Iterative linear regression by sector: renormalization of cDNA microarray data and cluster analysis weighted by cross homology.

David B. Finkelstein, Jeremy Gollub, Rob Ewing, Fredrik Sterky, Shauna Somerville, J. Michael Cherry

Abstract

Empirical evidence and observations validated by statistical tests have indicated that several distinct types of consistent measurement error can alter the interpretation of cDNA microarray data. Whenever possible models of error are derived during quality assessment and applied during data analysis. When measurement error is detectable and conforms to a defined model, corrections can be applied during renormalization.

However, some measurement errors are detectable but less well defined. In such cases, parallel analyses are required to determine the significance of such effects. Furthermore, supporting biological evidence from a distinct method designed to detect the problem may be required. In the specific case of the Spellman data both well-defined problems and ambiguous problems were examined.

First, the clearly detectable and definable measurement errors are corrected through renormalization. Reanalysis of the Spellman and Sherlock cell cycle data set begins with a new method of normalization that more accurately reduces the effects of outliers and spatial variation on the arrays. First, all intensity values are log transformed, then linear regression is performed separately on each sector. These sectors were produced by slotted printing pins. The Spellman data has four sectors and was printed with four distinct pins. Then these residuals are calculated for these four regression lines; one for each sector. Outliers (those residuals where |e| > 2 x std dev of e) are removed and the four regression functions are recalculated. If the difference between the value of r-squared of the new regression line is less than .001 of the old, then no further residuals are removed. Else, outliers are removed by the same test as above and the iterations continue. Once completely determined, the slope and intercept values are applied as correction factors to the log transformed channel 2 values. The result is that the function of log channel 1 and log channel 2 closely approximates y = x. Then these values are exponentiated, a new ratio is calculated and this ratio is put on the familiar log base2 scale. This renormalization alone has been demonstrated to substantially reduce the standard deviation of log2 ratios.

Next, the ambiguous task of detecting the effect of cross-hybridization was examined. The yeast genome is fully sequenced, thus the sequences of PCR fragments were known. Therefore it is possible, with some error, to determine the likely number of transcripts that could cross-hybridize to a given PCR fragment. The correlation between the likelihood of cross-hybridization and the frequency of transcripts with cross-homology is difficult to assess without empirical evidence. It is important to note that modeling the molecular events during hybridization has proven difficult. Therefore, no analysis can be used to correct data. However, a technique can be applied as an informed post hoc method. In this way, such analysis may indicate where biological confirmation experiments are warranted, rather than supply a mathematical solution.

Applying Linear Normalization

In all tested cases, applying a linear model of error combined with the iterative removal of outlying residuals reduces the standard deviation of the final

$\log_2$ratios. The range of the data is not substantially altered. However, the kurtosis increases and the skew may change in scale and in direction. Filtering iteratively normalized data without considering spatial bias, increased the number of genes that are consistently changed at the $|\log_2\_ratio| > 2$ for 1 of 11 Elutriation arrays by 4.3% (an increase of 9 genes) when compared to data normalized by the SMD default method. When the iterative method is applied each sector to correct spatial problems the number of genes that pass filtering criterion actually decreases. In both cases the overall standard deviation of the data is reduced. Only independent empirical methods can determine whether the differences in analysis methods are removing false positives.

Spatial Methods

Observation based on a spatial display tool developed for microarrays indicated that spatial problems may exist for several Spellman and Sherlock arrays. Renormalization by sector requires 4 parallel normalizations and assumes that functional groups of genes are not printed together. For many arrays the net result of spatial linear normalization is marginal. However, significant spatial effects have been detected in other cDNA arrays and therefore it is worth testing arrays for the effect.

Spatial bias is detectable with a simple ANOVA (y = $\log_2$ratio and X = grid #) that yields an F-test and r-squared value. Non-parametric methods such as the Kruskal-Wallis test also serve this function. Our current best estimate is that, if r-squared values are below .05, then spatial error is not significant. Best practice may indicate repeating experiments that are substantially altered, rather than applying sector specific normalization methods, which are *post hoc* and may only partially repair the effects.

