Yeast through the ages: A statistical analysis of genetic changes in aging yeast

A. Wise
J. Hardin
L. Hoopes

October 6, 2005

Microarray technology allows for the expression levels of thousands of genes in a cell to be measured simultaneously. The technology provides great potential in the fields of biology and medicine as the analysis of data obtained from microarray experiments gives insight into the roles of specific genes and the associated changes across experimental conditions (e.g., aging, mutation, radiation therapy, drug dosage, ...). The application of statistical tools to microarray data can help make sense of the experiment and thereby advance genetic, biological, and medical research. Likewise, microarrays provide an exciting means through which to explore different statistical techniques.

Our paper focuses on the analysis of data from a yeast DNA microarray experiment. The biological question that motivates our research is "What genetic changes in yeast happen over time?" In order to explore the research question of interest we first standardize the data to correct for errors that arise in the data due to biases from the complex microarray procedure.

Once we have data that accurately depicts the natural variability in the genes and arrays (as opposed to variability due to technical aspects of the microarray chip), we can focus on our primary interest: the analysis of yeast gene expression to further uncover the quantitative relationship between the gene expression levels and the generation (age) of the yeast cell. We use a statistical tool called predication analysis for microarrays (PAM) [3]; PAM is a classification tool that provides insight into different groupings of the generations of yeast. PAM isolates and identifies specific genes using a threshold value and creates a model to predict the generation for an independent sample array. We also attempt to improve the results obtained from PAM by using t-tests to pre-filter genes before introducing them into the PAM model.

Background Biology and Data

The yeast data were collected by Professor Hoopes of the biology department at Pomona College. Yeast are single cellular organisms, so for a given yeast sample all the genetic information is isolated to one cell. To study the genetic effects of aging, yeast were obtained from generations (g) 1.5g, 12g, and 18g. The genes from the experimental cells were then analyzed through 21 array experiments (see Table 1). Generation 1.5 is a mixture of cells from the first and second generation yeast cells, which is a consequence of the challenge for biologists to isolate cells that are strictly one generation old.

For each of the microarray experiments, 'spotted arrays' were used. On each glass slide (array) there is a single spot which measures the expression levels for a specific gene. The gene expression

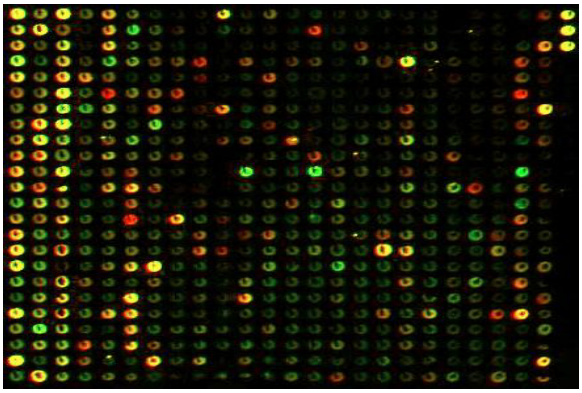

Figure 1: Segment of a typical microarray chip after being scanned. The colors reflect the amount of fluorescence or gene activity. Each spot represents a different gene.

levels for 7392 'spots' were measured for each of the 21 arrays. Of the 'spots', 6420 spots contained genetic information of interest, the remaining spots were either control genes or spots on the array that were empty of genetic information. Throughout the paper, we'll refer to 6420 spots of interest, though in some situations we analyzed fewer than 6420 values due to missing data. To measure the expression level of the genes, two mRNA samples are reverse transcribed into cDNA samples for a specific cell and compared on a single slide. The first cDNA sample is from the generation of interest (1.5g, 12g, 18g) and is labeled using a fluorescent dye of one color (typically red); the second sample is always from the base generation (1.5g) and is labeled using a different fluorescent dye (typically green). Note that, due to the characteristics of the dyes, the red and the green dyes will emit slightly different levels of intensities for the same gene expression level. To control for biases in dye intensity due to the differences in the dye color, for certain arrays the labels for the red and the green dyes were switched (see Table 1). Throughout our paper, when we refer to the red (R) and the green (G) dye intensities, we mean the dyes representing the interest and base generation respectively.

