An Exploration of Extensions to the RMA Algorithm

Kai-Ming Chang[1,*], Chris Harbron[2] and Marie C South[1]

[1]Cancer Bioscience, [2] Statistical Sciences, AstraZeneca, Alderley Park, Macclesfield, Cheshire, SK10 4TG, UK

ABSTRACT

There is frequently a requirement to analyze microarray data at one or more interim stages throughout the course of a study. The RMA algorithm for pre-processing Affymetrix microarray data has an undesirable property that the gene expression intensities on a microarray change when re-pre-processing is necessary due to the inclusion of additional microarrays at different stages of the study. The use of the RMA algorithm can also be limited by available computer memory. We describe and explore the properties of the RMA+ algorithm based on building an RMA model on a reference set of microarrays and storing the parameters of this model fit. The gene expression values for subsequent microarrays are calculated from these parameters without changing the gene expression values of previously calculated microarrays. We also propose an extension of the RMA+ method, RMA++, which makes lower demands on computer memory than RMA. Evaluating these methods on data in Bhattacharjee *et al.* (1) shows that RMA+ and RMA++ give a close approximation to RMA.

INTRODUCTION

cDNA microarrays, which measure gene expression, have been widely used in both academic research and the bio-technical industry for genomic studies. The high-density single-channel oligonucleotide microarray technology provided by Affymetrix® (2) is one of the most commonly applied methods which allows simultaneous measurement of the expression of an extensive number of genes. An Affymetrix GeneChip® microarray consists of many probesets, which are made up of a number of probe pairs comprising a perfect-match (PM) probe and a mismatch (MM) probe. By design, the abundance of a specific mRNA sequence is detected by a PM probe, and the MM probes are intended to measure the background signals. The probeset (gene) expression data are derived from the probe-level data by applying a pre-processing method. A number of pre-processing methods have been proposed, including single-chip methods such as MAS5 (3), and multiple-chip methods such as RMA (4), dChip (5), PDNN (6), GC-RMA (7), and PLIER (8). Among these methods, RMA is a widely accepted standard for pre-processing.

In the RMA method only PM measurements are used. A parametric model is assumed to describe the relationship between the PM measurements and the probeset intensities using the entire set of microarrays to be analyzed. A consequence of this is that the gene expression measurements of a microarray depend on the other microarrays that are pre-processed at the same time. In practice, sample collection for a study can be prolonged and microarray data may be generated at several points throughout the study. It is frequently desired to explore data at one or more interim stages during the study, as well as when the final data set is complete. As a result, microarrays may need to be pre-processed sequentially. One problem with using RMA is that the expression measurements of a chip will change when the current set of microarrays are re-pre-processed with additional microarrays. The gene expression measurements generated with different sets of microarrays are not directly comparable. This creates another problem when applying a predictive or classification model to a set of microarrays that were not pre-processed together with the original microarray set used to develop the model. In order to apply a model to the additional microarrays, the original and additional sets of microarrays have to be pre-processed together, but this changes the gene expressions of the microarrays used to develop the model, so the model is no longer valid. . The methods we describe in this paper provide solutions to these problems.

We give a detailed description of the RMA+ method, as an extension to RMA. This method has also been independently developed by Goldstein (Partition Resampling http://ludwig-sun2.unil.ch/~darlene/pub.html#Sub) as the Extrapolation Strategy. It avoids having to re-pre-process already pre-processed microarrays when new arrays are added to the data set, but still maintains many of the desirable properties of RMA. Under this framework, a reference set of microarrays which reasonably represents the chip population is chosen, and the parameters of the model are estimated using the reference set microarrays. Without having to re-fit the model when the dataset is updated, any future array can be processed independently without changing the gene expression measurements of the other chips.

The use of the RMA algorithm for processing large numbers of chips can be limited by available computer memory. One solution to this issue would be to apply RMA+, using a subset of chips as the reference set and processing the other chips using the parameters calculated from this reference set. We also propose the RMA++ algorithm as an improved approximation to RMA in this situation. RMA++ averages multiple RMA+ results based on several randomly selected reference sets. This approach provides an improved approximation to RMA. RMA++ can also be viewed as an extension of Partition Resampling by Goldstein. Both methods randomly select subsets of chips as reference sets: RMA++ averages over multiple

RMA+ results from both the reference and future chips, whilst Partition Resampling averages just over the set of reference chips.

