Published online 1 May 2008

Nucleic Acids Research, 2008, Vol. 36, No. 10 ***e58***
doi:10.1093/nar/gkn234

Enhanced identification and biological validation of differential gene expression via Illumina whole-genome expression arrays through the use of the model-based background correction methodology

Liang-Hao Ding[1,2], Yang Xie[3,4], Seongmi Park[2], Guanghua Xiao[3] and Michael D. Story[1,2,*]

[1]Simmons Comprehensive Cancer Center Genomics Core Facility, [2]Department of Radiation Oncology, Division of Molecular Radiation Biology, [3]Simmons Comprehensive Cancer Center Biostatistics Core and [4]Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center at Dallas, Dallas, TX, USA

Received November 11, 2007; Revised April 11, 2008; Accepted April 14, 2008

ABSTRACT

Despite the tremendous growth of microarray usage in scientific studies, there is a lack of standards for background correction methodologies, especially in single-color microarray platforms. Traditional background subtraction methods often generate negative signals and thus cause large amounts of data loss. Hence, some researchers prefer to avoid background corrections, which typically result in the underestimation of differential expression. Here, by utilizing nonspecific negative control features integrated into Illumina whole genome expression arrays, we have developed a method of model-based background correction for BeadArrays (MBCB). We compared the MBCB with a method adapted from the Affymetrix robust multi-array analysis algorithm and with no background subtraction, using a mouse acute myeloid leukemia (AML) dataset. We demonstrated that differential expression ratios obtained by using the MBCB had the best correlation with quantitative RT–PCR. MBCB also achieved better sensitivity in detecting differentially expressed genes with biological significance. For example, we demonstrated that the differential regulation of *Tnfr2*, *Ikk* and *NF-kappaB*, the death receptor pathway, in the AML samples, could only be detected by using data after MBCB implementation. We conclude that MBCB is a robust background correction method that will lead to more precise determination of gene expression and better biological interpretation of Illumina BeadArray data.

INTRODUCTION

With the advent of microarray technology, gene expression analysis has become a valuable tool in biological research from development to cancer. Researchers now can choose from a number of commercially available microarray platforms. While early results were limited by lack of reproducibility, the recently completed Microarray Quality Control (MAQC) project demonstrated that through the use of standardized protocols both intraplatform consistency and interplatform concordance could be achieved (1–6). However, one area of data processing where standardization and optimization lags is background signal correction, that is, the removal of nonspecific signals from total signal intensity in the process of microarray data analysis. Comprehensive comparisons have been conducted to evaluate the performance of different background correction methodologies in two-color arrays (7); however, because of the high extent of commercialization, background correction methods for single-color arrays are mostly platform dependent. This is a consequence of different array designs, image scanning and data extraction processes developed by different vendors. Even in this setting, there is a lack of standardization for background correction—even within the same platform. For example, the robust multi-array analysis (RMA), a popular algorithm to preprocess Affymetrix GeneChip data (8), uses only perfect match (PM) probes and ignores mismatch

*To whom correspondence should be addressed. Tel: +1 214 648 5557; Fax: +1 214 648 5995; Email: michael.story@utsouthwestern.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

(MM) probes when correcting for background signal. Although the empirical experience shows that RMA background correction works well in practice (8,9), it uses the *ad hoc* parameter estimation procedure and McGee *et al*. (8) have identified problems associated with its parameter estimation.

Another single color microarray platform tested in the MAQC project is from Illumina Inc., San Diego, CA, USA. This platform utilizes the company's BeadArray technology; 3-μm beads coated with hundreds of thousands of copies of 50-mer oligonucleotides sequences. The beads are randomly assembled into microwells on each array. A postscanning 'decoding' process is required to determine the location of each probe (10,11). One feature of Illumina arrays is that its noise is controlled by beads conjugated with nonspecific oligonucleotides. Illumina Inc. provides a method to perform background subtraction using the average value of control beads. Unfortunately, substantial negative data values can be generated by this method. For example, with the samples used in this study, where samples from one group are compared against another, use of the Illumina default background correction resulted in more than half of the probe values in one group being negative. More importantly, over 6000 probes had negative values in one group but positive values in the other group. Exclusion of probes with negative values can result in loss of large amounts of information on the chip, especially genes that 'switch' on and off between different sample groups. In this context, there has been report suggesting that method provided by Illumina has a negative impact on Illumina data quality and the use of the raw data before normalization was recommended (12). However, significant data compression was observed when expression ratios were calculated between two experimental groups when no background subtraction was performed. These data compression resulted in far fewer genes identified as differentially expressed than expected. For these reasons, our intent in this study was to address the necessity of performing background correction and, in the mean time, develop a robust background correction model to facilitate further statistical and data mining analysis.

