Quantification of sources of variation and accuracy of sequence discrimination in a replicated microarray experiment

J. Steven Brown[1], David Kuhn[2], Randall Wisser[3], Emilio Power[1], and Raymond Schnell[1]

BioTechniques 36:324-332 (February 2004)

cDNA microarray spot variability arises from many sources, and different systems have varying requirements for achieving the desired level of precision. We determined relative contributions to variance and investigated sequence discrimination using a multiple-array experimental design, with arrays subdivided to determine position and pin effect. Related fragments of 67 resistance gene homologs (RGHs) isolated from Theobroma cacao *L. and grouped by sequence similarity were spotted onto arrays, using two of the same RGHs in the fluorescent dye channels (Cy™3, Cy5) of the hybridization solution in a "dye-flip" design. A comprehensive statistical model accounted for variability well, giving a coefficient of variation (CV) based on experimental error of 2.12%. Although we were able to separate 85% of RGH group means clearly, some groups more similar to the target were indistinguishable due to nonspecific hybridization. Genetic factors together contributed 72.2% of the total variation, while position and pin effects and their interactions contributed 9.8%. Replication effect was statistically significant. Otherwise, no tests for position effects were significant. The results of the analysis indicate that our Genetic Microsystems 417™ arrayer and Affymetrix 428™ scanner are performing with sufficient precision, and we produced useful information for planning efficient future experiments.*

INTRODUCTION

Microarrays of cDNA have become an effective methodology for the analysis of gene expression (1) and are capable of producing data in large quantities for addition to expression databases. The Stanford Microarray Database (2) is an example of an expression database for storage and easy access of data for multiple experiments. In fact, microarray data from a given experiment are often more informative when used with sets of expression data from other experiments in a comparative fashion (3).

The integrity of data in these databases becomes paramount, as it arises from a set of highly technical and integrated processes. Each type of genetic data has many different sources of error constituting its error term. These sources of error differ in nature and over time for each type of data and for separate batches of data. It is critical that each process contributing data to the broader genetic perspective of any organism be shown to be capable of producing sufficiently accurate, precise, and repeatable data. In early testing of our Genetic Microsystems (GMS) 417™ arrayer and Affymetrix 428™ scanner, results showed relatively good repeatability and gave percent coefficient of variation (%CV) values in the range reported by Yue et al. (4).

We concur with the work by Kerr et al. (5) in using the analysis of variance (ANOVA) on original fluorescent intensity measurements, at least as a first-pass analysis as suggested by Wolfinger et al. (6), before analysis of differential expression (DE) ratios. Wolfinger et al. (6) found the use of ANOVA to be complementary to cluster methods previously proposed for interpreting microarray data and that its use prior to clustering or other data visualization methods would help to ensure statistically meaningful and repeatable interpretations. They also emphasized the importance of assessing statistical significance and showed clearly the disadvantages accepting 2- or 3-fold changes as cutoff values for assessing DE. Using a model-based error term removes unassociated variance and covariance from the error term used to test DE.

The Subtropical Horticulture Research Station (SHRS) plant genetics research group in Miami decided an initial validation experiment of the arrayer and scanner to be well worthwhile. Specific objectives of the validation were as follows: (*i*) to evaluate, quantify, and compare relative magnitudes of principal sources of error in our system; (*ii*) to assess our ability to distinguish signal consistently from background; (*iii*) to assess our ability to distinguish signal of specific hybridization from nonspecific hybridization; (*iv*) to evaluate the response of fluorescent signal relative to amount of target DNA spotted onto the slide (linearity);

[1]USDA-ARS, Subtropical Horticulture Research Station, Miami, [2]Florida International University, Miami, FL, and [3]Cornell University, Ithaca, NY, USA

Table 1. Slide Groups, Dyes, and RGH Fragments Used in the Probe Solution in Dye-Swap Experiment, and Frequency of Incorporation of Fluorescent Dye by Each Fragment

Slide Group	Slides Contained	Dye	Sequence	Frequency of Incorporation (No. of Labeled Nucleotides/1000)
1	4, 7	Cy3	RGH 4.1	12
		Cy5	RGH 8.8	12
2	8, 9	Cy3	RGH 8.8	15
		Cy5	RGH 4.1	18
	18	POPO-3	N/A	N/A

RGH, resistance gene homolog; N/A, not applicable.

and (v) to produce variance component estimates, which are useful for efficient future experimental design. The necessity of replication in microarray work has been clearly shown (7), thus obtaining an initial variance estimate for this factor is minimally necessary for efficient experimental design and power calculations. Positional effects on the array were also addressed, should it be necessary to include them in future experimental designs.

An important goal of the cacao (*Theobroma cacao* L.) project of the SHRS is to use existing molecular markers and to develop new markers for identifying resistance genes to witches' broom and frosty pod caused by *Crinipellis perniciosa* (Stahel) and *Moniliophthora roreri* (Cif. and Par.), respectively. In addition to the approach of interval mapping using segregating populations, the plant genetics laboratory at the SHRS has developed resistance gene homologs (RGH) and adapted them for use in marker-assisted selection (8). A great deal of sequence data and other bioinformatic data is obtained in this process, giving us useful a priori information about the RGH sequences and rendering these fragments well suited for this project. As the RGH fragments are developed using degenerate primers to find conserved sequence motifs as a first screen from target DNA, this experiment also provided an opportunity to investigate the effect of nonspecific hybridization among closely related sequences on fluorescent signal. Related families of genes continue to be discovered as genomic research progresses, and the presence of a related sequence in cDNA from expressed RNA could cause signal to emit from a related, but potentially functionally different sequence and thus be important for correctness of interpretation of microarray data.