In this paper, we demonstrate the performance of our methods using results from two studies. In study 1, we compare how well RMA+ and RMA++ approximate RMA. In study 2, we perform a differentially expressed gene analysis (9) to compare the gene lists obtained using the RMA and RMA+ methods. These studies were based on HGU95av2 microarrays of the lung tumor samples in Bhattacharjee *et al*. (1).

METHODS

MODELS

RMA

RMA consists of three steps:

1. ***Background Correction***: probe-level data for each chip are background corrected independently using a probabilistic model;
2. ***Quantile Normalization***: the background corrected probe-level data on each chip are normalized to a common set of quantiles, Q, derived from background corrected data from all chips;
3. ***Expression Calculation***: performed separately for each probeset. The logarithmic intensities for each microarray are estimated from the linear model $\log_2(N_{ij})=P_j+I_i+e_{ij}$ using median-polish, where I_i is the logarithmic intensity for the i^{th} microarray , N_{ij} is the background corrected and normalized intensity of the j^{th} PM probe of the i^{th} chip, P_j is the effect of the j^{th} PM probe of the probeset, and e_{ij} is a random error term.

For further details on the RMA algorithm refer to Irizarry *et al*.(4).

Figure 1 illustrates the relationship between variables involved in the RMA process. The probeset intensities of a chip depend on all chips through the calculation of P(the collection of all probe effects, P_j) and Q (the normalization quantiles). If any of the chips is altered or an additional chip is included, this will affect both P and Q and so all probeset intensities will change. Given P and Q, the calculated probeset intensities of all microarrays are conditionally independent of each other.

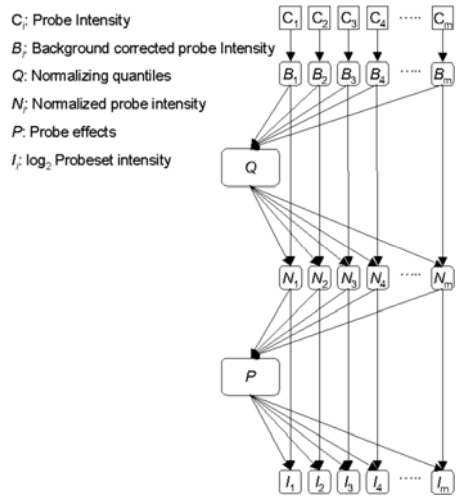

Figure 1. RMA

RMA+

Figure 2 shows a representation of the RMA+ method. The RMA+ method divides the set of microarray chips into two distinct sets: the reference set of chips which are used to generate reference sets of parameters for future processing, and the future set of all other chips which are subsequently processed. The RMA+ method follows 4 steps:

1. ***RMA*:** an RMA model is fitted to the reference set to obtain the estimated normalizing quantiles Q, probe effects P, and the probeset intensities of the reference set microarrays;
2. ***Background Correction***, as in RMA: probe-level data of each chip are background corrected independently using a probabilistic model;
3. ***Normalization***: the background corrected probe level data from the future microarrays are quantile normalized to the reference quantiles, Q;
4. ***Expression Calculation***: the probeset intensities of a future microarray are estimated using the linear model $\log_2(N_{fj})=P_j+I_f+e_{fj}$, assuming that the probe effects of the future arrays are the same as the probe effects of the reference set. The estimated logarithmic intensity $\tilde{I}_f$ of a probeset on a future array is:

$$\tilde{I}_f = \underset{j}{\text{median}}\left(\log N_{fj} - P_j\right) \qquad \textbf{Equation 1}$$

where N_{fj} is the background corrected and normalized intensity of the j^{th} PM probe of the future chip and P_j is the reference probe effect of the j^{th} PM probe of the probeset.

Given P and Q, all future chips are processed independently and only depend on the reference microarrays through the pre-stored P and Q, and it is not necessary to re-pre-process any previously processed chip. As a result, RMA+ allows microarrays to be pre-processed in several batches without changing the intensity measurements calculated in earlier batches.

