



A comparison of statistical methods for analysis of high density oligonucleotide array data

Dilip Rajagopalan

Bioinformatics Sciences, GlaxoSmithKline Pharmaceuticals R&D, 709 Swedeland Road, King of Prussia, PA 19406-0939, USA

Received on November 25, 2002; revised on February 27, 2003; accepted on March 7, 2003

ABSTRACT

Motivation: Gene expression profiling has become an invaluable tool in functional genomics. A wide variety of statistical methods have been employed to analyze the data generated in experiments using Affymetrix GeneChip[®] microarrays. It is important to understand the relative performance of these methods in terms of accuracy in detecting and quantifying relative gene expression levels and changes in gene expression.

Results: Three different analysis approaches have been compared in this work: non-parametric statistical methods implemented in Affymetrix Microarray Analysis Suite v5.0 (MAS5); an error-modeling based approach implemented in Rosetta Resolver[®] v3.1; and an intensity-modeling approach implemented in dChip v1.1. A Latin Square data set generated and made available by Affymetrix was used in the comparison. All three methods—Resolver, MAS5 and the version of dChip based on the difference between perfect match and mismatch intensities—perform well in quantifying gene expression. Presence calls made by MAS5 and Resolver perform well at high concentrations, but they cannot be relied upon at low concentrations. The performance of Resolver and MAS5 in detecting 2-fold changes in transcript concentration is superior to that of dChip. At a comparable false positive rate, Resolver and MAS5 are able to detect many more true changes in transcript concentration. Estimated fold changes calculated by all the methods are biased below the true values.

Contact: dilip.2.rajagopalan@gsk.com

INTRODUCTION

Gene expression profiling has become an invaluable tool in functional genomics. Using cDNA (Schena *et al.*, 1995) or high-density oligonucleotide (Lockhart *et al.*, 1996) microarrays, it is possible to simultaneously measure the relative abundance of thousands of mRNAs in a cell.

High-density oligonucleotide Gene Chip[®] microarrays are manufactured by Affymetrix for several species including rat and human. These microarrays are divided

into several hundred thousand probe cells grouped in pairs. Each pair has a Perfect Match (PM) and a Mismatch (MM) cell. The PM cell contains oligonucleotide sequence designed to hybridize to a specific target sequence. The corresponding MM cell contains an altered version of the PM sequence designed to increase the sensitivity of the method by capturing non-specific hybridization. A group of probe pairs designed to measure the abundance of a specific target molecule is known as a probe set. On the HG-U95A human arrays used to generate the data analyzed in this study, each probe set typically consists of between 16 and 20 probe pairs. Including control sequences, there are a total of 12 626 probe sets on the HG-U95A array.

In a typical eukaryotic gene expression profiling experiment, a mixture of labeled cRNA molecules isolated from a tissue of interest is hybridized to the array. The array is then scanned to determine fluorescent intensity at each probe cell which must be converted to a measure of relative transcript abundance. The analysis methods compared in this work take the information generated by the scanner and calculate a signal estimating relative transcript abundance. This process is known as absolute analysis. These methods can also detect differential expression by comparing arrays hybridized to different cRNA samples, and this is known as comparison analysis. The goal of this work is to compare three different types of analysis methods on a common data set and benchmark their performance. The analysis methods considered in this work are:

- (1) Non-parametric algorithms implemented by Affymetrix in Microarray Analysis Suite v5.0 (MAS5; Liu *et al.*, 2002; Hubbell *et al.*, 2002).
- (2) Algorithms built around an error-modeling approach implemented in Rosetta Resolver[®] v3.1 (Stoughton and Dai, 2002; Roberts *et al.*, 2000)
- (3) Algorithms built around an intensity-modeling approach implemented in dChip v1.1 (Li and Wong, 2001a,b)

These analysis methods are based on very different statistical approaches and benchmarking their performance

on the same data set provides a useful way to understand their relative strengths and weaknesses. Comparison analysis results of these methods are also compared to *t*-test results.

Lemon *et al.* (2002) compared the model-based intensity approach to Affymetrix MAS4, which is based on very different statistical methods than MAS5, and concluded that the model-based approach was superior. Irizarry *et al.* (2002) compared MAS5 to the model-based approach, and they also introduced a new expression measure in a model-based context. They concluded that 'there is no obvious downside to summarizing the expression level' by their new measure, relative to the measure proposed by Li and Wong (2001a), or to the results from Affymetrix MAS5. Building on these comparisons, the present work includes the error-modeling approach implemented in Rosetta Resolver[®] 3.1 along with the rank-based algorithms of Affymetrix MAS5 and the intensity-modeling approach of dChip 1.1.