MATERIALS AND METHODS

Sixty-seven related sequence fragments from candidate RGHs were spotted onto microarray slides (the term slide is used interchangeably and is equivalent to array from this point on) (GenBank® accession nos. AF403606–AF402768). The RGH were developed from *T. cacao* L. DNA isolates at the SHRS (8) and were placed into 11 groups based on nucleotide sequence similarity. Cloned DNA was amplified with M13 forward and reverse primers as described in Kuhn et al. (8). Products were ethanol-precipitated, resuspended in 50% dimethylsulfoxide (DMSO), and pipeted in four adjacent wells (e.g., A1, A2, B1, B2) so that each pin of the four-pin head was dipped into the same solution. The concentration of products for each of the 67 RGH clones was not adjusted. Products were spotted onto CMT-GAP slides (Corning, Corning, NY, USA) at five hits per spot with a GMS 417 arrayer (Affymetrix, Santa Clara, CA, USA). RGH 4.9 (AF403706) was spotted in a 2-fold dilution series over six dilutions to provide information about linearity of response, accuracy of dose and response, and lack of fit (bias). Target solutions were prepared using RGH 4.1 and RGH 8.8 with primers internal to the cloned fragment. For RGH 4.1, a 374-nucleotide portion of the original 519-base homolog (72%) was amplified from nucleotide 114 to 487. For RGH 8.8, a 144-nucleotide portion of the original 524-base homolog (27%) was amplified from nucleotide 356 to 499. Both amplifications included the presence of Cy™3-dCTP or Cy5-dCTP (Amersham plc, Little Chalfont, Buckinghamshire, UK). Two slides were hybridized with a mixture of RGH 4.1 Cy3 and RGH 8.8 Cy5, and two slides were hybridized with the same fragments having been labeled with the opposite dyes in what is commonly called a dye-swap or latin square design (9). A complete listing of slide groups, slides contained in each group, dyes, and sequences is given in Table 1, along with the frequency of incorporation of fluorescent dyes by each sequence. An additional slide was stained with POPO™-3, a dimeric cyanine nucleic acid stain (Molecular Probes, Eugene, OR, USA) to measure relative quantities of DNA delivered to each spot on the slide. Fragments within RGH groups are often very close, being greater than 80% identical in sequence, and even closer in some groups, especially group 4. The fragment amplified from RGH 4.1 was identical over its length with the corresponding sequences from RGH 4.2, 4.3, 4.6, 4.7, and 4.9 and differed by one nucleotide (position 224) from RGH 4.4, 4.5, 4.8, and 4.10. The fragment amplified from RGH 8.8 was identical with RGH 8.7 over its length and differed by three nucleotide substitutions (positions 387, 408, and 440) from RGH 8.1, 8.2, 8.3, and 8.4. The RGH 8.8 fragment differed by an additional two nucleotide substitutions from RGH 8.5 and 8.6 (positions 441 and 450). RGH 8.8 differed by 17 nucleotide substitutions from RGH 8.10 and 20 nucleotide substitutions from RGH 8.11. Groups of RGH fragments, members of each group, and relative relationships of groups to each other based on sequence similarity have been previously reported in Kuhn et al. (8). Each slide was divided into four quadrants, and three replications within each quadrant were spotted for each sequence by each pin of the 4-pin arrayer head. The same randomization pattern was used for each replication. An illustration of the plate map is available upon request.

Our work corresponds to analytic steps prior to forming DE ratios, as suggested by Wolfinger et al. (6). The two response variables analyzed were the natural log of the fluorescent signal

after background subtraction (by the manufacturer's software) in each channel, and log of background as measured by the Affymetrix 428 scanner analysis software, Jaguar 2.0® (Affymetrix). Two split-plot ANOVAs were used with nesting structure for partitioning and comparing sources of variance, followed by a third analysis using the first model and appropriate regressors to analyze the dilution series.

The first analysis (Model 1) consisted of a mixed model with two internal nesting structures and grouped the two slides (nos. 4 and 7) together, which contained RGH 4.1 and RGH 8.8 in Cy3 and Cy5, respectively, followed by an identical analysis grouping the other two slides (nos. 8 and 9), which had the RGH segments swapped to the opposite dye in the hybridization target. It was our objective to first study both RGH selected to be in the hybridizing solution individually in each dye, thereby avoiding the confounding effects that would result from combining slide groups, as discussed in Kerr and Churchill (9), stemming from the latin square design. Main plots corresponded to replications nested within quadrant by pin sector combinations, and subplots corresponded to individual microarray spots (RGH segments) nested within their respective groups. Three nesting structures existed within the model. Quadrants were nested within slides, replications were nested within quadrant by pin sector combinations, and RGHs were nested within their respective groups. The resulting model became quite complex, however the physical structure of microarray experimentation together with the meaningful group structure of the RGH based on nucleotide similarity necessitated this triple nesting. The statistical model consisted of all main effects and all interactions deemed to be of interest a priori as well as those with non-negligible effects. Some interactions were impossible to obtain due to the complex nesting structure of the model. Fixed effects included the following factors: pins, groups, RGH (group),