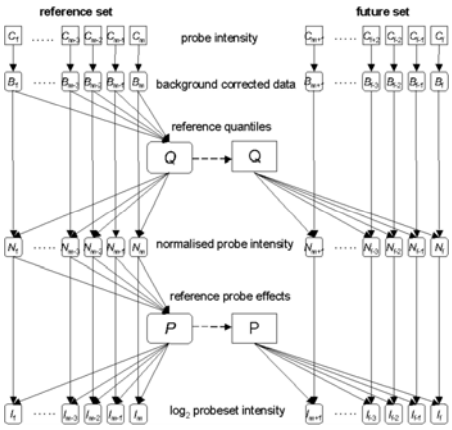

Figure 2. RMA+

RMA++

The use of the RMA algorithm for processing large numbers of chips can be limited by available computer memory, meaning that with limited computing power large sets of microarrays cannot be processed in one batch by RMA. Both the number of probes on chips and the number of chips studied simultaneously are increasing, meaning that this problem may become more acute in the future. One solution to this issue would be to apply RMA+, using a subset of chips as the reference set and processing the other chips using the parameters calculated from this reference set. We also propose the RMA++ algorithm which averages multiple RMA+ results calculated from randomly selected reference sets as an improved approximation to RMA in this situation. RMA++ is based on the idea that probe effects and normalization quantiles derived from multiple random reference sets could be more representative of the entire set of microarrays than a single reference set. RMA++ consists of the following steps:

1. Randomly select n chips as a reference set, the remainder of the chips form the future set. n should be selected to be the maximum number of chips that can be processed in one batch by RMA within the available computer memory;
2. Apply RMA+ to this reference and future set;

3. Repeat steps 1 and 2 several times;
4. Take an average (on the $\log_2$ scale) of the gene expression profiles calculated in steps 1-3;
5. Any additional chips can be pre-processed by using RMA+ to calculate a gene expression profile based on the saved parameters from all of the randomly selected reference sets and averaging these gene expression profiles across the reference sets.

Since the RMA++ algorithm involves multiple RMA processes, it is more computational intensive than RMA+. However implementing a parallel processing scheme would reduce the computing time to the same order as RMA+.

IMPLEMENTATION

All computing procedures were coded in the R language in R v2.1.1 [www.r-project.org (10)], which is a statistical computing environment. The AFFY v1.6.7 (11) and affyPLM v1.3.3 (12) packages in Bioconductor v1.6 [www.bioconductor.org (13)] were used to build the models. The probe intensities in the CEL files were background-corrected and normalized using the AFFY package, which was also used to apply RMA to the entire set of microarrays for each study. For RMA+ and RMA++, the probe effects of the reference set(s), and the probeset intensities were obtained using the rmaPLM function in the affyPLM package.

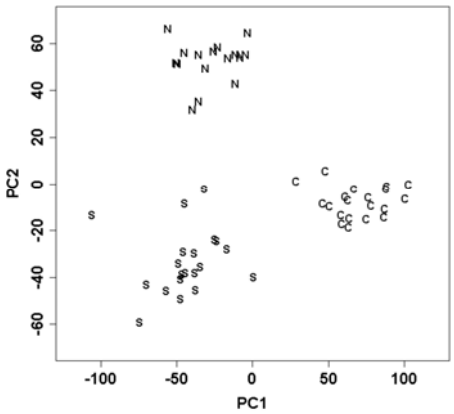

Figure 3. Principal Component Plot of 1st and 2nd Principal Components (35.7% of variation explained) of RMA measurements of 21 squamous cell lung carcinomas (S) , 20 pulmonary carcinoids (C) and 17 normal lung (N) microarrays.

EVALUATION

In this paper, we evaluated the reference set methods in two ways. Firstly, we looked at how well RMA+ and RMA++ approximate RMA over a range of different sizes of randomly selected reference sets using the root mean squared difference (RMSD) in expression values (log2 scale). Secondly, we compared the list of differentially expressed genes (DEGs) identified by the RMA+ method with the DEG list identified by the RMA method under an unfavorable scenario where the reference set is selected from a different population to the microarrays of interest.

In both studies we applied our methods to subsets of the HG-U95aV2 microarrays of lung tumor samples published by Bhattacharjee *et al.* (1). The CEL files of their study are available at http://www.broad.mit.edu/mpr/lung/. In their study, the microarrays are categorized as normal lung cells, adenocarcinoma tumor cells, pulmonary carcinoids, squamous cell lung carcinomas, SCLC, and other adenocarcinomas suspected to be extrapulmonary metastases. Information on the samples used in our two studies is provided in the supplementary information.