Applying the Linear Method by Sector

For each the four independent sectors of each DNA microarray the iterative simple linear regression technique is applied. As expected many arrays, are not substantially altered by this approach. However in instances, where outliers are detectable by the F-test differences in normalization are noticeable (Figure 1). Note that the four sectors each have independent patterns with respect to background corrected channel 2 intensity (CH2D). The differences between the SMD method and Iterative method are consistently greater at low intensities: below 150. Each pattern is at a minimum where the linear regression equation for a given sector is equal to the SMD global mean. In this case, there is a clear difference in the minimum of one pattern, which may indicate spatial bias in that sector.

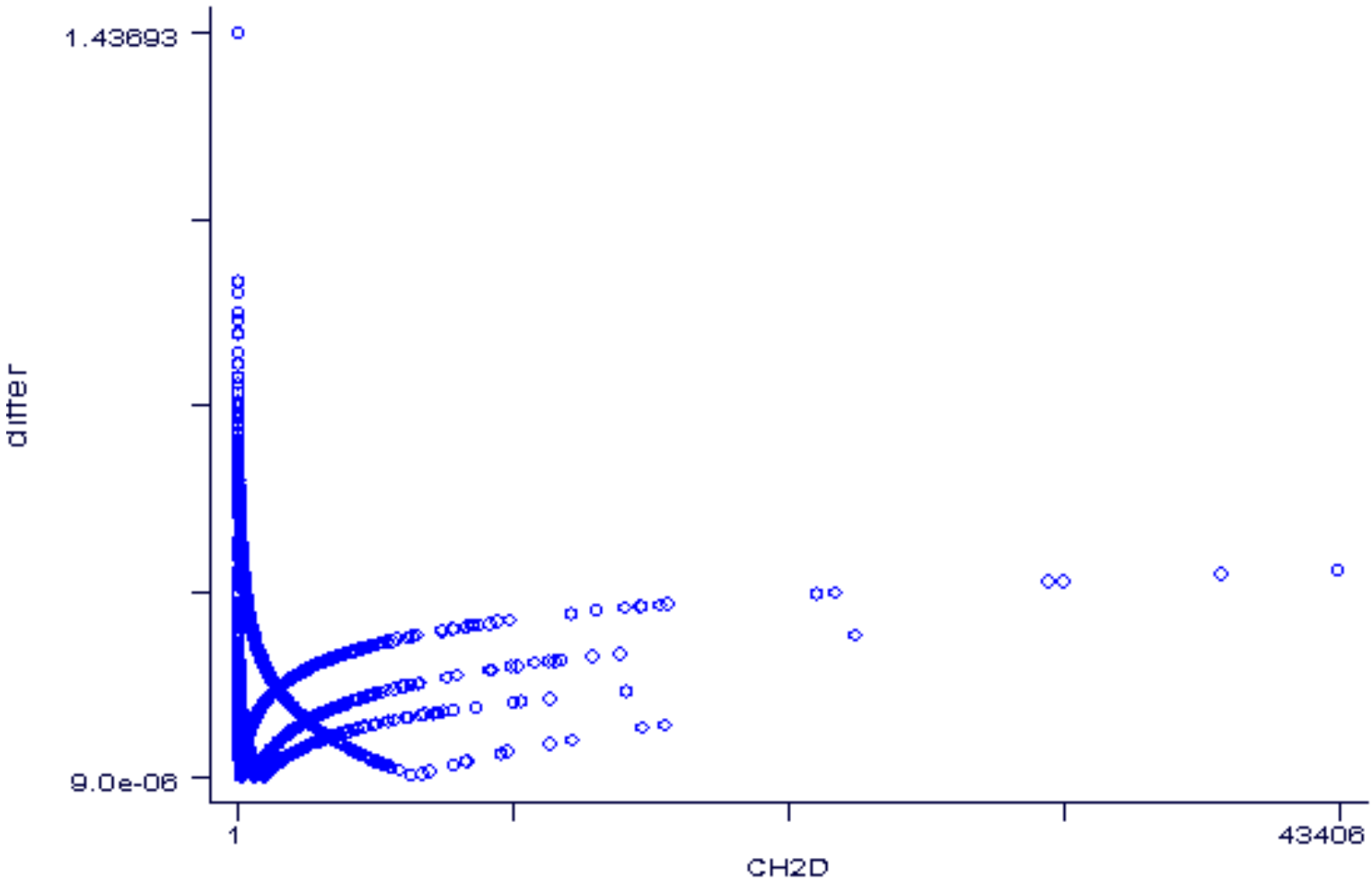

Figure 1. *The absolute value of the difference between log_2 ratio calculated by the SMD method and the Iterative method is plotted on the y-axis. The background-corrected channel 2 intensity is plotted on the x-axis*

Filtering results

Filtering parameters: all spots that have an average intensity of 100 in each channel and a $|log_2 ratio|>2$ in at least 1 array were selected.