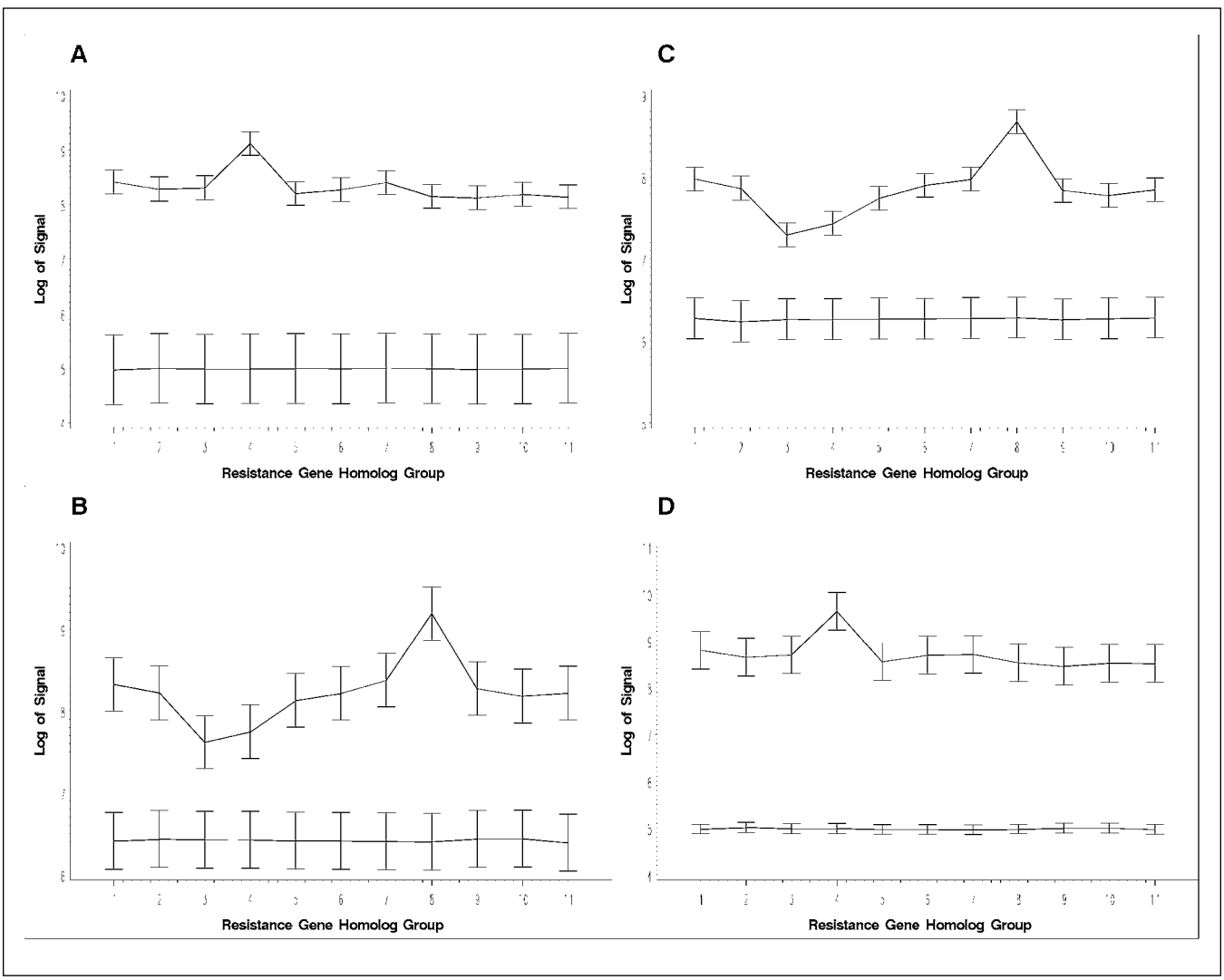

Figure 1. Means and confidence limits of resistance gene homolog (RGH) group means. (A) Slide group 1, dye = Cy3, target sequence = RGH 4.1. (B) Slide group 1, dye = Cy5, target sequence = RGH 8.8. (C) Slide group 2, dye = Cy3, target sequence = RGH 8.8. (D) Slide group 2, dye = Cy5, target sequence = RGH 4.1. Log of signal, upper line; log of background, lower line; 95% mean confidence interval is shown by crossbars.

pin × group, and pin × RGH (group). The factors, slide, quadrant (slide), pin × slide, and replication [pin × quadrant (slide)] were considered as random effects. Letting y_{ijklm} denote the signal measurement from the i^{th} slide, j^{th} quadrant, k^{th} pin, l^{th} replication, and n^{th} *RGH* within the m^{th} group, the complete statistical model for Model 1 can be written as follows:

$$log\,(y_{ijklmn}) = \mu + S_i + Q\,(S)_{ij} + P_k + S \times P_{ik} + R\,[P \times Q\,(S)]_{ijkl} + G_m + RGH\,(G)_{mn} + P \times G_{km} + P \times RGH\,(G)_{kmn} + \varepsilon_{ijklmn}$$

where μ is the experimental grand mean, S_i is the effect of the i^{th} slide, $Q\,(S)_{ij}$ is the effect of the j^{th} quadrant nested within the i^{th} slide, P_k is the effect of the k^{th} pin, $S \times P_{ik}$ is the interaction of the i^{th} slide with the k^{th} pin, $R\,[P \times Q\,(S)]_{ijk}$ is the effect of the l^{th} replication nested within the k^{th} pin in the ij^{th} nesting of quadrant (slide), G_m is the effect of the m^{th} group, $RGH\,(G)_{nm}$ is the effect of the n^{th} *RGH* in the m^{th} group, $P \times G_{km}$ is the effect of the interaction of the k^{th} pin with the m^{th} group, $P \times RGH\,(G)_{kmn}$ is the effect of the interaction of the k^{th} pin with n^{th} *RGH* in the m^{th} group, and ε_{ijklmn} is the individual error term for each spot on the slide.