DATA AND METHODS

Data used in the comparison

The human Latin Square data set generated by Affymetrix was used in this comparison, and detailed information on the data set is available at www.affymetrix.com. This data set consists of 14 labeled transcripts spiked at varying concentrations into a labeled mixture of RNA from a tissue in which the spiked transcripts are known not to be present. Each transcript concentration series or experiment consists of 13 transcripts spiked at successively doubling concentrations between 0.25 and 1024 *pM* and one transcript not spiked (concentration of 0 *pM*). Each RNA mixture was hybridized in triplicate. The concentration of each transcript was also varied in a systematic manner between successive experiments. Twelve of the transcripts double in concentration, one changes from 0 to 0.25 *pM*, and one changes from 1024 to 0 *pM*. There are a total of 14 experiments over which each of the transcript concentrations are varied between 0 and 1024 *pM*.

This data set enables evaluation of the various analysis methods by comparing the trends in calculated transcript abundance with the known trend in transcript concentration. By comparing results from one experiment to another, this data set can also be used to evaluate how the various analysis methods perform in detection and quantification of differential expression. Comparing sequential experiments, all of the spiked transcripts should be detected as being changed in concentration, and for the 12 transcripts that double in concentration, the concentration ratio or fold change computed by the methods should be close to 2. Since the concentrations of only 14 out of a possible 12 626 targets vary in this data set, it does not pose a severe test of normalization algorithms, a subject that

is being addressed in other work (Hoffmann *et al.*, 2002; Bolstad *et al.*, 2002). Affymetrix recommends discarding data for 2 of the 14 spiked transcripts due to data quality concerns. This data set has been used by Affymetrix (Hubbell *et al.*, 2002; Liu *et al.*, 2002) to assess the performance of their non-parametric algorithms implemented in MAS5.

Non-parametric algorithms

In their MAS5 software, Affymetrix (Liu *et al.*, 2002; Hubbell *et al.*, 2002) has implemented algorithms for absolute and comparison analysis that are based on non-parametric statistical techniques. The normalization option of global scaling of all probe set intensities to a target value of 35 was used for the Latin Square analysis presented in this work.

In the absolute analysis, detection calls are made on the basis of a discrimination score R_i which is defined for each probe pair as

$$R_i = \frac{PM_i - MM_i}{PM_i + MM_i} \quad (1)$$

where PM_i and MM_i are the intensity of the perfect match and mismatch cells of probe pair i , respectively. Zero or negative values of this discrimination score imply that the target cannot be reliably detected, because the intensity due to specific hybridization at the perfect match cell is less than the mismatch cell intensity arising from non-specific hybridization. The detection call for a probe set is made via a Wilcoxon signed rank test to determine whether the mean value of the discrimination score over all probe pairs in the probe set is greater than 0.015.

The signal or expression level of the target molecule associated with each probe set is defined as the average value of $PM_i - CT_i$ over all pairs in the probe set. To avoid sensitivity to outliers, a one-step Tukey's biweight mean is used. CT_i is known as the contrast value and is usually equal to the mismatch cell intensity MM_i . However, to avoid computing negative values for signal that can arise when many probe pairs have MM_i values larger than PM_i , an imputed value of CT_i is used for such probe pairs. If only few of the probe pairs in a probe set have MM_i values larger than PM_i , then the CT_i values for these pairs are imputed from the corresponding PM_i values using the average ratio of PM_i/MM_i over the entire probe set. If, on the other hand, most of the MM_i values are smaller than PM_i , the corresponding CT_i values are simply set equal to a value that is slightly smaller than the corresponding PM_i , under the assumption that the detection call will generally determine that such transcripts are not present in the RNA mixture.

In a comparison analysis between a 'baseline' and 'treatment' array, a Wilcoxon signed rank test is done to compare the matched probe pair intensities from baseline to

treatment arrays and determine whether the target concentration increased, decreased or remained unchanged. To compensate for the limitations of the simple global normalization procedure discussed above, three such comparison analyses are performed. One is based on the normalized intensities derived from the scaling procedure, and the other two are based on perturbing the intensities on one of the two arrays up and down by a perturbation factor (default value of 1.1). The most conservative of the three calls made in this manner is selected as the comparison result. Thus, probe sets are declared unchanged if their intensities on the two arrays are close enough to each other such that a 10% perturbation up or down reduces the significance of the signed rank call.