TABLE I.

SMD Method Iterative Method Proportional Change

α **-Factor**: 334 269 0.805

Elutriation: 179 135 0.754

CDC: 1204 1099 0.913

Note that the Iterative method consistently reduces the number of genes that pass the filters. It also consistently lowers the standard deviation of the log_2 ratios in these studies. It does not, however, consistently improve the global

correlation between the $\log_2$ ratios of any two arrays.

Examples of Changed Arrays

Column 1:**SMD Method** Column 2:**Iterative Method**

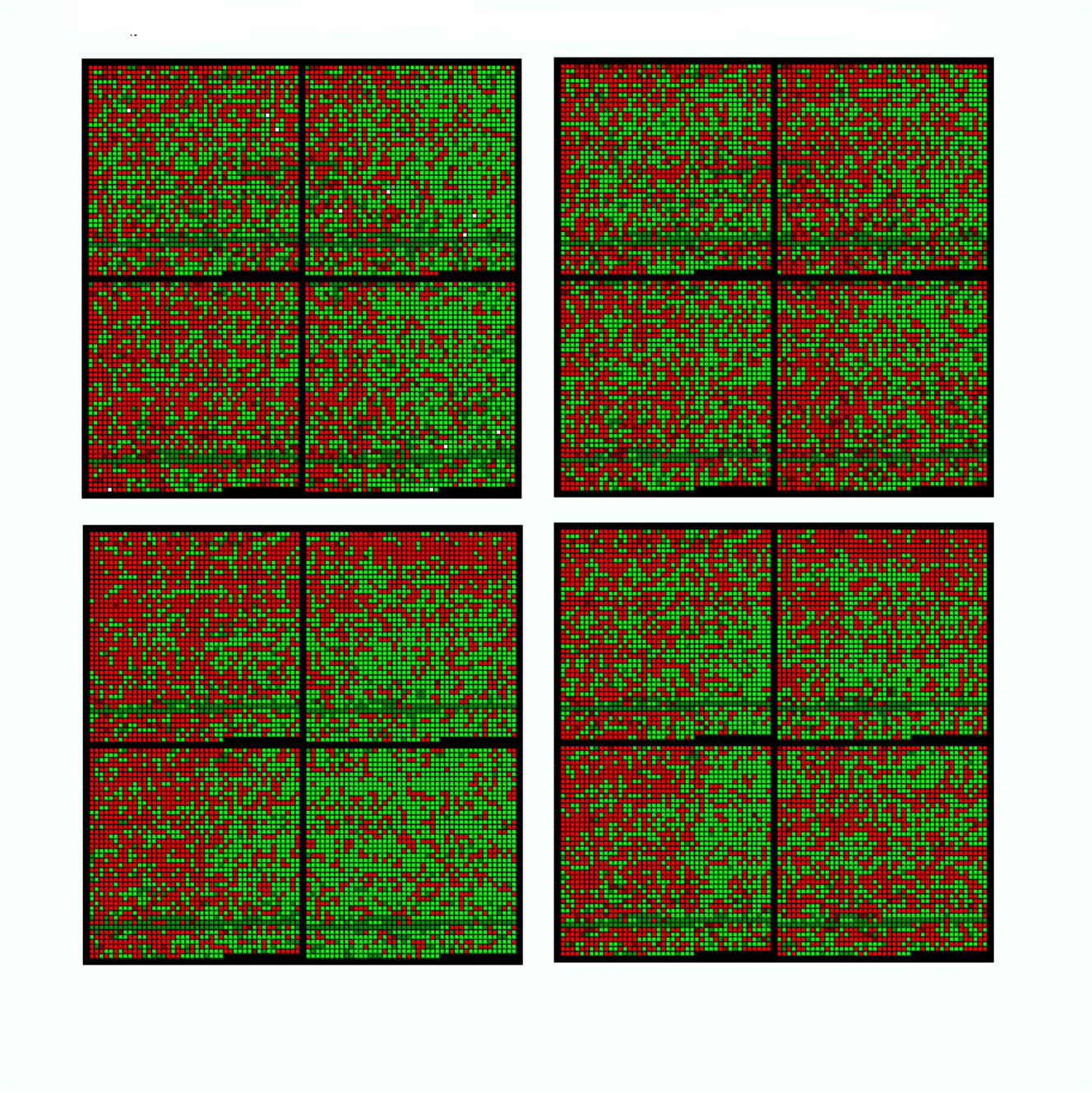

Figure 2. *The plots below show the spatial pattern of $\log_2$ ratios on two Elutriation arrays (SMD EXPID 56 (**row B**) and 57(**row A**) normalized by the SMD method on the left and by the Iterative method on the right. All spots with a $\log_2$ ratio greater than 1 appear in red. All spots with a ratio below 1 appear in green. Black spots indicate a flagged spot, white spots have a ratio of 1. Note that the iterative method (Column 2) partially corrects the spatial bias seen in the SMD method (Column1)for both expt. 56 and 57.*

Sequence Similarity in Yeast Arrays

The degree to which cross-hybridization might influence microarray expression data was also examined. First, a preliminary analysis was performed that related sequence similarity to the degree of correlation between expression profiles. Several assumptions are made. First, it was assumed that the full length ORFs available from SGD (*Saccharomyces* Genome Database) approximate