Initial exploratory analysis indicated that a log transformation sufficed well for conformity of both response variables, signal, and background, to a normal distribution as many other researchers have found (1,9,10) and to the usual stochastic assumptions (i.e., random effect normally distributed with zero means and variance components as given in the model) and standard assumption for the fixed effects (11).

Proc Mixed of SAS® (SAS Institute, Cary, NC, USA) (12) was used with restricted maximum likelihood (REML) fitting of variance components from random effects. The Bonferroni adjustment was then used as a second step to correct the error term for multiple mean comparisons of RGH group mean comparisons. The Bonferroni adjustment was chosen, as firstly, not all group mean sample sizes were equal and secondly, group means constituted specifically the particular set of estimated contrasts of interest to be compared (13). Least-squares means estimates from Proc Mixed construct the standard error of each mean and corresponding confidence intervals based on mixed model theory (11) and appropriate estimable functions. A detailed explanation of least-squares means as calculated in Proc Mixed of SAS can be found in the documentation of the SAS Institute (12). Error terms constructed in this manner reflect the variance of the mean being tested based on variances (and covariances, if present) coming from the model structure, as opposed to a simple standard error of the mean.

The second analysis (Model 2) combined both slide groups and target solutions, with all effects considered random, and the objective being to compare relative magnitudes of sources of variance from a more comprehensive perspective. This is a common and convenient practice in statistical analysis for the above stated goal (14). In this combined analysis, we can include more main effects and interactions, enabling their comparison as contributors to total variance, though due to confounding and nesting structure, it was still impossible to resolve some interactions of interest. The second model included the following additional main effects and interactions: dye (Cy3 versus Cy5), sequence (RGH 4.1 versus RGH 8.8, in the target solution), dye × group, dye × pin, pin × sequence, and sequence × RGH (group). This constituted the most saturated model possible, as found by sequentially testing models with interaction terms.

The dilution series of RGH 4.9 was analyzed separately using a scan of one slide in which RGH 4.9 was in the Cy3 channel. At a gain setting of 60 decibels, log scale, the signal of the lower dilutions in this scan was visible, while the upper concentration did not appear saturated until the most concentrated member of the series. The regression analysis started with a linear model and added exponential terms until a cubic regression was reached, after which no higher order regressors were significant. Observations were considered to be outliers in this analysis, as the absolute value of the student *t*-value for the residual of the observation was $\geq |3|$.

RESULTS

Upon scanning the hybridized arrays, it was noticed that quadrant 1 was consistently visibly abnormal from other quadrants, and preliminary statistical analyses confirmed this. Quadrant 1 was therefore dropped from all analyses. This resulted from a failure of the coverslip to fit over the entire addressable area of the slide within quadrant 1. RGH group 2, the only group with a single member, also had a high number of spots that were either outliers or very close to being such, but was retained in the analysis nonetheless. We saw no clear justification for its elimination.

RGH segments that consistently had signals outside the overall range of segments in their respective group were observed to also have correspondingly lower or higher signal on the POPO-3 stained slide. These relative outliers were concluded therefore to be due to the amount of DNA bound to the slide. The spot-to-spot correlation coefficient between individual log signal of spots from the POPO-3 stained slide with the corresponding spot on other slides (calculated by slide) ranged between -0.30–0.31. Although certain outliers could be accounted for by DNA amount, as reflected by POPO-3 stain, the association between the POPO-3 signal and hybridization signal was essentially zero over all spots. When

Table 2. Number and Percentage of Least Square Means Comparisons with *P* > 0.05, after Bonferroni Group Mean Correction

Slide Group(s)	Dye	RGH	Total Comparisons	Total No. of Means with *P* > 0.05	% > 0.05
1	Cy3	4.1	55	7	12.70
1	Cy5	8.8	55	10	18.10
2	Cy3	8.8	55	6	10.90
2	Cy5	4.1	55	10	18.10

Comparisons over Slide Group(s)	Total Comparisons	Total No. of Means with *P* > 0.05	% > 0.05
4.1	110	17	15.45
8.8	110	16	14.55
4.1 and 8.8	220	33	15.00

RGH, resistance gene homolog.