To quantify the magnitude of differential expression, for each probe pair, the log ratio of PM_i-MM_i value from the treatment array to the corresponding value from the baseline array is first computed. A one-step Tukey's biweight mean of log ratios for the probe pairs within a probe set is then calculated.

As implemented in MAS5, these algorithms cannot be readily applied to comparison of two experiments with replicates. For the Latin Square data set where three replicates per experiment were available, these algorithms were applied by first forming all nine pairwise comparisons between the individual arrays for each pair of experiments. The consensus call was defined as the call made in the majority (five or more) of the individual comparisons. If no majority was found, the target concentration was determined to be unchanged. The consensus log ratio was found by averaging the ratio values for each probe set over all nine comparisons and taking the log.

Error-modeling approach

Error modeling is the central concept behind the algorithms implemented in Rosetta Resolver[®] v3.1 (Stoughton and Dai, 2002; Roberts *et al.*, 2000). Since replication error cannot be accurately estimated with few repeats typically performed in microarray experiments, an empirical intensity-based error model is employed to obtain a conservative estimate of signal variability. This model, which depends on array type, is fit using data from arrays hybridized to the same RNA sample.

In the absolute analysis, the signal for each probe set is computed by averaging the PM_i-MM_i values for the probe pairs after discarding values that lie outside 3 standard deviations from the mean. No attempt is made to adjust MM values that are larger than their corresponding PM values, so a negative signal may be estimated for many probe sets. As part of the absolute analysis, the error model is used to compute the variability of each probe set signal.

Similar to MAS5, Resolver also attempts to determine whether or not each target is present in the RNA sample. This is done by comparing the signal of each probe set

to the average signal of negative controls on the array. The error-normalized difference in signal between each probe set and the average of the negative control signals is used to determine whether the difference is statistically significant. The target is called present if the probe set signal is found to be significantly different than the signal of the negative control probe sets.

For comparison analysis between 'baseline' and 'treatment' experiments with replicate hybridizations, Resolver automatically performs comparison analysis between all pairs of baseline and treatment arrays and computes a consensus result for change calls and fold changes. For each comparison analysis, a single global normalization is performed between the two arrays, followed by smaller adjustments for regional effects within an array and non-linear effects. The fold change is computed as the log of ratio of the probe set level signals in the two arrays being compared. The error of this log ratio is computed by appropriately combining the modeled error of the individual signals and the variability of the signal within replicates. As the number of replicates available increases, the contribution of the modeled error decreases. The log ratio value and its error are used to determine whether the target concentration changed between the two experiments.

Intensity-modeling approach

A model-based analysis for intensity has recently been proposed by Li and Wong (2001a,b) and implemented in dChip 1.1. For every probe set in a group of arrays being analyzed, a model is estimated that captures the dependence of probe-pair level intensities on expression levels of the target molecule in the samples hybridized to each array.

For every probe set in a group of $i = 1, 2 \dots I$ arrays, an intensity model is constructed of the form

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \epsilon_{ij} \quad (2)$$

where PM_{ij} and MM_{ij} are the perfect match and mismatch intensities of the j th ($j = 1, 2 \dots J$) probe pair in the i th array, θ_i is the expression index or signal of the probe set in the i th array, ϕ_j is the coefficient that captures the dependence of cell level intensities of the j th pair on the target abundance, and ϵ_{ij} is the error term. The model parameters are fit using data from a group of representative arrays as a training set. For the Latin Square data set analyzed in this study, data from all 59 hybridizations was used to fit the model and calculate the probe-pair coefficients and estimated target abundance.

Once the coefficients (ϕ_j) have been estimated for a large enough group of arrays, they can be directly used to quantify target abundance on new arrays without the need for the model fitting process. The signal or target

abundance for a new array is given by

$$\theta = \frac{\sum_j \phi_j (PM_j - MM_j)}{J}. \quad (3)$$

The residuals between the model estimate and the actual values of $PM_j - MM_j$ are used to compute a standard error for the above estimate. This estimate of standard error is useful in probe selection and outlier detection. It is also used to assess the reliability of clustering results (Li and Wong, 2001b) as well as in comparison analysis. Presence calls made in dChip are not considered in this work, because they are based on the procedure implemented in Affymetrix MAS4 which has been shown to be inferior to the presence call algorithm used in MAS5. (Liu *et al.*, 2002).