the targets actually used on the microarray. This assumption is deemed reasonable, as yeast primer pairs were designed to include as much of the ORFs as possible (Gavin Sherlock, pers. comm.). Second, it was assumed that the degree of sequence similarity between a pair of sequences, as measured by an alignment program such as BLASTN, would approximate the degree of cross-hybridization between those sequences.

First, 2,690 ORFS were selected from the original 6,178 yeast ORFs. The selected ORFS were those with the fewest missing expression data values (that is ORFs with greater than 8 missing values across the 62 experiments were excluded). For all pairs of the 2,690 ORFs, the correlation coefficient between the expression profiles was calculated and a BLASTN alignment of the sequences created. For all pairs of ORFs with some degree of homology, the correlation coefficients were extracted and are plotted as two histograms in Figure 2. ORF pairs are divided according to their BLASTN e-values. Correlation coefficients for ORF pairs with BLASTN e-value greater than 1×10^{-4} are shown in white and those with BLASTN e-value less than 1×10^{-4} are in red.

Relatively few ORF pairs showed significant sequence similarity. 1991 ORF pairs had e-values greater than 1×10^{-4} and 59 pairs had e-values less than 1×10^{-4}. The set of 1991 ORF pairs had a mean pairwise correlation coefficient of 0.036, whereas the set of 59 ORF pairs with lower e-values had a mean pairwise correlation coefficient of 0.419.