For the first study, we selected 57 arbitrary microarrays from the adenocarcinoma tumor group (the first three compressed files on the website). One of the 57 microarrays was then excluded from further use in this study since principal component analysis (14) of the RMA measurements identified it as a clear outlier. The remaining 56 chips were then pre-processed again using RMA. The 56 microarrays used are from 54 patients, with 2 patients having replicated microarrays.

For the second study, we used all of the microarrays from the normal (17 samples) , pulmonary carcinoid (20 samples) and squamous cell lung carcinoma (21 samples) categories. The principal component plot in figure 3 shows three well-separated clusters corresponding to sample type. In this study, the 21 squamous cell lung carcinomas were used as the reference set for RMA+ and the other two groups made up the future set.

STUDY 1: COMPARISON OF THE RMA, RMA+ AND RMA++ MEASURES

STUDY DETAILS

In this study, we investigated how well RMA+ and RMA++ approximated RMA for the 56 lung adenocarcinoma sample microarrays. The following steps were used to generate the RMA+ and RMA++ results:

1. Randomly order the microarrays;
2. Form the initial reference set using the first 8 microarrays. The remaining microarrays form the future set;
3. Perform RMA using the reference set. Store the RMA measurements of the reference set, reference quantiles (*Q*), and the reference probe effects (*P*);
4. Calculate the RMA+ measurements of the future set microarrays using *Q* and *P*. The RMA+ measurements of the reference set microarrays are simply their RMA measurements;
5. Extend the reference set by adding 6 microarrays to the current reference set from the current future set. The remaining future microarrays form the next future set. Repeat 3-5 until the size of the reference set is equal to 56;
6. Repeat 1 to 5 20 times;
7. Average the 20 RMA+ gene expression profiles of the same reference set size to obtain the RMA++ measurements.

To measure the difference in expression levels between RMA and RMA+, we calculated the RMSD of the future set ($RMSD_F$) and of the complete set of all microarrays ($RMSD_A$). The RMSD of the expression measures is defined as

$$RMSD_{\Phi}=\sqrt{\frac{1}{a\times b}\sum_{\Phi}\sum_{\text{all probesets}}\left(\text{RMA expression} - \text{RMA+ expression}\right)^2} \qquad \textbf{Equation 2}$$

where Φ is either the set of future chips, F, or all chips A, *a* is the number of microarrays, and b is the number of probesets. $RMSD_A$ is defined similarly to measure the difference in expression levels between RMA and RMA++.

We also use RMSD to measure the difference between the estimated parameter values (*P* and *Q*) for different methods. In this case RMSD is defined as

$$RMSD=\sqrt{\frac{1}{c}\sum_{\text{all probes}}\left(\text{RMA parameter} - \text{RMA+ parameter}\right)^2} \qquad \textbf{Equation 3}$$

where *c* is the total number of the probes on a microarray.

RESULTS

Figure 4 shows the $RMSD_F$ and $RMSD_A$ for RMA+ and RMA++ for a range of reference set sizes. Figure 5 shows the percentage quantiles of the absolute difference (abs(*D*)) between RMA and RMA+ across the range of reference set sizes, calculated over all 20 different reference sets. The values of RMSD calculated over all microarrays and just for the future set are almost identical. This is unsurprising as the differences in expression levels between RMA and RMA+ are driven by differences in the normalization quantiles *Q* and the probe effects *P*, and these same differences apply to the expression measures calculated for both the reference and the future sets. Figures 4 and 5 both show that, even using a moderately sized reference set, RMA+ gives a good approximation to RMA. As the size of the reference set increases this approximation improves and the variability due to different selections of reference sets decreases. Figure 4 also shows that RMA++ gives an improved

approximation to RMA and that the quality of this approximation improves only marginally as the size of the reference set increases.