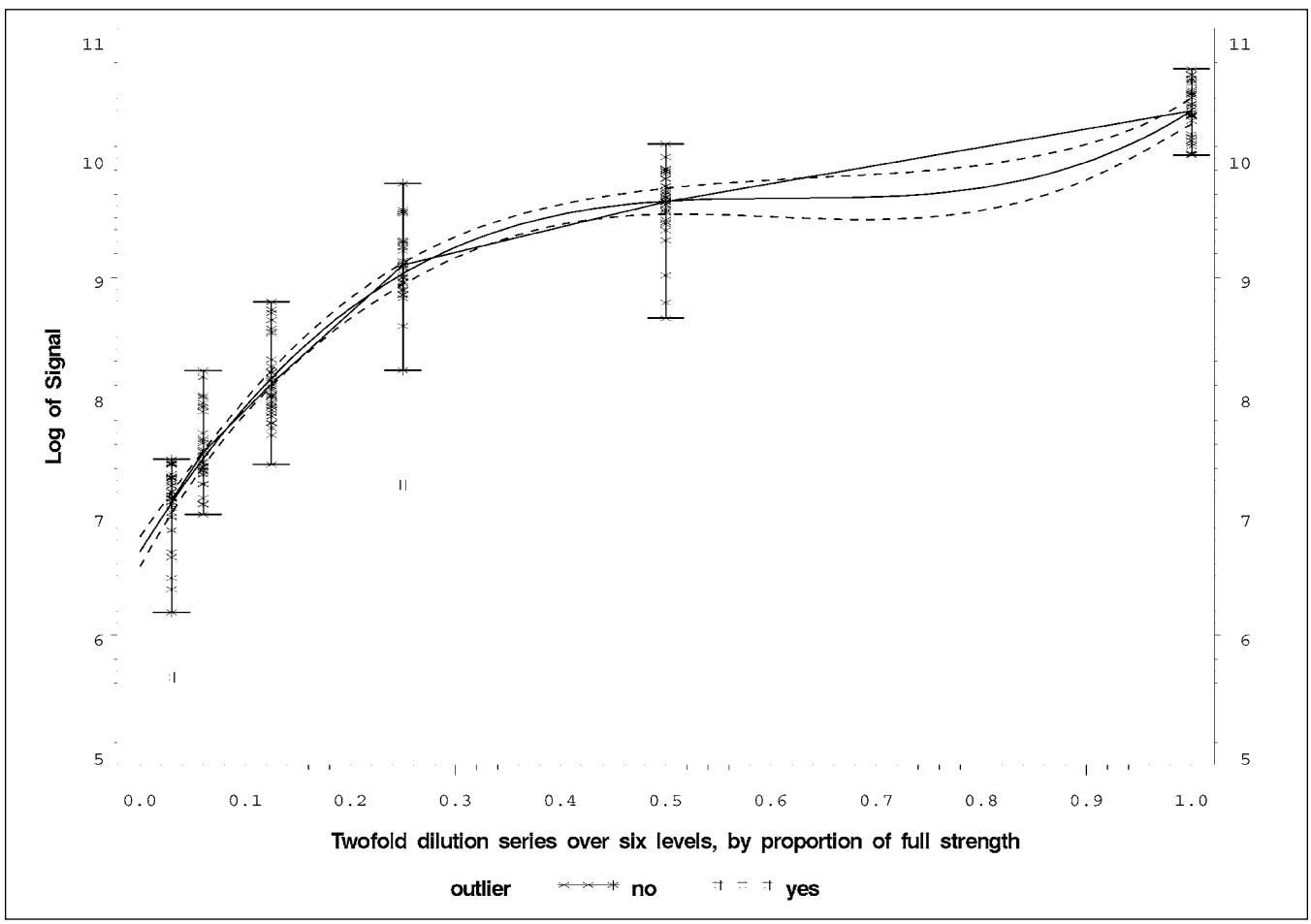

Figure 2. Plot of log of signal versus dilution. Upper and lower confidence intervals (95%) of predicted means. Regression equation: log (signal) = 6.700755 + 14.34237 × dilution - 23.22118 × dilution^2 + 12.57545 × dilution^3.

Table 3. Tests of Fixed Effects and Estimates of Variance Components: Mixed Model by Florescent Channel and Target Mixture

Slide Group 1, Dye = Cy3, Sequence = RGH 4.1				
Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F-Value**	**Pr > F**
Pin	3	3	3.74	0.1537
Group	10	4488	2628.35	<0.0001
RGH (Group)	56	4488	444.72	<0.0001
Pin × Group	30	4488	13.20	<0.0001
Pin × RGH (Group)	168	4488	7.20	<0.0001
Variance Component Estimates				
Source	**Estimates**	**Standard Errror**	**Lower CL[b]**	**Upper CL**
Slide	0.0180	0.0352	0.0027	9722.8300
Quadrant (Slide)	0.0147	0.0107	0.0052	0.1321
Slide × Pin	0.0070	0.0062	0.0021	0.1356
Rep[a] [Pin × Quadrant (Slide)]	0.0045	0.0009	0.0032	0.0069
Residual	0.0191	0.0004	0.0183	0.0199

Slide Group 2, Dye = Cy5, Sequence = RGH 4.1			
Tests of Fixed Effects			
Effect	**Num DF**	**Den DF**	**F-Value**
Pin	3	3	1.41
Group	10	4488	3646.59
RGH (Group)	56	4488	517.27
Pin × Group	30	4488	3.16
Pin × RGH (Group)	168	4488	2.70
Variance Component Estimates			
Source	**Estimates**	**Standard Errror**	**Lower CL[b]**
Slide	0.0727	0.1191	0.0127
Quadrant (Slide)	0.0238	0.0174	0.0083
Slide × Pin	0.0130	0.0011	0.0039
Rep[a] [Pin × Quadrant (Slide)]	0.0094	0.0018	0.0067
Residual	0.0183	0.0004	0.0176