In addition to using no background correction or the default manufacturer's protocol, there are two background subtraction methods for BeadArrays proposed in the Bioconductor 'lumi' package. The first method is to raise all data uniformly to a 'floor' value, which assures that all signals are positive. The other approach applies the Affymetrix RMA background correction model to BeadArray data. The former method is arbitrary and induces data compression that results in diminished expression ratios. The latter method ignores the nonspecific oligo-beads contained on Illumina arrays. By examining the histogram of signal intensities on BeadArrays, we noted that gene and control signal intensity values exhibited different distributions. The intensity values from genes did not follow the same distribution as that of the control beads in that they exhibited much heavier tails than the control intensity values (Figure 1). It was our contention that the intensity values associated with the nonspecific oligo-labeled beads is valuable information and by using the information from the nonspecific oligo-labeled beads, we could develop a model to estimate a true background and perform first-round data processing for Illumina arrays. This background correction method, referred to as model-based background correction for BeadArrays (MBCB), incorporates the negative control bead information into a statistical algorithm for background correction of Illumina arrays. Using the same set of array data, we compared the results after implementing the MBCB method versus using the data without background subtraction (RAW) as recommended by Barnes *et al*. (12). We also adapted the Affymetrix RMA algorithm and compared it with MBCB as well.

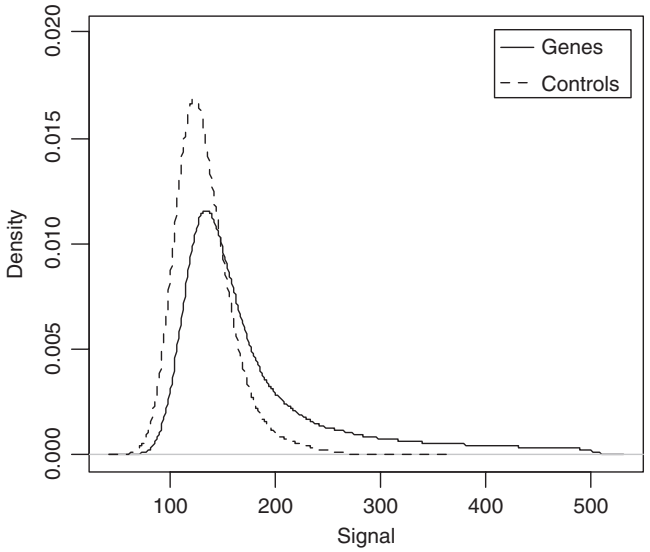


Figure 1. Distribution of control signal and total gene-expression signal. The smoothed histograms represent the observed intensities for both genes and negative controls from Illumina whole-genome expression arrays.

Although we could not apply it to Illumina data, another approach we considered was the Li and Wong model (13,14) found in the popular Affymetrix expression analysis package dCHIP. Their model-based analysis algorithm has a similar philosophical approach to MBCB in that it is a model-based approach to oligonucleotide expression array analysis. Unfortunately, it is not directly applicable to Illumina arrays because of inherent platform differences. For instance, unlike the Affymetrix MM probe that has single nucleotide differences to perfect match probes, Illumina control beads are conjugated to nonspecific sequences that are not associated with gene-specific probes. Furthermore, the main feature of Li–Wong's model is to take account of the probe-specific effects into account for the computation of expression index. They use a parameter (Φ_j) to represent the sensitivity of PM probe of probe pair j, and the parameter is estimated by using information from multiple arrays, a minimum of 10 being optimal. This is appropriate for Affymetrix array because each probe in the probe set is different, and some probes are more sensitive for hybridization. However, for Illumina BeadArray data, the beads in each bead type are identical and there is no sensitivity difference among beads within same bead type. Therefore, the parameter (Φ_j) in

Li–Wong's model is not appropriate for Illumina BeadArrays and dCHIP software could not be adapted to analyze output data from the Illumina platform.