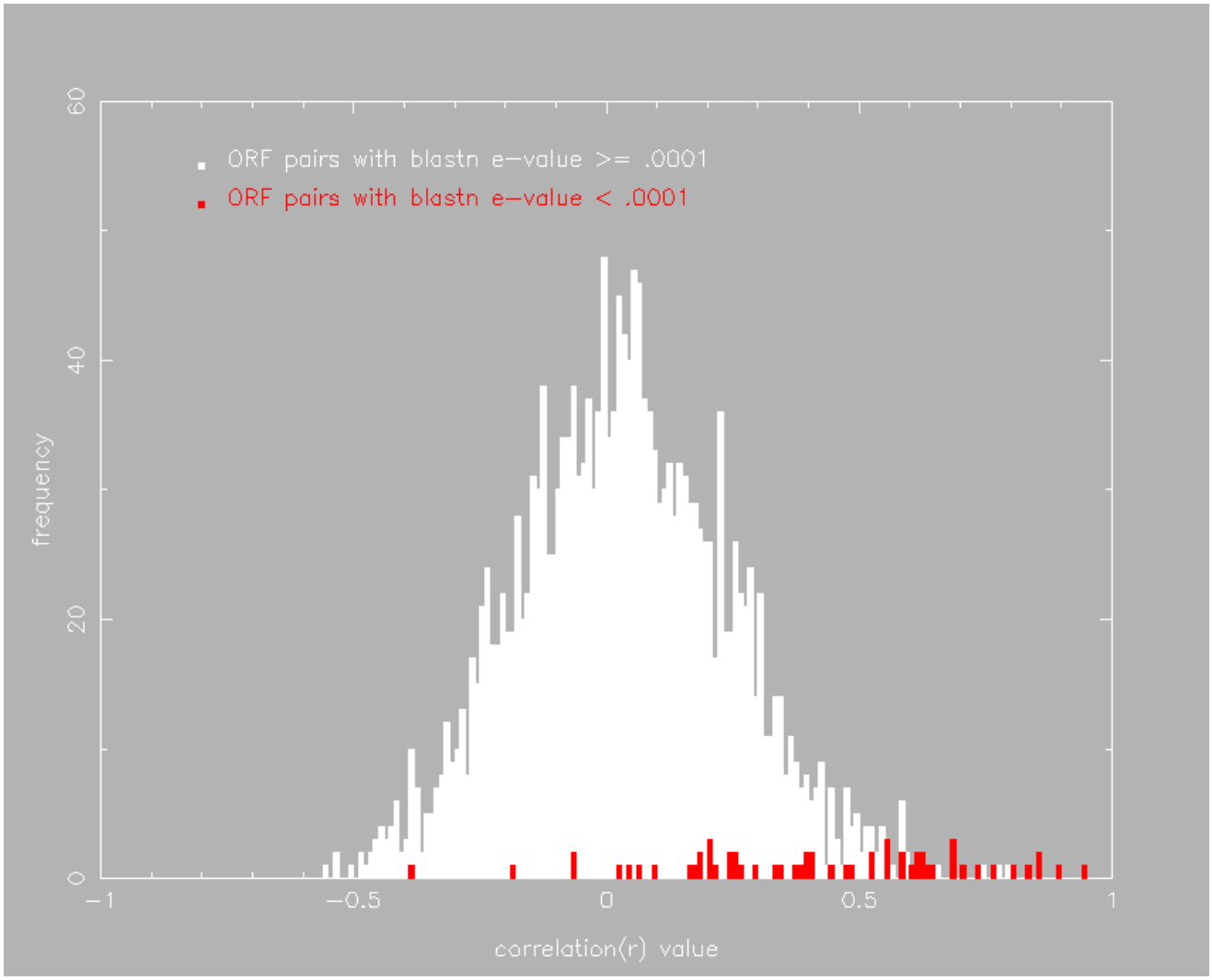

Figure 3. *Pairwise correlation coefficients of the expression pattern across 62 experiments of yeast ORFS. Red comparisons are highly similar pairs. Note the relative rarity of cross homology and the relative high degree of co-expression amongst the highly similar ORFs.*

Despite the small numbers, it appears that ORF pairs with a higher degree of sequence similarity are also more likely to exhibit a higher degree of correlation between their expression profiles. The e-value indicates, but does not prove cross-hybridization. It is also possible that genes with high sequence similarity may have similar function and therefore may be authentically co-expressed. Cross-hybridization and the degree to which this may confound results from genome-wide microarray experiments should definitely be considered in the design of future microarrays by printing gene specific probes used wherever possible.

Calculating the Weights

Weights were assigned to only those genes, which passed two criterion. The pairwise expression with another ORF had to exceed 1e-4 and their BLASTN score had to exceed 100. The BLASTN score was the more

stringent criterion, resulting in no expression correlation below 1e-21. 782 pairwise comparisons passed this cut off, representing some 678 ORF's. Weights were first calculated for each pairwise comparison. If an ORF was part of more

than one pairwise comparison then the weights were multiplied. Weights were calculated as 0.5 + 0.5(minimum exp/exp). In this case the minimum exponent of the data set was -21. If a given pairwise correlation value was 1e-42 then the weight would be 0.5 + 0.5(-21/-42) = 0.75. The maximum weight was 1 and the minimum weight was 0.16. This method is one of many that should give a reasonable approximation of the range of cross-hybridization. However, a method based on empirical evidence of cross-hybridization would be preferable.