Slide Group 1, Dye = Cy5, Sequence = RGH 8.8				
Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F-Value**	**Pr > F**
Pin	3	3	27.66	0.0109
Group	10	4488	3342.46	<0.0001
RGH (Group)	56	4488	291.87	<0.0001
Pin × Group	30	4488	5.7	<0.0001
Pin × RGH (Group)	168	4488	3.66	<0.0001
Variance Component Estimates				
Source	**Estimates**	**Standard Errror**	**Lower CL[b]**	**Upper CL**
Slide	0.0468	0.0783	0.0080	700.2600
Quadrant (Slide)	0.0238	0.0170	0.0085	0.2030
Slide × Pin	0.0016	0.0015	0.0004	0.0578
Rep[a] [Pin × Quadrant (Slide)]	0.0024	0.0005	0.0016	0.0040
Residual	0.0369	0.0008	0.0354	0.0384

Slide Group 2, Dye = Cy3, Sequence = RGH 8.8			
Tests of Fixed Effects			
Effect	**Num DF**	**Den DF**	**F-Value**
Pin	3	3	4.54
Group	10	4488	3910.4
RGH (Group)	56	4488	481.53
Pin × Group	30	4488	4.44
Pin × RGH (Group)	168	4488	2.92
Variance Component Estimates			
Source	**Estimates**	**Standard Errror**	**Lower CL[b]**
Slide	0.0058	0.0154	0.0007
Quadrant (Slide)	0.0102	0.0077	0.0035
Slide × Pin	0.0052	0.0049	0.0015
Rep[a] [Pin × Quadrant (Slide)]	0.0072	0.0014	0.0051
Residual	0.0243	0.0005	0.0233

[a]Rep, replication.
[b]95% Confidence limits.
DF, degrees of freedom; F, F test value; PR, probability; CL, confidence limits.; Num, numerator; Den, denominator; RGH, resistance gene homolog.

Table 4. Sources of Variation, Associated Variance Component Estimates, Upper and Lower Confidence Limits, and Percent Contribution to Total Variance of Each Source

Source	Estimate	Std. Err.	Lower CL	Upper CL	% Total
Dye	0.07191	0.1019	0.01429	75.2387	11.92
Pin	0.01334	0.01241	0.003845	0.3453	2.21
Sequence	0.1494	0.217	0.02902	219.23	24.76
Group	0.02835	0.01993	0.01022	0.2298	4.70
RGH (Group)	0.001765	0.03287	—	—	0.29
Dye × Group	0.000379	0.000184	0.000176	0.001318	0.06
Dye × Pin	0.000428	0.00036	0.000134	0.006754	0.07
Pin × Sequence	0.000744	0.000618	0.000235	0.01111	0.12
Pin × Group	0.000393	0.000234	0.00016	0.00204	0.07
Sequence × RGH (Group)	0.2577	0.0449	0.1882	0.3745	42.72
Pin × RGH (Group)	0.002236	0.000288	0.001764	0.002929	0.37
Slide	0.01874	0.02125	0.004549	1.8452	3.11
Quadrant (Slide)	0.01754	0.009007	0.007872	0.06745	2.91
Rep (Slide × Quadrant × Pin)	0.005404	0.000727	0.00422	0.00717	0.90
Slide × Pin	0.004078	0.002219	0.001762	0.01749	0.68
Residual	0.03087	0.000318	0.03026	0.0315	5.12
Total	**0.603277**				**100.00**

Std. Err., standard error; CL, confidence limit; RGH, resistance gene homolog.

we attempted to use POPO-3 signal as a covariate to adjust fluorescent signal within the context of Model 1 (data not shown), the covariate was insignificant.

We were consistently able to separate signal from background statistically significantly as calculated by the Affymetrix scanner software, Jaguar 2.0. This was possible when the comparison was based on group mean data and also nearly always with individual spot data (not presented due to volume), except for a very few outliers. For these few instances, the background calculation algorithm resulted in a negative signal, an odd situation known to happen with many background calculation algorithms. Initially, we were able to consistently separate signal of segments of related groups from signal of less related groups, using RGH group mean data and its appropriate error term to construct the confidence limits. Figure 1, A–D, shows the means and confidence limits of group means. The absence of overlap of the confidence limits of group means indicates statistically significant mean separation. However, upon applying the Bonferroni adjustment for multiple group mean comparisons, the mean signal of the group corresponding to the fragment in the target is indistinguishable from some other groups 15% of the time on average, as summarized in Table 2. RGH group mean comparison tests on slides treated with target solution containing RGH 4.1 failed to distinguish groups when, the groups were in fact different (Type II error), at a rate of 15.4%. The Type II error rate among RGH group means on slides with RGH 8.8 in the target was 14.5%, giving an overall Type II error rate of 15%. Our ability to consistently distinguish among cDNA fragments with less than 80% similarity (minimum similarity within RGH group fragments) by hybridization is not yet completely reliable, based on these results. However, one must consider that the RGH were selected on common sequence motifs, hence, results from RGH fragments would not correspond exactly to cDNA from more generally expressed RNA. Detailed sequence comparisons given in Kuhn et al. (8) of certain specific RGH show that one would expect more nonspecific hybridization among the more similar RGH groups. Further experimentation to find more optimal hybridization conditions (ionic strength, temperature, etc.) can also reduce nonspecific hybridization, however knowing initially that signal was affected in such a manner was important.