For our approach, samples from spleens of CBA mice, positive or not for acute myelogenous leukemia (AML), were used to generate comparative data sets. The MBCB, RMA and RAW methods were used to process probe signal for subsequent statistical analysis. Relative gene expression values were validated using quantitative RT–PCR (QPCR). In some cases, western analysis of protein expression was also examined. Data preprocessed using the MBCB method had the highest correlation with QPCR results and provided a greater sensitivity for detecting differentially expressed genes. The result was a better discrimination of gene expression between the AML and non-AML groups, and subsequent to that a better interpretation of differences in signal transduction pathway activation in AML-positive spleen samples.

MATERIALS AND METHODS

RNA isolation

In this study, total RNA from samples of spleen cells from four normal CBA mice and samples of spleen cells from four mice confirms by a pathologist as positive for AML were used. The tissues were homogenized in Trizol Reagent (Invitrogen, CA, USA) using an Omni Tip Disposable Generator Probe (Omni International Inc., Marietta, GA, USA). RNA isolation was performed according to the Qiagen RNeasy Mini Kit column (Qiagen, CA, USA) protocol. Chloroform was added to the homogenate for phase separation, the aqueous phase was removed and mixed with 1 Vol. 70% ethanol and the mixture was loaded into the Qiagen column, which was then centrifuged at 11 000 r.p.m. for 1 min. The flow-through liquid was discarded, buffer RW1 was added, the column was washed and the RNA eluted. The RNA was quantified and the quality was checked by using an Experion automated electrophoresis system and Experion RNA StdSens Chips (Bio-Rad Inc., CA, USA).

RNA labeling and microarray hybridization

Illumina Mouse-6 V1 BeadChip (Illumina, Inc.) mouse whole-genome expression arrays were used in this study. Of the eight samples, four were hybridized twice and were used as technical replicates. Each RNA sample was amplified using the Ambion Illumina RNA amplification kit with biotin UTP (Enzo) labeling. The Ambion Illumina RNA amplification kit uses T7 oligo(dT) primer to generate single stranded cDNA followed by a second strand synthesis to generate double-stranded cDNA, which is then column purified. *In vitro* transcription was done to synthesize biotin-labeled cRNA using T7 RNA polymerase. The cRNA was then column purified. The cRNA was then checked for size and yield using the Bio-Rad Experion system. A total of 1.5 μg of cRNA was hybridized for each array using standard Illumina protocols with streptavidin-Cy3 (Amersham, Piscataway, NJ, USA) being used for detection. Slides were scanned on an Illumina Beadstation and analyzed using BeadStudio (Illumina, Inc).

Model-based background correction for BeadArrays (MBCB)

The expression data described by Figure 1 motivated a background plus signal model to explain the observed intensity of gene i, where $S_i = X_i + Y_i$, and S_i represents the observed intensity for gene i, X_i is the signal intensity for gene i, and Y_i is the noise intensity for gene i. The goal was to adjust the observed intensity S_i by removing the effects of background Y_i. The observed intensity of the negative control bead comes only from the noise intensity. In order to estimate X_i, we assume the signal X_i comes from an exponential distribution with mean α and the noise intensity for both genes and negative controls comes from a normal distribution with mean μ and variance σ^2. We applied the model to the observed intensity of genes and negative controls, and used Markov chain Monte Carlo simulations to estimate the parameters. The estimated signal intensity X_i is used as the background corrected expression for gene i. The negative control values were extracted from the bead-level intensity data of all nonspecific control beads and summarized using median values of each control bead type. The summarized control bead-type values for each array were used for the modeling. The R script and data files used for MBCB model was posted as Supplementary materials at the journal website. The model details could be found in the Supplementary material.

Adaptation of RMA to BeadArrays

We adopted the convolution model in RMA for Affymetrix Arrays to Illumina BeadArrays. The parameter estimation also follows RMA's procedure: first, a nonparametric density function was fitted to the observed intensities, the mode of this density was used as the estimate of the mean of the noise; second, the lower tail of the mode was used to estimate the variability of noise and finally, the right tail of the mode was used to estimate the rate of exponential distribution of the signal. The model was applied to the observed intensity of each gene. The expected true signal value condition on the observed total signal is used as the background corrected gene expression level.