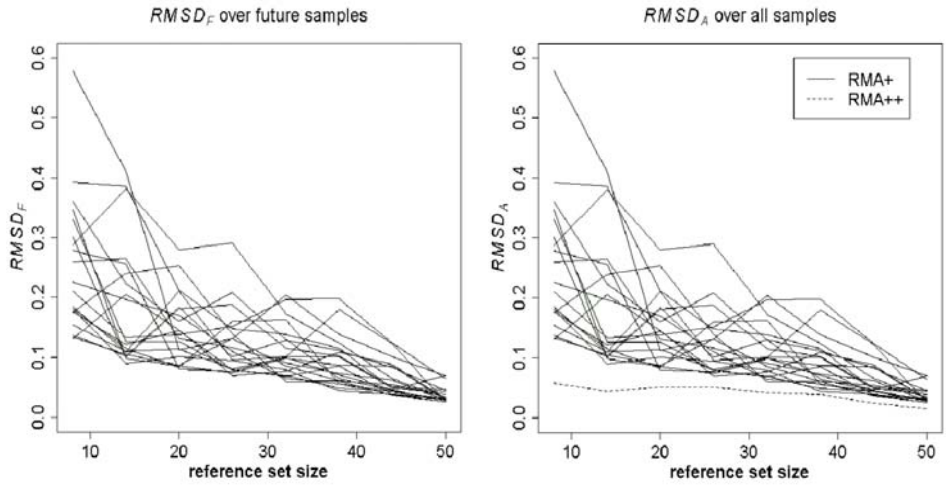

Figure 4 $RMSD_F$ (solid curves in the left plot), $RMSD_A$ (solid curves in the right plot) of 20 runs with different reference sets between RMA+ and RMA measurements, and $RMSD_A$ between the RMA++ and RMA measurements (dashed curves) with different reference set size.

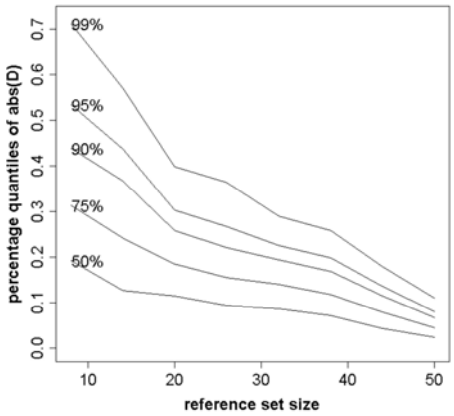

Figure 5. The absolute differences between the RMA and the RMA+ measurements for the given percentage (labelled on the curve) of probesets for each reference set size are not greater than the corresponding value on the curve.

Figures 6 and 7 show the RMSD for the probe effects and the reference quantiles in different runs for different reference set sizes, respectively. Figure 6 shows that the RMSD of the probe effects is fairly independent of the selection of the reference set, but decreases as the reference set size increases. Figure 7 shows a large variation in the RMSD of the reference quantiles for different reference sets, with a smaller relationship with reference set size. Differences in both the reference quantiles and the probe effects appear to be important in how closely RMA+ approximates RMA.

Overall, the results of study 1 have shown that, even with moderate reference set sizes, RMA+ is able to approximate RMA well, and RMA++ yields a further improvement in the approximation. Most of the variation which arises from using different reference sets of the same size is due to variation in the normalization quantiles, rather than the probe effect estimates. The improvement in the probe effect estimates from modeling with a larger number of microarrays contributes most strongly to the improved approximation associated with increased reference set size.

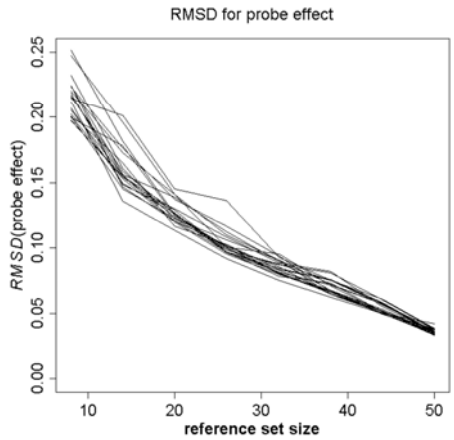

Figure 6 Variation of *RMSD* of the estimated probe effects at different reference set sizes.

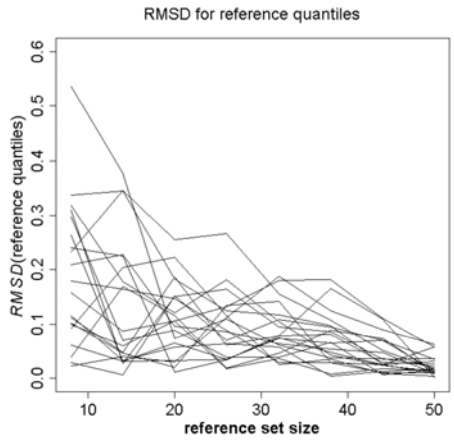

Figure 7. Variation of the *RMSD* of the reference quantiles at different reference set sizes

STUDY 2: DIFFERENTIALLY EXPRESSED GENE ANALYSIS

Study Details

In this study, RMA+ measurements for normal samples and pulmonary carcinoid samples were calculated using squamous cell lung carcinoma samples as the reference set. For comparison, RMA measurements of the normal samples and the pulmonary carcinoid samples were calculated in one batch, firstly excluding the squamous cell lung carcinoma samples (RMA_{future}) and secondly including the squamous cell lung carcinoma samples (RMA_{all}). This is an extreme and challenging scenario for RMA+ as the reference set of squamous cell lung carcinoma samples is very different from both the normal and the pulmonary carcinoid samples, as shown in Figure 3.