On each of the 21 arrays, for each of the 6420 spots of interest, the intensity ratio between the generation of interest and the base generation is measured; the ratio reflects the change in the gene activity between the two generations. For example, consider array 10 in which the genetic information from 1.5g was labeled with green dye and the genetic information from 12g was labeled with red dye. For a given spot, the measurement of red dye intensity reflects the amount of that gene present in 12g yeast cells; the measurement of the green dye intensity reflects the amount of that gene in 1.5g yeast cells. For a particular spot, if the red dye intensity is greater than the green dye intensity, the expression level for the 12g is greater than the 1.5g; the ratio of the red to the green dye (or 12g to 1.5g) is greater than one and so the gene measured at the spot of interest has become more active as the yeast has aged to its 12g. Along with removing control and empty spots, we also preprocessed the data by removing data from spots that were flagged as unacceptable or that had almost no observable signal. See figure 1 for an example of a section of a typical microarray.

Normalization

Before we can begin to compare and analyze the data from the different arrays, the values are adjusted to remove biases. The microarray procedure is subject to biases both within arrays, for

Array Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Generation	1.5	1.5	1.5	1.5	1.5	1.5	1.5	12	12	12	12	12	12	12	18	18	18	18	18	18	18
Base Dye Color	R	G	R	R	R	R	R	R	R	G	R	G	R	G	R	G	R	G	R	G	R

Table 1: The generation and base dye color of each of the 21 arrays. If the base dye color is red (R), then the expression level of the generation of interest was measured by the green(G) dye intensity.

example, the location of the gene on the array and the overall dye intensity on each spot, and between the arrays, for example, the specific slide a cell is on. We can apply a normalization technique that uses loess smoothing [2] and scaling to minimize systematic location variations in the gene expression levels. Once the spatial location and dye biases have been accounted for, expression levels across the slides and the biological differences between the genes and samples can be better analyzed.

MA-plots [4] are a helpful way to observe the biases described previously and assist in the normalization of the data. To create the plot, we first transform the data into values representing dye intensity ratios (M) and overall dye intensity (A). We want to modify our data so that our ratios of interest (M) are not dependent on overall intensity (A). The adjustment will also help remove the effects of the differences between the red and green dye. For each of the 21 arrays, the red (R) and the green (G) intensity values have been measured for each gene (i.e., at each spot.) Therefore, for each gene on each array we set:

$$M_{ij} = log_2(\frac{R_{ij}}{G_{ij}}) \text{ and } A_{ij} = \frac{1}{2} log_2(R_{ij} \cdot G_{ij}) \tag{1}$$

where $i = 1, 2, \ldots, 21$ is the array and $j = 1, 2, \ldots, 6420$ is the gene. Note that for samples with dye swaps, $M_{ij} = log_2(\frac{G_{ij}}{R_{ij}})$. Hence, M_{ij} represents the log of the ratio of the dye intensities of interest while A_{ij} represents the average log intensity (overall brightness) for a specific gene on a specific array. Because M_{ij} is the log ratio, $M_{ij} = 0$ reflects no change in gene intensity over the generation. In the calculation of A_{ij}, the $(\frac{1}{2})$ term is used so that A_{ij} will have the same range as $log_2(R_{ij})$ and $log_2(G_{ij})$, while log-base two is used because it is convenient to express ratios in powers of two (twice as big, four time as big,...). After transforming the yeast data, for each array i a scatter plot of the 6420 M_{ij} values is plotted against the corresponding A_{ij} values. In total there are 21 plots with 6420 spots, assuming that no data points are missing. See figure 2(a) for an example MA-plot of array 19.

Box plots of the unscaled M values for each of the 21 arrays, as shown in figure 3(a), reveal the bias that exists. Note how the spreads within a given generation vary. Also, the arrays for which the red dye was used to measure the generation of interest have lower centers than arrays for which the green dye was used. The difference in center for the dye-swapped arrays results from the greater intensity of the green dye, so that even if a gene's ratio should be zero due to no change, if the green measured the base generation, the dye ratio will be less than zero. Such a bias will be corrected.