The probe set signal computed using the above model can be negative if enough mismatch cell intensities are larger than the perfect match intensities. To circumvent this problem, dChip provides an analysis option based on the perfect match intensities alone in which the model takes the following form:

$$PM_{ij} = \theta_i \phi_j + \epsilon_{ij}. \quad (4)$$

This model is termed the PM-only model to distinguish it from Equation (2) which is the PM-MM model.

For comparison analysis, dChip provides two methods to identify targets whose change in abundance is statistically significant. The first method is based on assuming that the ratio of abundance estimated using the model-based procedure follows the χ^2 distribution (Li and Wong, 2001b) and thus obtaining a confidence interval for the fold change. Targets whose confidence interval of expression ratio does not include unity are designated as significantly changed in expression. The second option, which is used in this work, is based on an unpaired t -test on the means. If replicates are not available, the standard error calculated from the model residual is used in the t -test. This estimate of standard error is also used when replicate baseline and treatments are available to downweight unreliable signals when computing group means and standard errors for the t -test.

RESULTS

Absolute analysis

The target abundance level or signal estimated by each of the methods is plotted against the actual target concentration in Figure 1. The calculated signal for each data point represents a global average over all transcripts in the various experiments that are spiked at the concentration value shown on the x -axis.

The most common practice in microarray analysis is to assume that the ratio of estimated transcript abundance

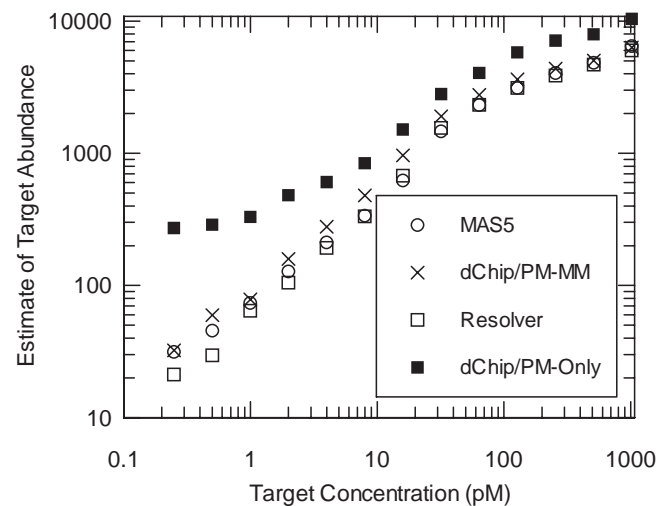


Fig. 1. Comparison of estimated target abundance plotted against actual target concentration.

is equal to the true ratio of transcript concentration. For this to be valid, the log-log plot in Figure 1 should be linear across the entire concentration range with a slope of unity. All the analysis procedures compared in this work show evidence of saturation at high concentration. All four methods appear to saturate at roughly the same concentration (between 64 and 128 pM), indicating that the analysis methods are not responsible for this saturation effect. Below this concentration, MAS5, Resolver and the PM-MM version of the dChip model show a linear regime essentially down to the lowest concentration of 0.25 pM . However, the PM only version of the dChip model performs poorly and saturates at the low end as well. The linear range calculated by this model is extremely narrow. The reason for this poor performance is that the background signal due to non-specific hybridization is not being subtracted, and this has the largest effect at low concentrations. The signal being estimated at low concentrations with this method is largely dominated by background. The importance of subtracting MM values at low concentrations has also been noted by others (Liu *et al.*, 2002; Hubbell *et al.*, 2002). Based on this trend, it does not appear advisable to use the PM only model for microarray analysis, and the remaining comparisons in this work are based on the PM-MM model of dChip.

The log-log trend of signal versus concentration is approximately linear for MAS5, Resolver and dChip/PM-MM below about 100 pM . However, the slope of these curves is lower than unity in all cases. As shown in Table 1, when all points including and below 64 pM are included in the slope calculation, Resolver with a slope of 0.88 comes closest to unity. When the lowest

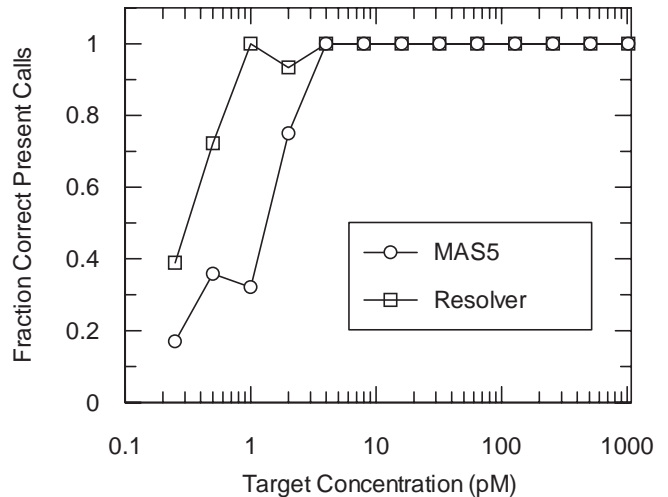
Table 1. Log–log slopes of the signal versus concentration plot