The AFFX prefixed probesets were excluded from the data after pre-processing. For both the RMA and RMA+ processed data a t-test assuming unequal variance in both groups was used to compare the normal and pulmonary carcinoid samples for each probeset in turn. The correlation between the RMA t statistics of the probesets and the RMA+ t-statistics of the probesets reflects was used to assess the degree of agreement between these two methods. We examined the degree of overlap between the gene lists generated from the RMA or RMA+ values based on a variety of p-value cut-offs. .

We also used $RMSD_F$ to measure the overall difference between the RMA+ measurements and the RMA_{future} and RMA_{all} measurements.

Results

The $RMSD_F$ between RMA+ and RMA_{all} is 0.2862. This is slightly higher than observed in study 1 for the same reference set size reflecting the greater heterogeneity between samples used in study 2. The $RMSD_F$ between RMA+ and RMA_{future} is 0.6788. This is considerably larger, reflecting that the model parameters for RMA+ and RMA_{future} are calculated from non-overlapping sets of chips and that this is designed to be an extreme situation.

Figure 8 shows the t-statistics based on RMA_{future} and RMA+ which show a high degree of agreement (Pearson correlation coefficient R=0.99). Table 1 shows the number of probesets selected given different *p*-value cut-offs. Using a *p*-value cut-off of 10^{-20}, 10 probesets would be selected based on RMA_{future} and 12 probesets would be selected based on RMA+ with 9 probesets selected regardless of the pre-processing method. Using a *p*-value cut-off of 10^{-12}, 136 probesets would be selected based on RMA_{future} and 141 probesets would be selected based on RMA+ with 125 probesets selected regardless of the pre-processing method. These results show a large degree of overlap between the probesets selected by RMA_{future} and RMA+.

We conclude that the results of DEG analysis of RMA_{future} and those of RMA+ are statistically very similar even though the reference set of microarrays is very different from the set of microarrays of interest.

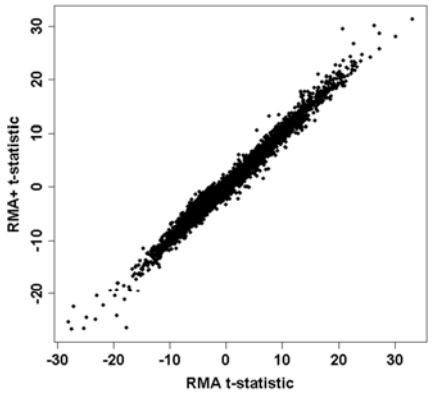

Figure 8. t-statistics of RMA and RMA+.

	Probesets Selected Using RMA+		
$\log_{10}$ p-value cut-off	Probesets Selected Using RMA_{future}		
$\log_{10}(P)$	$p_{RMA}<P$	$(p_{RMA}<P)$ and $(p_{RMA+}<P)$	$p_{rma+}<P$
-4	552	2185	152
-8	98	480	43
-12	11	125	16
-16	3	37	4
-20	1	9	3
-24	2	1	2

Table 1: Number of probesets selected for a given p-value cut-off P, with different gene expression measures, where p_{RMA} and p_{RMA+} represent the p-values based on RMAfuture and RMA+, respectively.

DISCUSSION

In this paper we have described two extensions to the RMA algorithm, RMA+ and RMA++. RMA+ avoids having to re-pre-process data for existing microarrays when additional microarrays are included in the study. RMA+ and RMA++ also permit any number of chips to be pre-processed with no restrictions due to computer memory. The RMA+ algorithm is fundamentally the same as for RMA, except that in RMA+, the model parameters are estimated by applying RMA to a reference set of microarrays as opposed to all microarrays. This allows interim analyses of microarrays to be performed before all data collection is complete and the gene intensities of already processed microarrays will not change when additional microarrays are included. RMA++ is an extension of RMA+ where the RMA++ gene expression profile of a microarray is the average of

multiple RMA+ gene expression profiles of the same microarray each derived from a different randomly sampled reference set.