Normalization, clustering and significance analysis

After performing MBCB and RMA background subtraction, we applied quantile normalization across samples using the 'Affy' package in Bioconductor. For comparison of groups without background subtraction (RAW), we carried out quantile normalization using the raw intensity data on the arrays. The scatter plots and Venn diagram were generated using GeneSpring GX 7.3.1. Significance analysis was done using significant analysis of microarray (SAM) and BRB-ArrayTool. Pathway analysis was performed using ingenuity pathway analysis (IPA) (www.ingenuity.com).

Real-time quantitative RT-PCR

QPCR was performed on one splenic sample of AML-positive mouse and one normal mouse splenic sample. Fourteen genes were randomly selected to test for

validation of microarray results. RNA solutions were treated with DNase I before reverse transcription. Complementary DNA was synthesized from the treated RNA solution in a reaction containing SuperScript III reverse transcriptase (Invitrogen) and random hexamer primers. The gene-specific primers were designed by using Primer3 software. PCR reactions were performed using a SYBR PCR master kit (AB Biosystems, Inc., Foster City, CA, USA), and a Chromo4 Fluorescence Detector (Bio-Rad, Inc.). The PCR protocol was designed with an initial denaturing step of 95°C 10 min, followed by 40 cycles of 95°C 15 s and 60°C 1 min. Mouse 18S RNA was used as an internal control between samples. The PCR reactions were performed in triplicates for each gene being validated. Primers used in this experiment can be found in the Supplementary materials.

Western blots

Samples of frozen mouse spleen tissues were lyzed with lysis buffer (150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS and 50 mM Tris pH = 8.0), homogenized and then sonicated. After centrifugation for 10 min at 4°C, the protein concentration was measured. For western blot analysis, 25 μg of total protein was resolved on a 7.5% polyacrylamide gel and transferred onto immunobilon-FL membrane (Millipore, Billerica, MA, USA). Western blotting was performed as follows. Briefly, after 1-h incubation with blocking buffer at room temperature, membranes were incubated with the primary antibody overnight at 4°C. The membranes were then washed five times with washing buffer and incubated with the secondary antibody for 1 h at room temperature. After washing with washing buffer three times, the membrane was incubated with ECL plus reagents and exposed to X-film. The antibodies for *Tnfr2* and *NF-kappaB* were purchased from Santa Cruz Biotechnology (Santa Cruz, CA, USA), and the actin antibody was purchased from Sigma, St. Louis, MO, USA.

RESULTS

Model-based background subtraction (MBCB) and quantile normalization

We compared the normalization results of median polish and quantile normalization, in order to determine which method to use in combination with our background correction models. The difference between the two normalization methods was best manifested in two technical replicate samples. These replicates are hybridizations that used the same amplified cRNA samples. The comparison indicated that the MBCB method, in combination with quantile normalization, demonstrated a better correction of variations between chips than the commonly used median polish normalization (Figure 2). In subsequent analyses, quantile normalization after implementation of MBCB was always performed. We also performed quantile normalization when using data without background subtraction (RAW) and with the RMA method.