Method	Concentration range (pM)	
	1–64	0.25–64
MAS5	0.84	0.79
dChip/PM-MM	0.87	0.83
Resolver	0.90	0.88
dChip/PM-Only	0.62	0.51

two concentration points are dropped, all the calculated slopes increase, indicating saturation at low concentrations as well. The Resolver slope is least affected by dropping the lowest two concentration points, and this indicates that the Resolver curve is closest to being linear down to the lowest concentration. Thus, one would expect the best sensitivity from the Resolver signal at low concentrations. The imputation procedure in MAS5 and the model-based smoothing and outlier rejection in dChip are probably responsible for their slightly reduced sensitivity. Resolver does not discard or impute any cell intensities which improves sensitivity but results in many more negative signals than the other methods. For example, for this data set, Resolver calculates the signal associated with about 3000 (out of 12 626) probe sets to be negative, dChip/PM-MM only about 60 to 70, and MAS5 does not result in any negative signals. Although Resolver does appear to be the most sensitive at low concentrations, it too shows evidence of saturation which could be due to the spiked transcripts being present in the complex background mixture of RNAs at very low constant levels.

Since the slope of the log–log signal versus concentration curve is less than unity, the true ratio of transcript concentrations will always be higher than the estimate obtained from the ratio of signal values. For example, with a slope of 0.88, a 2-fold change in calculated signal actually implies a true concentration ratio of 2.2, and a 50-fold change in signal implies a true concentration ratio of 85. Outside the linear portion of the curve, the ratio estimate obtained from the calculated signal values is even less representative of the true ratio. One possible way around this problem is to use the curve obtained from the spiked transcripts as a calibration curve and incorporate a calibration procedure into routine microarray analysis. Such a curve could be constructed in entirety for each hybridization, or perhaps, this curve could be constructed periodically and used for a batch of hybridizations. In addition to obtaining more accurate estimates of fold change, such a calibration procedure would also be extremely useful in normalization if performed for every hybridization.

Presence calls made by Resolver and MAS5 at the recommended p -value cutoffs (0.01 and 0.04, respectively) are compared in Figure 2 as a function of spiked target

**Fig. 2.** Effect of actual target concentration on presence calls.

concentration. Among the transcripts that are part of the Latin Square series, the algorithms should ideally call all spiked transcripts as present. The one transcript in each series that is not spiked (zero concentration) should not be called present. The fraction correct present calls plotted in the graph is defined as the ratio of the number of times spiked transcripts at a certain concentration were called present to the total number of present calls that should have been made at that concentration. At a concentration of 4 pM and above, both methods correctly identify all spiked transcripts as present. The performance of the MAS5 presence calls starts to deteriorate below this concentration, and Resolver's presence calls deteriorate below 1 pM . At the lowest concentrations, precisely where the presence calls are most important, neither method performs at a level that justifies their use in data analysis. Comparing the trends in Figures 1 and 2, it is evident that the signal estimates are reliable down to lower concentrations than the presence calls. Given the inaccuracy of the presence calls at low concentrations, it is not advisable to eliminate or filter data based on the presence call. Considering the Latin Square transcript that was not spiked in each experiment, the fraction of false presence calls is 0 for MAS5 and 0.167 for Resolver.

Comparison analysis

Performance of the comparison analysis algorithms in detecting and quantifying 2-fold changes in concentration was assessed by applying the algorithms to data from sequential experiments. In each comparison of sequential experiments, all transcripts in the Latin Square set should be detected as changed. The fold change for 12 spiked transcripts is expected to be 2. Of the remaining two spiked transcripts, the concentration of one changes from

0 to 0.25 pM and the other from 1024 to 0 pM . Since the background RNA mixture was unchanged from one experiment to the next, none of the transcripts other than those in the Latin Square set should be called as changed. The number of transcripts not in the Latin Square set that are determined to have changed provides a measure of the false positive rate.