The normalization technique applied to the data expressed in the MA-plots requires multiple steps [1]. The first step applied was within-print tip group normalization. When our cDNA array was printed, the spots were broken up into 16 groups based on their location. Each group contains $1/16^{th}$ of the genes and uses a different print tip for printing the spots onto the array. Because the 16 print tips give spatial information about the chip layout, we can compare print-tip groups in order to reduce variability. For example, a print-tip group located on the edge of the chip might be slightly drier than an internal print-tip group causing the external group to have a lower overall signal. By normalizing the chip within print-tip groups, we reduce spatial variability. Within-print-tip group normalization removes the intensity dependence of M and the unwanted variability within each print-tip group that results from the location of the print-tip group on the array and the amount

of dye the group received by assuming that the genes should be centered at $M = 0$. If the data values were all without the potential errors caused by the aforementioned sources of variability, or "true", then the intensity log ratios M in an MA-plot should be symmetrically distributed around the horizontal line at zero, indicating no ratio dependence on intensity. The line should be centered at $M = 0$ as we assume that most genes do not change from generation to generation, so the R to G ratio is 1 for most genes. Figure 2(b) shows the box plot for each of the 16 print-tip groups of array 19. The non-zero centers and varying spreads show the effects of bias and need for normalization.

Within print-tip group normalization uses a loess smoothing technique. The specific smoother used on our data was derived by Cleveland and is often referred to as a locally weighted running-line smoother [2]. The idea behind the loess smoother is that a regression line will not accurately fit the data over the entire range of the explanatory variable (here, A) because the data are not linear; however, over small intervals of A the data model can be approximated with a linear regression line. The regression lines for each print-tip group on each array are calculated by locally weighting that print-tip group's data at given points, creating a local linear regression line, and then connecting the local regression lines to create a smoothing spline. Due to the local weighting, the loess smoother is robust as it will not be greatly affected by a small percentage of differentially expressed genes that appear as outliers in the MA-plot. Figure 2(a)shows the 16 smoothing splines for array 19.

The adjusted data values are calculated to be the difference between a spot's initial M value and the M value from the corresponding smoothing spline at A, thus normalizing each gene within its print-tip group. The results of the normalization are shown for array 19 in figure 2(c). As can be observed from comparing the box plots of the print-tip groups in figure 2(c) to those in figure 2(b), the print-tip normalization centers the data around zero. The within print-tip group normalization removed the print-tip and some of the spatial bias as well as the dye intensity bias. However, we still note that there are some differences in the spread of the print-tip groups.

To more thoroughly remove the spatial bias we scale across each of the print-tip groups. Scaling data that has undergone within print-tip group normalization is called scaled print-tip normalization. The scaling method assumes that the unscaled log ratios from each print-tip group k, $k = 1, 2, \ldots, 16$, follow a distribution with mean zero and variance s_{ik}^2. Furthermore, as mentioned before, we assume that the print-tip groups should have the same variability across a given array. If we let $s_{ik}^2 = a_{ik}^2 \sigma_i^2$, where σ_i^2 is the variance of the log ratios if there is constant scale within the i^{th} array and a_{ik}^2 is the scale factor for the k^{th} print-tip group, then to scale the data we want to find a_{ik}^2 so that s_{ik}^2/a_{ik}^2 is constant, and all print-tip groups on array i have the same variance, σ_i^2. To estimate a_{ik}, the median absolute deviation (MAD) is used. MAD is a robust alternative method to the maximum likelihood estimate for a_{ij}, which is affected by outliers. For each print-tip group on each array the MAD is calculated as the median of the absolute deviation between each M value within a print-tip group and the median M value over the entire print-tip group.

Once all the $\hat{a}_{ik}$ have been calculated, the print-tip groups on all of the arrays can be scaled. Figure 2(d) displays the results of the scaled print-tip normalization for array 19. We can see how normalization remedied the varying spread as seen in 2(c). Furthermore, as depicted in figure 3(b) as compared to figure 3(a), the loess smoothing technique combined with scaled print-tip normalization consistently normalizes the data within each of the arrays. (Note: we did not try to create constant scale across arrays as that would not be meaningful for data of different generations.) Through the methods described and illustrated by the figures, it can be seen that within a given array, the average log ratio values, M, are no longer dependent on the log intensity values, A. We have also removed unwanted data biases due to differences in location on array, dye type, and print-tip group. We can also see how spreads for the arrays within a given generation are now more similar. Generation 1.5 arrays have very low spreads around zero as would be expected given that the red and green dye are both measuring 1.5g. However, that there is a spread at all (and not simply a ratio of 1 for each