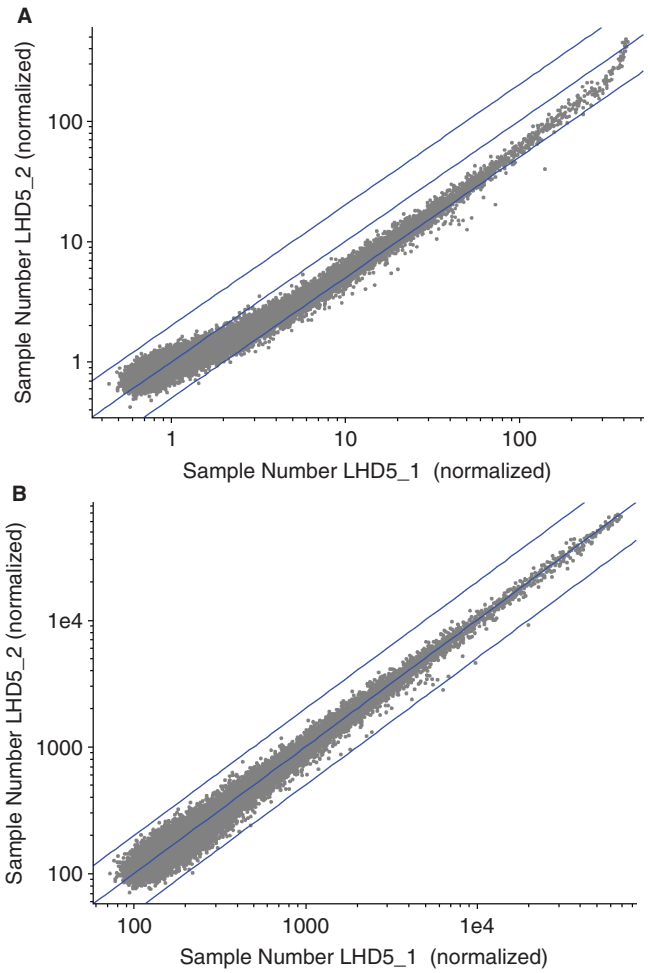


Figure 2. Scatterplot of signal intensity for the same cRNA sample (LHD5) hybridized on two arrays. (**A**) Two arrays normalized to the median of each array after MBCB. (**B**) Quantile normalization was implemented after MBCB.

The MBCB method identifies more genes as differentially expressed

Two different statistical models, widely used to identify differentially expressed genes, were used to interrogate normal spleen samples versus the AML samples. In the first analysis the random variance *F*-test, developed by the NCBI and implemented in BRB-ArrayTools, was used. The default significance ($P < 0.001$) provided by the software was used. In the second analysis, the SAM model, a widely used algorithm developed by groups in Stanford University, was used. The cutoff value for SAM is median false discovery rate (FDR) <0.01. In both of the analyses, more genes were identified as differentially regulated when the MBCB versus either the RMA or RAW methodologies was used (Table 1). When compared to the RAW method, there were 24 or 42% more genes identified when the MBCB background subtraction model was used in combination with either SAM or the BRB ArrayTools, respectively. The RMA method also detected 20 or 42% more genes than the RAW methodology. Plot of FDRs and numbers of significant genes generated from SAM output indicated that the MBCB method facilitated the lowest

Table 1. Number of differentially expressed genes and the percentage increase in detection over the RAW methodology in parentheses

Methods	RAW + Quantile	RMA + Quantile	MBCB + Quantile
BRB ArrayTool	2393	3391 (42%)	3398 (42%)
SAM	3648	4384 (20%)	4521 (24%)

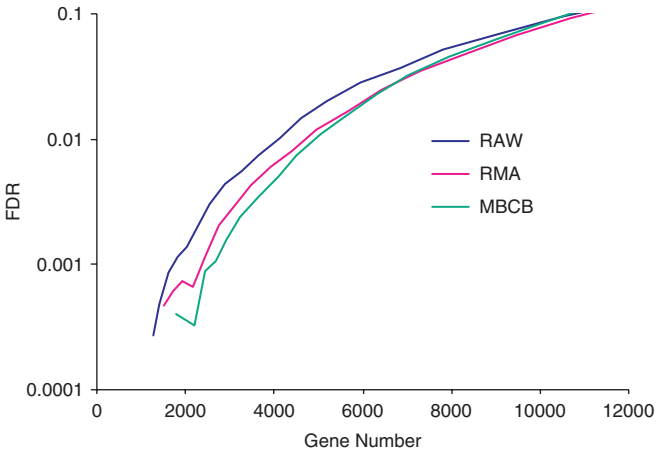


Figure 3. Plot of FDR and numbers of significant genes called by SAM software. The data indicate that the MBCB method resulted in the lowest FDR for a given number of significant genes.

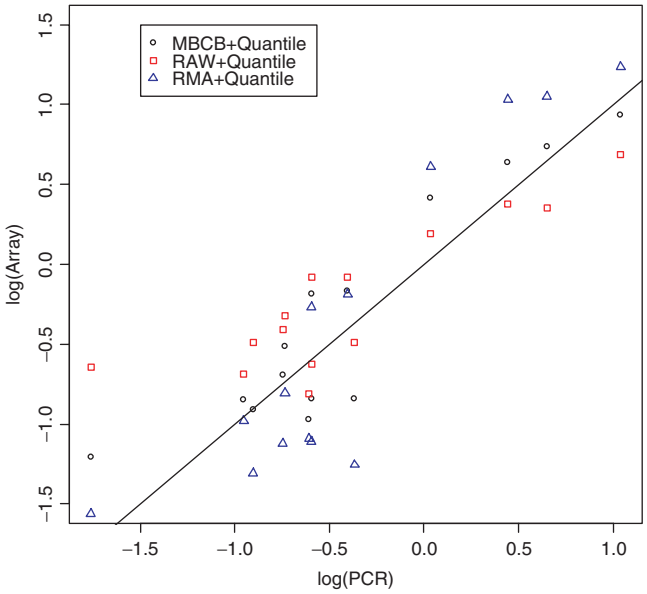
FDR for a given number of significant genes when comparing all three background correction methods (Figure 3).