The accuracy of the three analysis methods in making 2-fold change calls is compared in Figure 3. The recommended p -value cutoffs of each method (MAS5:0.0025, Resolver:0.01, dChip:0.05) were used in this comparison. The accuracy is measured by the fraction of 2-fold changes in the entire Latin Square data set correctly called by each method. For example, there are a total of 12 different instances in the Latin Square data set where the comparison will be done between two experiments in which one of the transcripts was spiked at 0.25 pM in one experiment and 0.5 pM in the other. The y -axis in Figure 3 is the number of times out of 12 that each method detects this change divided by 12. This accuracy measure is plotted against the lower target concentration to assess any systematic concentration effects. The lower target concentration is the lower of the two spiked concentrations being compared (0.25 pM in the above example). Resolver comes closest to the desired behavior of fraction correct calls equal to unity across the entire concentration range. MAS5 performs almost as well, and the performance of dChip is not as good. All three methods have difficulty correctly detecting changes at low and high concentrations. At the high end, this difficulty is most likely due to the saturation of the signal, whereas at the low end, the problem is due to the variability of the signal. The instances where a transcript was not spiked in one of the experiments being compared, and spiked at a 0.25 pM concentration in the other are not represented in Figure 3. Of a total of 12 such changes, MAS5 and dChip correctly identified only 1, and Resolver identified 4.

The methods are more systematically compared in terms of sensitivity and specificity by using five p -value cutoffs for each (0.0025, 0.005, 0.01, 0.05, 0.1) and constructing the Receiver Operating Characteristics (ROC) curves. In this comparison shown in Figure 4, the overall accuracy or sensitivity across all concentrations is plotted on the y -axis. This estimate of accuracy is calculated using all the spiked transcripts with 2-fold concentration change as well as those that change from 0 to 0.25 pM . It does not include the transcript that changes in each comparison from 1024 to 0 pM . The false positive rate shown on the x -axis is assessed from the number of transcripts not included in the Latin Square set that are determined to have changed significantly. Results of comparing sequential experiments using the Student's t -test are also included in the figure for reference. The t -test was performed on the signal values for each probe

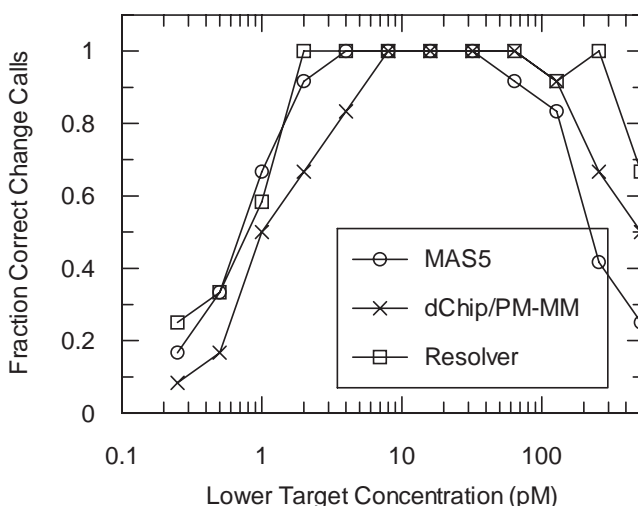


Fig. 3. Effect of lower target concentration on 2-fold change calls.

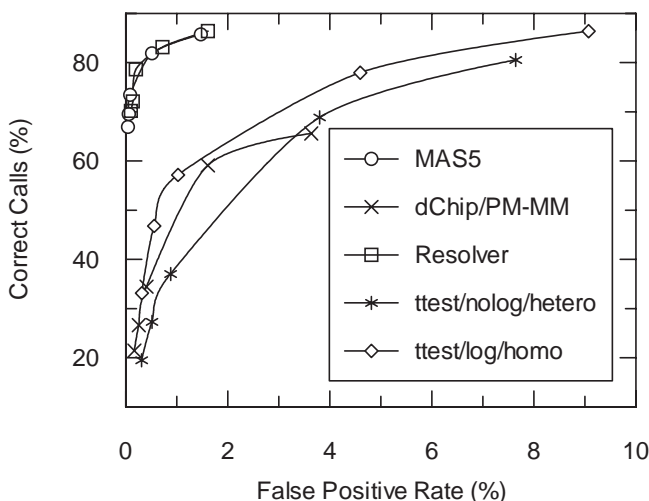


Fig. 4. Comparison of rate of correct 2-fold change calls plotted against false positive rate.

set calculated using the absolute analysis procedure of MAS5 on each array. Two versions of the two-tailed t -test were performed: the heteroscedastic version without log transforming the intensities, and the homoscedastic version after log transformation. It is evident from this comparison that the performance of MAS5 and Resolver is far superior to dChip or t -tests in making change calls. At a fixed false positive rate, both MAS5 and Resolver detect many more true changes. Conversely, to obtain a comparable level of accuracy, dChip would generate many more false positive calls.