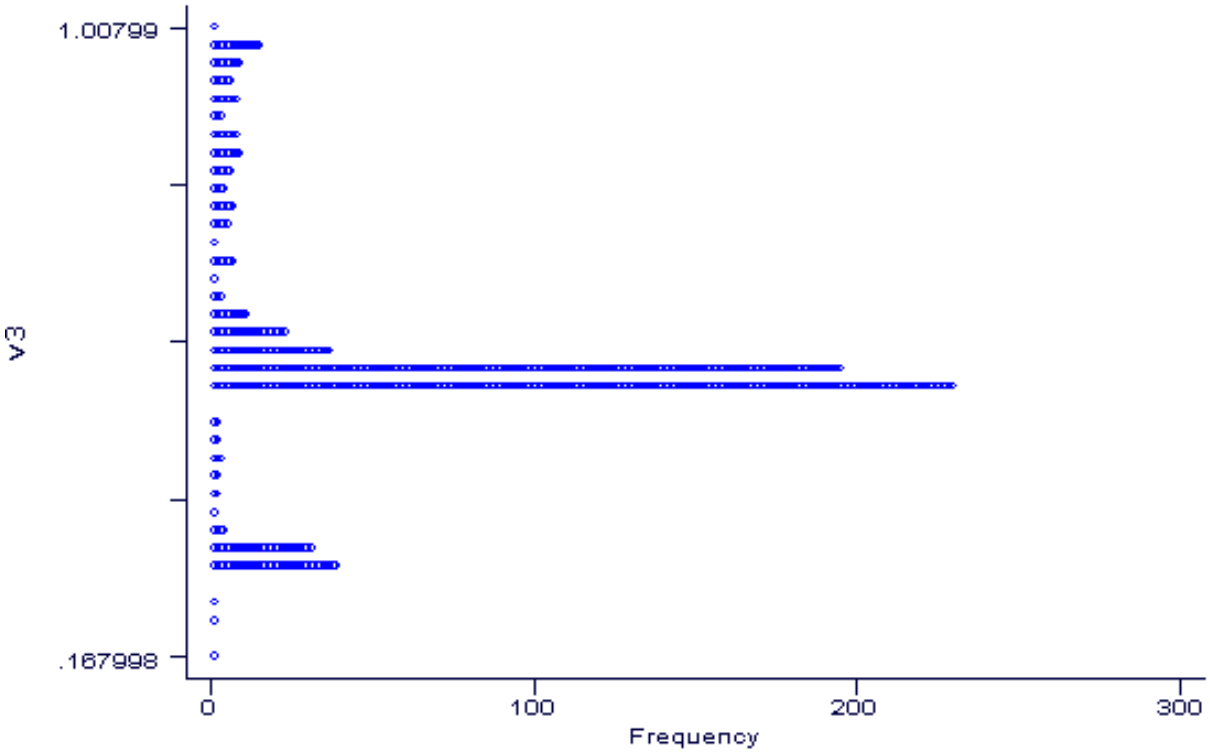

Figure 4. *Histogram of the weights applied to potentially cross-hybridizing genes. Note that most weights were nearly 0.5, and only a small sub-population is weighted less than 0.25.*

Determining Best Practice

Best practice of microarray data analysis is directly tied to the application of the data. If the arrays are to be used as rapid screening tools then sophisticated normalization and analysis may not be necessary. If, however, the object of the experiment is to model subtle biological patterns across gradients of time or treatment, much more complex analysis is required. Furthermore, while statistical measures are useful in measuring and correcting some errors, the most accurate means of determining gene transcript behavior will require empirical evidence. When DNA microarrays are designed to assist the statistical analysis, best practice can be achieved. For example, the replicate printing of a core control group of elements in several locations throughout an array would greatly simplify detection of spatial bias. Also, doping controls that consist of a class of non-homologous RNA transcripts could serve as independent verification of normalization methods. Finally, the correlation of expression amongst elements with high degrees of cross-homology does not prove cross-hybridization. In summary, the combination of careful array design, empirical verification and accurate mathematical models of error will result in the best practice of microarray data analysis.