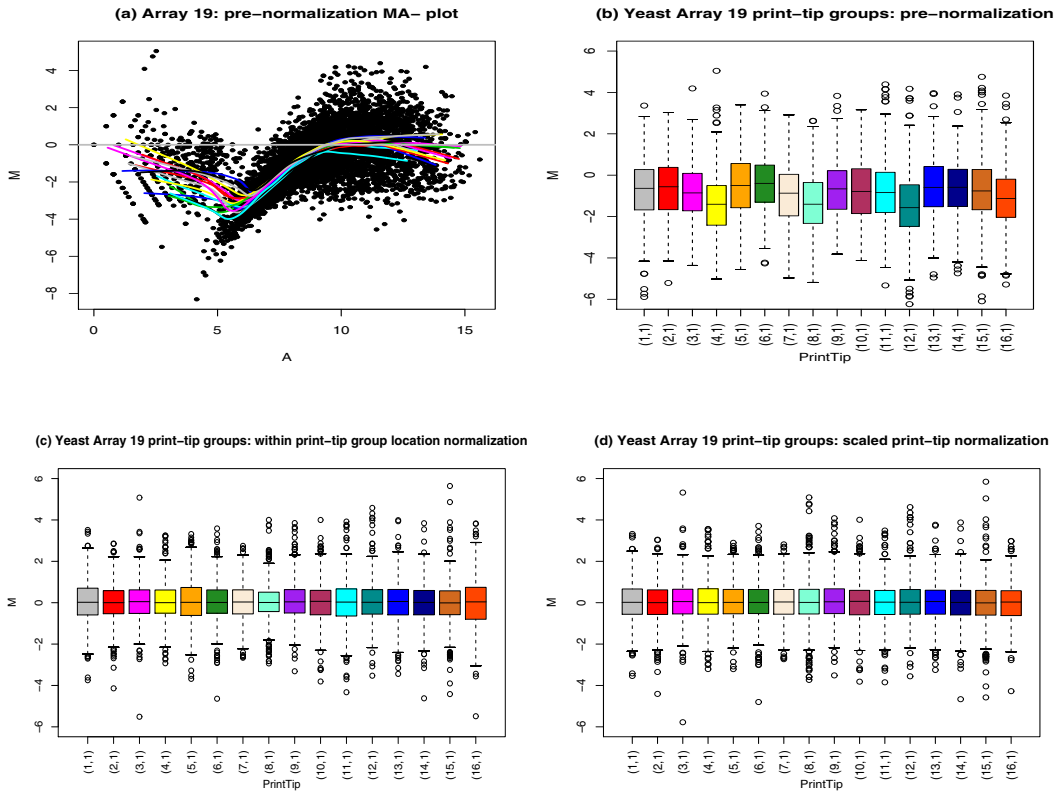

Figure 2: Plots of array 19 and the 16 print-tip groups. a) The MA-plot of array 19 before normalization with all 16 print-tip group smoothing splines shown. b) Box plots of the 16 print-tip groups of array 19 before normalization. c)Box plots of the 16 print-tip groups after within-print-tip normalization. d)Box plots of the 16 print-tip groups of after scaled print-tip normalization.

spot) shows that not all of the variability due to the microarray procedure can be removed. The normalization technique has improved the quality of the data and allows for the natural variability of the data to be better observed.

PAM

Prediction analysis for microarrays (PAM) is a classification technique used for class prediction [3]. For example, a PAM model can predict class membership (i.e., 1.5g, 12g, or 18g) for a new observation given the data on the 21 arrays. For each gene, PAM uses gene expression data to calculate a shrunken centroid for each class and for all classes overall. PAM performs soft-thresholding to shrink the class centroids toward the overall centroids. For example, consider gene # 47; PAM calculates $\text{cen}_{1.5}$, cen_{12}, cen_{18}, and cen_{ov} (centroids for classes 1.5g, 12g, 18g, and overall, respectively.) If cen_{12} and cen_{18} are close to cen_{ov}, but $\text{cen}_{1.5}$ is sufficiently far from cen_{ov},then the values in groups 12g and 18g are replaced with the value of cen_{ov} while the values in group 1.5 are left alone, thus "shrinking" groups 12g and 18g toward the overall centroid. Determining "close to the