Although the Latin Square data set includes replicate hybridizations for each experiment, it can also be used to

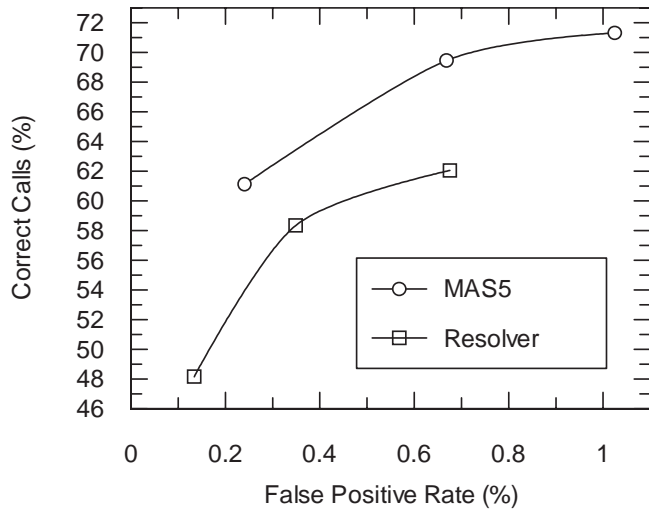


Fig. 5. Comparison of the average rate of correct 2-fold change calls from the nine individual pairwise comparisons between experiments B and A plotted against the mean false positive rate over the nine comparisons.

gauge the performance of the algorithms when replicates are not available. Each of the nine pairwise comparisons between the individual hybridizations from experiments A and B were constructed using Resolver and the number of correct calls and false positives at p -value cutoffs of 0.01, 0.05 and 0.1 were calculated. These results were compared to the corresponding individual comparisons carried out using MAS5 at p -value cutoffs of 0.00025, 0.00125 and 0.0025. The results from the nine comparisons were averaged and plotted in Figure 5. As in the comparison with replicates, both MAS5 and Resolver perform well with a low false positive rate and good accuracy in detecting actual changes in target concentration.

The fold change values calculated by the various methods are plotted versus the lower target concentration in Figure 6. The value plotted is the median value of the fold change restricted to only those transcripts that each method called as changed in concentration. The fold change values are all biased below the correct value of 2, as expected from the slope of the intensity versus concentration curve discussed above. Above a concentration of about 100 pM where the intensity curves saturate, the fold change values consistently decrease away from the correct value of 2.

CONCLUSIONS

A wide variety of statistical methods have been employed to analyze the data generated in microarray experiments using Affymetrix GeneChips[®]. Three different analysis approaches have been compared in this work: non-

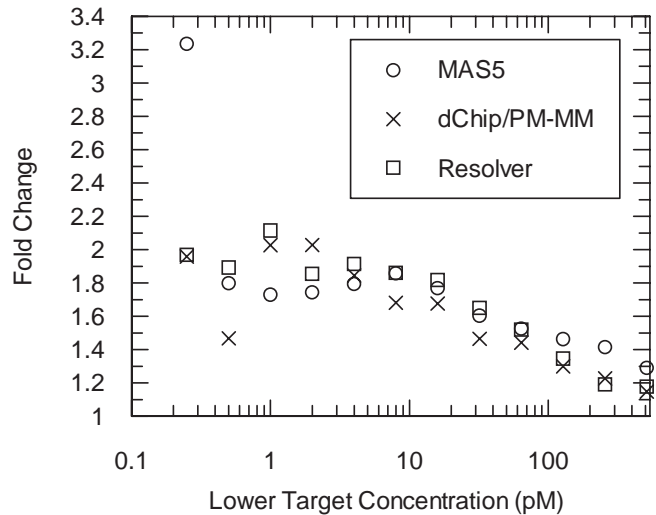


Fig. 6. Effect of lower target concentration on calculated fold change.

parametric statistical methods implemented in Affymetrix Microarray Analysis Suite v5.0; an error-modeling based approach implemented in Rosetta Resolver[®] v3.1; and an intensity-modeling approach implemented in dChip v1.1. These methods are compared in terms of their accuracy in detecting and quantifying relative gene expression (absolute analysis) as well as changes in gene expression (comparison analysis). A Latin Square data set generated and made available by Affymetrix was used in the comparison. This data set consists of 14 labeled transcripts spiked at varying concentrations into a labeled mixture of RNA from a tissue in which the spiked transcripts are known not to be present. The spiked concentrations range from 0.25 to 1024 pM . This data set enables evaluation of both absolute and comparison analysis capabilities of the various methods.

