L. and Schuler, G. D. (2003). UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): *National Center for Biotechnology Information.*

Quackenbush, J. 2002. Microarray data normalization and transformation. *Nature Genetics* 32, pp. 496-501.

Rosenzweig, B., *et al.* (2004) Dye-bias correction in dual-labeled cDNA microarray gene expression measurements. *Environ. Health Perspect.*, 112, 480–487.

SAS Institute Inc. 2004. SAS OnlineDoc® 9.1.3. Cary, NC: SAS Institute Inc.

Yang,Y.H., *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*. 2002 Feb 15;30(4):e15.

Appendix - SAS Code used to perform Analysis of Variance resulting in residuals used for figures.

Case Study I - Aflatoxin in Corn

Figures 4 and 5 – Residuals for simple within array variability – Dye was not included in the model and rep was coded so that the dye swap microarray for rep2 was treated as if it was an additional biological rep. This should have no effect on the simple within slide variability. Residuals from this model represent how each spot deviates from the mean of the three spots for a given gene x genotype x inoc x array combination.

```
PROC MIXED DATA=A NOBOUND;BY GENE;
CLASS REP GENOTYPE INOC DYE SPOT;
MODEL LOG2INT=GENOTYPE INOC GENOTYPE* INOC /*DYE*/ /OUTP=RESID_F4n5;
RANDOM REP(GENOTYPE) REP*INOC(GENOTYPE) ;
RUN;QUIT;
```

Figures 6 and 7– Residuals for adjusted within array variability – Each spot on the array contained measurement for both INOC treatments. This analysis is the same as the one for figures 4 and 5 with an additional random effect for SPOT within REP and GENOTYPE. This will remove any differences in common to both INOC treatments from the residuals.

```
PROC MIXED DATA=A NOBOUND;BY GENE;
CLASS REP GENOTYPE INOC DYE SPOT;
MODEL LOG2INT=GENOTYPE INOC GENOTYPE* INOC /*DYE*/ /OUTP=RESID_F6n7;
RANDOM REP(GENOTYPE) REP*INOC(GENOTYPE) SPOT(REP GENOTYPE) ;
RUN;QUIT;
```

Figures 8and 9– Residuals for between array variability –Data was first averaged over the three technical replications within an array. Residuals from this analysis measure REP x INOC within GENOTYPE or Var(within slide)/3 + Var(between arrays).

```
PROC MIXED DATA= NOBOUND;BY GENE;
CLASS REP GENOTYPE INOC DYE;
MODEL LOG2INT=GENOTYPE INOC GENOTYPE* INOC DYE OUTP=RESID_F8n9;
RANDOM REP(GENOTYPE);
RUN;QUIT;
```

Case Study II - Age of Cotton Fiber

Figures 11a and 11d – Residuals for simple within array variability – Dye was not included in the model. This should have no effect on the simple within slide variability. Residuals from this model represent how each oligo (or spot) deviates from the mean of the multiple oligos for a given gene x treatment x array combination.

```
PROC MIXED DATA=A NOBOUND;BY GENE;
CLASS ARRAY OLIGO DYE TRT;
MODEL LSIGNAL= TRT /OUTP=RESIDS;
RANDOM ARRAY TRT*ARRAY;
RUN;QUIT;
```

Figures 11b and 11e – Residuals for adjusted within array variability –Each spot on the array contained measurement for two treatments. This analysis is the same as the one for figures 11a and 11d with an additional random effect for OLIGO within ARRAY AND TRT. This will remove any differences in common to both treatments from the residuals.

```
PROC MIXED DATA=A NOBOUND ;BY GENE;
CLASS ARRAY OLIGO DYE TRT;
MODEL LSIGNAL= TRT /OUTP=RESIDS;
RANDOM ARRAY TRT*ARRAY OLIGO(ARRAY TRT);
RUN;QUIT;
```

Figures 11c and 11f – Residuals for between array variability –Data was first averaged over the multiple oligos within an array. Residuals from this analysis measure ARRAY X TRT error term or Var(within slide)/(# oligos per array) + Var(between arrays).

```
PROC MIXED DATA=B NOBOUND;BY GENE ;
CLASS ARRAY  DYE  TRT;
MODEL LSIGNAL= DYE TRT /OUTP=RESIDS;
RANDOM ARRAY ;
RUN;QUIT;
```