RGH group means of log background (Figure 1, A–D) are quite constant, as are the confidence intervals for each analysis, though the size of background and its error differed among different slides and fluorescent channels. Background signal calculation appears to be very constant within slide groups and independent of signal size in this experiment. Furthermore, correlation analyses between log signal and log background done on individual spot data were insignificant (not shown), calculated by slide or over the entire data set.

Statistical tests for fixed effects in the mixed model analysis of each RGH/dye combination resulted in pin-to-pin differences not statistically significantly different from zero in three of the four analyses. The exception was the solution containing Cy5 and RGH 8.8, in which pin-to-pin differences were highly significant. We can pose no assignable cause for this one exception, though pin-to-pin differences were observed in earlier runs of this system. All other fixed effects [group, RGH (group), pin by group, pin by RGH (group)] were highly significant (Table 3), in particular the first two effects, indicating that a large portion of the log signal variance is coming from genetic differences (72.2%), as is desired. Variance components due to the first three random effects in the model [slide, quadrant (slide), and slide by pin] were very small and not significantly different from zero in any of the four analyses. Taken together with the lack of significance of pin effects, this indicates no statistically measurable position effect on the slide. This simplifies future experimental design for DE research with this system, indicating no need for planned systematic placement or randomization of expressd sequence tags (ESTs) over the array. The random effect due to replication nested with pin [quadrant (slide)] was, however, consistently highly significant (from the *z*-test in Proc Mixed or from full model-reduced model testing, not shown), indicating the importance of replication within an array, as other researchers have emphasized (6,7).

Table 4 contains the results of Model 2, the model in which dye-swap treatments were combined, treating each factor as a random effect, and

fitting a maximally saturated model. Looking at the percent of total variation contributed by each factor, the genetic components, sequence, group, and the interaction, sequence × RGH (group), contribute the highest percentages of variance (24.76%, 4.70%, and 42.72%, respectively), followed by dye, slide, and quadrant (slide) (2.21%, 3.11%, and 2.91%). Though visible pin-to-pin differences were noted in preliminary analyses, pin effect accounted for only 2.21% of the total variance. Components due to slides and quadrant effects within slides were also small, though larger than the component for replication effect, however tests in both models were significant only for replications. No two-way interactions contributed unusually large amounts of variance other than sequence × RGH (group). The interaction of dye × sequence was confounded with the slide effect in our analysis (see Table 1) and was nonestimable. Though this could inflate the slide component, since there were only two fragments in the target solution, any possible effect would be relatively small compared to a set ESTs derived from mRNA from a living cell. It is known that the rate of incorporation of dyes differs over fragments, an important point that Kerr and Churchill (10) discussed, and its variance would come out in the dye × sequence interaction. The percent of total variance explained by the residual term, 5.12%, is reasonable, giving a CV of 2.1% from a parallel least squares analysis (11), compared to spot-to-spot variances that can often have %CV values greater than 50%.

Figure 2 contains a plot of the dilution series for the scan previously mentioned with the third degree regression equation. The cubic fit gave an R^2 of 0.969, which is quite acceptable. A term for lack-of-fit (bias) was included and was insignificant, indicating very little bias. This dilution series spanned a range from the linear response region at low cDNA concentrations to the nonlinear region of saturation at the top. Yue et al. (4) reported linearity within any three orders of magnitude, their dilution series having spanned three orders of magnitude, while the dilution series in this experiment spanned a 12-fold range. Our results are difficult to compare directly with those of Yue et al. (4), though ours appear to span a larger range. In the lower portion of the response (approximately 0.00- to 0.25-fold), we observe a mostly linear response, though a quadratic term could be fit in this span, also. Still, we find the response of signal to dilution to be generally acceptable, and it has no statistically significant bias. We consider it advisable to test new systems for dose-response and to observe the type of response curve obtained, as well as accuracy and bias for a reasonable range of inputs, and to know in what range the system operates well.

The issues of heterogeneity of variance among genes and spot-to-spot variability have been well discussed by several researchers (6,10) along with ways of dealing with it in data analysis. Residual analysis of our models, however, did not indicate heterogeneity of variance among residuals associated with RGH fragments on the slide or the size of log signal, therefore there it appears unnecessary to apply first a normalization model, followed by a gene model (6). The fact that fragments in this experiment are more similar to one another than what one would likely obtain in set of ESTs derived from organismal RNA can explain this result. Future use of the microarrayer at SHRS will, in all likelihood, necessitate a two-step modeling process, removing all effects prior to gene and gene by slide interaction effects in the model, followed by a second ANOVA by a gene and gene by slide model.

DISCUSSION

This test and series of analyses indicated that our microarray system is functioning sufficiently well to easily distinguish foreground signal from background and usually to distinguish specific hybridization from nonspecific hybridization to less related segments (more than 20% different), with sufficient replication and proper modeling of the variation. RGH sequences could potentially be used in future experiments as reference genes, to guide the separation of specific and nonspecific hybridization. Relating the degree of similarity of segments to strength of

hybridization signal in a quantitative manner was outside the scope of this experiment, but could constitute a potentially interesting issue for future investigation, especially in how nonspecific hybridization may inflate the signal in microarray analysis and hence affect Type I error. The fact that signals were at times unable to be differentiated from one another when the sequences were greater than 20% different has important ramifications about Type I error rates when there is no control for nonspecific hybridization to closely related sequences in DE experiments, and it is imperative to know when such a problem exists. Conclusions about the identity or difference of any segment without a large signal difference should be taken with caution at this point in this system. Given the amount of replication of each spot (48 times per slide per channel), the decrease in ability by 15% to statistically discern the RGH group matching the fragment in the target when the Bonferroni adjustment was applied point out the need for improved hybridization conditions. The magnitude of background signal on some slides/channel combinations and its variance could also affect sensitivity. This would obviously be important when looking for genes that express at low levels, such as some genes that are hypothesized to induce the expression of other genes.