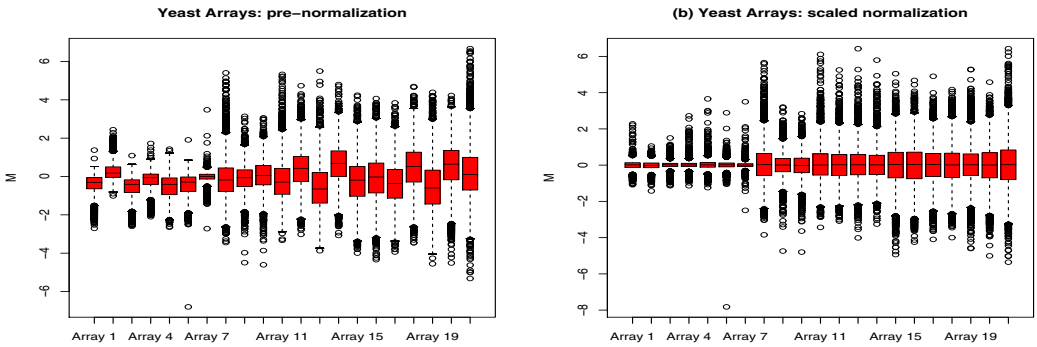

Figure 3: Box plots of M for all 21 arrays. a) Before normalization. b) After scaled print-tip normalization. Notice both the centering and the spread of the box plots.

overall centroid" is done through a threshold parameter which is adjusted in the model building process. Sometimes all values are close to the overall centroid and PAM substitutes values in all groups with cen_{ov} which eliminates all information given by that gene. The soft-thresholding technique will reduce the number of genes used in the class prediction as well as ignore the insubstantial deviations of a class from the overall centroid. Hence, we eliminate excess noise from genes that do not vary (or vary less relative to other genes) across the classes. For each threshold, the class of a new observation is predicted by the closest shrunken centroid. PAM also identifies the specific genes that most determine the centroid (genes that had group centroids far from cen_{ov}.

As we are interested in studying the genetic differences in the generations of yeast, our data will be broken into three classes, one for each generation. We applied PAM class prediction 21 times, each time setting aside a single array from the sample set to be our "new observation", and then using the remaining 20 samples to predict the generation of the "new observation" (leaving out one observation at a time to determine misclassification rates is a standard application of cross validation.) We can illustrate the results we obtained from PAM with a misclassification plot, see figure 4. The figure shows three lines, one for each generation. The Y-axis represents the misclassification error, or the proportion of times that the model misclassified the test samples from a generation. The X-axis shows how the misclassification error of each generation changes as the threshold is raised and the number of genes which remain in the model decreases. We can see that for 1.5g and 18g, until the threshold becomes quite large, the PAM model predicts each of the samples accurately. However, for 12g, for most threshold values, the majority of samples were misclassified. To observe the actual classifications for each generation, we can look at the misclassification error at a single threshold (we use a threshold of 4.75) through the use of a confusion matrix (see table 4) which shows that 12g is misclassified as both 1.5g and 18g.

One technique we employed in an attempt to improve the classification of 12g was incorporating t-tests into the PAM model. For each of the 21 test runs, before the prediction and the threshold was applied, a series of t-test (comparing each group to every other group) was run for each gene using generation as the response variable thereby measuring the effect of generation on the intensity of a gene. Genes for which generation had a significant effect ($p < 0.01$) in at least one t-test comparison advanced into the PAM cycle. We thought that the additional t-test criteria would help to further reduce the noise and help isolate genes that contribute to a more accurate PAM model.