The results in study 1 show that RMA+ provides a good approximation to the RMA algorithm. As the number of microarrays included in the reference set increases, the closer the RMA+ measure is to the RMA measure. RMA++ approximates RMA even more closely than RMA+ even for small reference set sizes. These results are as expected, as more information about the entire set of microarrays is contained within the reference set, the RMA+ model will approximate the RMA model better. With RMA++, a larger proportion of microarrays are used to build the model regardless of reference set size, so it will approximate RMA even more closely.

The DEG analysis in study 2 showed that the conclusions from analyses where the microarray data have been pre-processed using RMA and RMA+ are very similar, even when the reference microarrays are from a sub-population that is very different from the two groups being compared in the DEG analysis. This result increases our confidence to use the RMA+ gene expression profiles for additional microarrays when comparing the additional microarrays with the microarrays processed with RMA at earlier stage.

The approach of RMA+ is also very compatible with the modelling and subsequent application of predictive or classification models, where a group of samples with known phenotype or response are used to develop a mathematical model which is then used to predict other samples with unknown phenotype or response. The unknown samples predicted by the model may be generated after the model has been developed. Using RMA+ the samples used to develop the model may be pre-processed as the reference set and any unknown future samples can be pre-processed using the calculated parameters, leaving the intensities of the reference set of samples and the generated model and its properties unchanged. If RMA were to be used then either the unknown samples would have to be pre-processed separately in which case the model may not perform well as the parameters from which these samples intensities were derived would be different leading to systematic differences, or the new samples would be pre-processed together with the samples used to generate the model which would change the intensities of those samples and any model generated from them and its properties.

As microarray technology continues to improve, the volume of microarray data, both in terms of the number of probes on a chip and the number of chips that will be processed in a single study, and the computational requirements to process this data will continue to increase dramatically. RMA+ and RMA++ provide an approach for overcoming any computer memory limitations whilst maintaining confidence in the results of subsequent analyses. RMA++ has the additional benefit that it may be parallelized, allowing efficient computation in a parallel computing environment.

Although the methods we developed are based on RMA, the same concepts can also be applied to other related approaches to pre-processing Affymetrix chips such as GCRMA.

R code for RMA+ and RMA++ are available from the first named author.

SUPPLEMENTARY DATA

Supplementary files are available from the first author.

ACKNOWLEDGEMENTS

The authors would like to acknowledge colleagues within AstraZeneca who provided valuable suggestions and comments, and thank the authors of Bhattacharjee, et al. (2001) who permitted the use of their microarray data.

REFERENCES

1. Bhattacharjee,A., Richards,W.G., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.
2. 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. *Nature Biotechnology*, **14**, 1675–1680.
3. Hubbel,E., Liu W.M. and Mei R. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585-1592.
4. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, Normalization, and Summaries of High density Oligonucleotide Array Probe Level Data. *Biostatistics*, **4**, 249-264.
5. Li,C. and Wong,W.H. (2001) Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection. *Proc. Natl Acad. Sci. USA*, **98**, 31-36.

6. Zhang,L., Miles,M.F. and Aldape,K.D. (2003) A Model of Molecular Interactions on Short Oligonucleotide Microarrays. *Nature Biotechnology*, **21**, 818–821.
7. Wu,Z., Irizarry,R.A., Gentleman,R., Martinez-Murillo,F. and Spencer,F. (2004) A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, **99**, 909-917.
8. (2004) GeneChip® Expression Platform: Comparison, Evolution, and Performance. *Technical Note*, Santa Clara, USA
9. Lönnstedt,I. and Speed,T. (2002) Replicated Microarray Data. *Statistica Sinica*, **12**, 31-46.
10. Ihaka,R. and Gentleman,R. (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**, 299-314.
11. Gautier, L., Cope, L.M., Bolstad,B.M. and Irizarry,R.A. (2004) affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307-315.
12. Bolstad,B.M. (2004) Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. *Dissertation*. University of California, Berkeley.
13. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B.M., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.*, (2004) Bioconductor: Open Software Development for Computational Biology and Bioinformatics. *Genome Biology*, **5**, R80.
14. Venables,V.N. and Ripley,B.D. (2002) *Modern Applied Statistics with S*, 4th ed., Springer, London.