The fact that no two-way or higher-order interactions other than sequence × RGH (group) were of large importance is encouraging, indicating the possibility to produce quality microarray data using relatively simple models for analysis. The fact that the two fragments in the target solution come from different RGH groups and comprise such different amounts (74% versus 27%) of the entire RGH sequence can explain the interaction of sequence × RGH (group), together with the fact that they would be expected to differentially hybridize across RGH fragments. The dye by sequence interaction, though nonestimable, still appears to be low in this trial, as mentioned above, probably lower than normal. Though blocking the array by quadrants and pins did not give statistical significance, the fact that quadrant 1 was able to be identified as an outlier, and hence eliminated, increased the precision of the experiment. Otherwise, we found no need to consider array positional effects based on these results. Significance of the replication effect within arrays demonstrates again the importance of replication. Given the relatively small size of the variance component due to replications (<1% of total variation), we consider three replications per array to be sufficient, in agreement with other researchers (6,7). The use of mixed model analysis together with the least square means is shown to be another method by which sensitivity and precision can be increased, and in situations with a nongenetic factor exerting a large effect, the increase could be large. Calculation of least squares means in well developed software, such as Proc Mixed of SAS or the mixed model capability of Genstat® (NAG LTD, Oxford, UK), with the option for multiple comparison adjustments and several different methodological choices, allows the researcher to produce statistically correct family-wise error rates very easily when testing several ESTs or groups of ESTs simultaneously for a variety of situations. The dilution series showed that our system is functioning accurately from the lower ranges of detection to saturation and generally gives a linear response within any 3-fold range. In summary, the results of this exploratory validation analysis indicate that our system is functioning well and that we have sufficiently flexible software for data analysis. We should be able to obtain data with this system in the future with three replications per slide and with sufficient accuracy and precision to identify differentially expressed genes.

ACKNOWLEDGMENTS

We would like to acknowledge Huiben Yue and Cuauhtemoc Cervantes for statistical consulting and Cecile Olano for helpful editorial comments.

REFERENCES

1. **Coombes, K.R., E.W. Highsmith, T.A. Krogmann, K.A. Baggerly, D.N. Stivers, and L.V. Abruzzo.** 2002. Identifying and quantifying sources of variation in microarray data using high-density cDNA membrane arrays. J. Comp. Biol. *9*:655-669.
2. **Sherlock, G., T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. Matese, S. Dwight, M. Kaloper, S. Weng, et al.** 2001. The Stanford microarray database. Nucleic Acids Res. *29*:152-155.
3. **Quackenbush, J.** 2001. Computational analysis of microarray data. Nat. Rev. Genet. *2*: 418-427.
4. **Yue, H., P.S. Eastman, B.B. Wang, J. Minor, M.H. Doctolero, R.L. Nuttall, R. Stack, J.W. Becker, J.R. Montgomery, et al.** 2001. An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. Nucleic Acids Res. *29*: e41.
5. **Kerr, K.M., M. Mitchell, and G.A. Churchill.** 2000. Analysis of variance for gene expression microarray data. J. Comp. Biol. *7*:819-837.
6. **Wolfinger, R.D., G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R.S. Paules.** 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J. Comp. Biol. *8*:625-637.
7. **Lee, M.L., F.C. Kuo, G.A. Whitmore, and J. Sklar.** 2000. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc. Natl. Acad. Sci. USA *97*:9834-9839.
8. **Kuhn, D.N., M. Heath, R.J. Wisser, A. Meerow, J.S. Brown, U. Lopez, and R.J. Schnell.** 2003. Resistance gene homologues in *Theobroma cacao* as useful genetic markers. Theor. Appl. Genet. *107*:191-202.
9. **Kerr, K.M. and G.A. Churchill.** 2001b. Statistical design and the analysis of gene expression microarray data. Genet. Res. *77*: 123-128.
10. **Kerr, K.M. and G.A. Churchill.** 2001a. Experimental design for gene expression microarrays. Biostatistics *2*:183-201.
11. **Searle, S.R.** 1971. Linear Models. John Wiley, New York.
12. **SAS® Institute.** 2000. SAS/STAT Software Version 8. SAS Institute, Cary, NC.
13. **Neter, J. and W. Wasserman.** 1976. Applied Linear Statistical Models. Richard D. Irwin, Homewood, IL.
14. **Brown, J.S. and B. Glaz.**. 2001. Analysis of resource allocation in final stage sugarcane clonal selection. Crop Sci. *41*:57-62.

Received 2 September; accepted 3 November 2003.

Address correspondence to:

J. Steven Brown
United States Department of Agriculture-Agricultural Research Service (USDA-ARS)
Subtropical Horticulture Research Station (SHRS)
13601 Old Cutler Road
Miami, FL 33158, USA
e-mail: miajb@ars-grin.gov