The results of the second set of runs are shown in figure 5 which looks similar to the misclas-

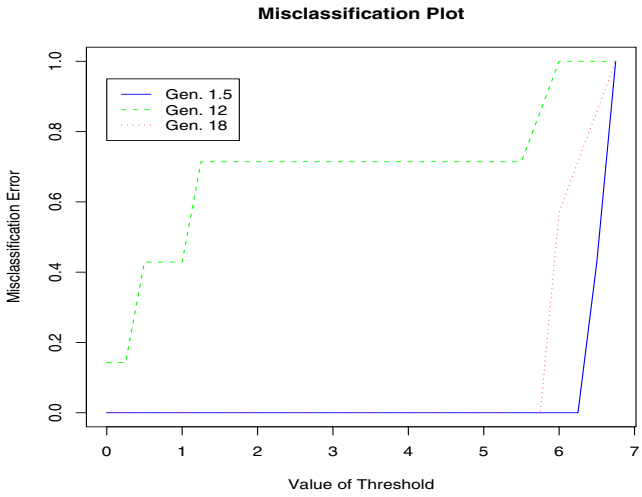

Figure 4: Misclassification error over various thresholds which induce different numbers of genes used in predicting the generation of each yeast sample.

generation	predicted			class
observed	1.5	12	18	error rate
1.5	7	0	0	0
12	2	2	3	0.714
18	0	0	7	0
Overall error rate = 0.238				

Table 2: The confusion matrix for the data at a threshold of 4.75. Fewer than 200 genes made the threshold cut for each of the 21 iterations of the PAM classification models. The values represent the expected generation on the top, and the observed generation on the left. The overall error rate is 0.238.

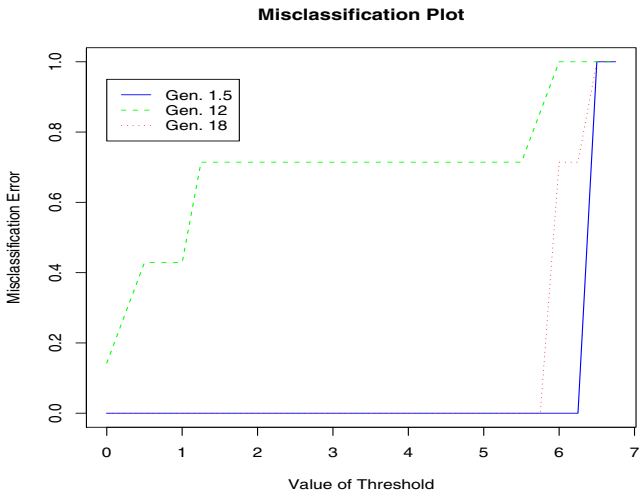

Figure 5: Plots of the misclassification error over the various thresholds which correspond to different numbers of genes used in predicting the generation of each yeast sample after the t-tests were applied.

sification plot in figure 4. The addition of the t-tests into the procedure does not seem to have improved the quality of the output as can be seen from the misclassification of 12g in both models. One reason the t-tests may have failed to improve the PAM model is that if there are much larger differences across 1.5g and 18g (than across either compared with 12g), the genes that differentiate those groups will be chosen in the PAM model. Furthermore, in both examples it may be the case that with such a small sample size of only 20 arrays at a time, it is difficult to accurately describe the differences across all 3 generations.

Discussion

This paper has pointed out several issues concerned with the analysis of microarray with an emphasis on normalization and class prediction. The normalization and PAM techniques provide useful tools for preparing and performing statistical analysis on microarray data. PAM was very effective at predicting 1.5g and 18g. However, arrays from 12g do not provide very promising results. The addition of t-tests into the PAM procedure did not seem to improve the classification outcomes. Possibly another microarray experiment could show whether t-tests could, in practice, be used as an additional noise filter.

The true test will be for genes which passed the higher thresholds to be further studied in the lab to provide stronger evidence of true genetic changes across time. Statistical analyses on microarrays will enable the biologists to focus on a few of the yeast genes, a much less time consuming task than studying thousands of genes individually.

Acknowledgements

The project was supported in part by an NIH AREA grant (#1 R15 AG021907-01A1).

References

[1] Bengtsson, H., Calder, B., Mian, I.S., Callow, M., Rubin, E. and Speed, T.P. (2001) Identifying differentially expressed genes in cDNA microarray experiments: making aging visible", *Science of Aging Knowledge Environment*, 12, vp8.

[2] Cleveland, W. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association,* 74, 829-836.

[3] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression, *PNAS*, 99, 6567-6572.

[4] Yang, Y., Dudoit, S., Luu, P., David, L.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research*, 30, e15.