All three methods—Resolver, MAS5 and the version of dChip based on the difference between PM and MM intensities—perform well in absolute analysis. The log-log plot of signal versus actual transcript concentration is linear over almost three decades in concentration, with slopes slightly less than unity. Resolver generates a slope value closest to unity (0.88), and the slope calculated using the other methods are lower. Since the slope of these plots are less than unity, ratios of signal will always be lower than the corresponding ratio of actual transcript concentration. All methods show evidence of saturation above a concentration of about 100 pM . The version of dChip based on the PM intensity alone performs very poorly at low intensities showing a pronounced loss of sensitivity. This version of the intensity model was introduced to avoid negative signals that arise from MM

intensities larger than the corresponding PM intensity. Unfortunately, the accompanying loss in sensitivity is far too severe. Resolver provides the best sensitivity at low concentrations, but it also generates the most number of negative signals. MAS5 generates no negative signals, and dChip PM-MM generates very few, and this is accompanied by only a small loss in sensitivity at low concentrations. Presence calls made by MAS5 and Resolver perform well at high concentrations, but neither can be relied upon at low concentrations where this information could be most useful. Both methods fail to reliably detect transcripts spiked at low concentrations (MAS5 below 4 pM , Resolver below 1 pM).

The performance of Resolver and MAS5 in detecting 2-fold changes in transcript concentration is far superior to that of dChip or t -tests. At a comparable false positive rate, Resolver and MAS5 are able to detect many more true changes in transcript concentration. Each method performs best at detecting concentration changes at intermediate concentrations, and their accuracy falls off at low and high concentrations. The falloff at low concentration is due to increased variability of the signal, whereas at high concentrations, it is due to saturation of the signal. Applying the comparison methods to experiments where there are 2-fold changes in spiked transcript concentrations, the estimated fold changes are always biased below the correct value of 2. The deviation of the estimated fold change from the correct value of 2 increases with concentration at high concentration due to saturation of the signal.

It is important to note that except for the false positive rates, other aspects of the algorithms have been evaluated based on the behavior of only 12 spiked transcripts. However, the false positive rates in detecting differential expression were determined using the more than 12 000 non-spiked transcripts. As more data sets become available in which other transcripts are spiked in, it will be important to verify that the conclusions of this work hold.

ACKNOWLEDGEMENTS

I'd like to thank Lee Weng (Rosetta Inpharmatics, LLC), Robert Hnatuk (Affymetrix Inc.), and Cheng Li (Harvard University) for useful discussions on the different analysis methods. Several colleagues at GlaxoSmithKline contributed to this work. Steve Clark proposed this comparison and suggested using the non-spiked transcripts to gauge false positive rates. Sujoy Ghosh suggested using the Latin Square data set for the comparison. Mike

Lonetto provided software assistance, and Shawn O'Brien loaded the data into the MAS5 and Resolver systems. Bob Gagnon suggested performing the log-transformed, homoscedastic t -test. I'd also like to thank Kay Tatsuoka and Pankaj Agarwal for several useful discussions about this work.

REFERENCES

- Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2002) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, in press.
- Hoffmann, R., Seidl, T. and Dugas, M. (2002) Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol.*, **3**, 0033.1–0033.11.
- Hubbell, E., Liu, W.-M. and Mei, R. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.
- Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2002) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, in press.
- Lemon, W.J., Palatini, J.J.T., Krahe, R. and Wright, F.A. (2002) Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics*, **18**, 1470–1476.
- Li, C. and Wong, W.H. (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
- Li, C. and Wong, W.H. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.*, **2**, 0032.1–0032.11.
- Liu, W.-M., Mei, R., Di, X., Ryder, T.B., Hubbell, E., Dee, S., Webster, T.A., Harrington, C.A., Ho, M.-H., Baid, J. and Smeekens, S.P. (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, **18**, 1593–1599.
- Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
- Roberts, C.J., Nelson, B., Marton, M.J., Stoughton, R., Meyer, M.R., Bennett, H.A., He, Y., Dai, H., Walker, W.L., Hughes, T.R. *et al.* (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, **287**, 873–880.
- Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary cDNA microarray. *Science*, **270**, 467–470.
- Stoughton, R. and Dai, H. (2002) Statistical combining of cell expression profiles. US Patent 6351712.