MBCB-derived data correlates better with quantitative RT–PCR results

QPCR for 14 randomly picked genes was performed in order to validate the Illumina array results. Our comparisons indicate that the MBCB method has the best overall correlation with quantitative PCR results than RMA and RAW method (Figure 4). In the plot of Figure 4, each gene was plotted according to the log-expression ratio generated by RT–PCR and one of the three background correction methods. The MBCB has the smallest mean square error (MBCB: 0.09; RMA: 0.19; and RAW: 0.17). Assuming a hypothetical perfect fit for array data versus PCR data for the same genes, the slope of the fitted line would be 1.0, with a *y*-intercept equal to 0. The fit of the MBCB/PCR data are the closest to meeting that hypothetical condition, see Figure 4. Both the RAW and RMA fits have slopes that significantly deviate from 1.0 and both have *y*-intercepts that deviate from 0. The values for each slope can be found in Figure 4.

AML-related gene-expression changes detected only by the MBCB method

SAM analysis in combination with the MBCB model identified 873 more genes as differentially regulated between the two groups than the combination of RAW and SAM (FDR < 0.01), and 44 of these 873 genes have an



	MSE	Beta (95% CI)	R^2
RAW	**0.17**	**0.56(0.38–0.74)**	**0.74**
RMA	**0.19**	**1.24(0.9,1.58)**	**0.81**
MBCB	**0.09**	**0.91(0.69,1.13)**	**0.83**

Figure 4. Ratios of differential gene expression between AML and normal tissue generated by QPCR show better correlation with array results in MBCB data set than in the RAW data set. MSE is mean square of error when comparing with RT–PCR results; Beta is the slope of fitted linear line and R^2-values indicate how well the linear regression of array data approximates the QPCR results, with a value of 1.0 indicating the perfect fit.

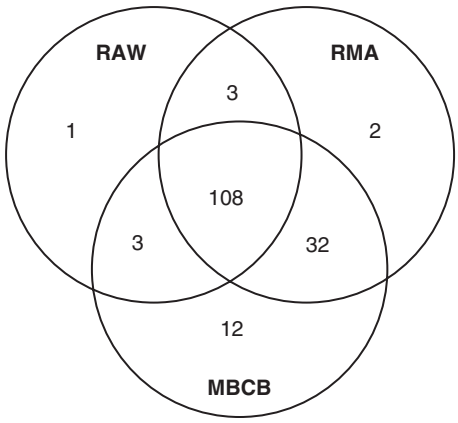


Figure 5. Venn diagram shows the numbers of known leukemia-associated genes that had been detected to be significantly changed in AML samples by SAM (FDR < 0.01) in the data set using different background correction methods.

association with AML according to the designations identified by the IPA software package. If RMA was used in combination with SAM analysis 32 of these 44 genes were identified (Figure 5). In comparison, there were only small portions of genes that could be detected by either

Table 2. AML-related genes, and their expression ratios, identified through the use of the MBCB methodology and not seen when no background correction was performed

Gene	Ratio (AML versus Normal)	Description
Gfi1	14.24	*Mus musculus* growth factor independent 1 (Gfi1).
Ccl3	9.97	*Mus musculus* chemokine (C-C motif) ligand 3 (Ccl3).
Nfil3	7.54	*Mus musculus* nuclear factor, interleukin 3, regulated (Nfil3).
Cdkn2a	5.63	*Mus musculus* cyclin-dependent kinase inhibitor 2A (Cdkn2a).
Scgf	5.34	*Mus musculus* stem cell growth factor (Scgf).
Aml1	4.94	*Mus musculus* runt-related transcription factor 1 (Runx1).
Cdkn1a	4.23	*Mus musculus* cyclin-dependent kinase inhibitor 1A (P21) (Cdkn1a).
Cebpb	3.97	*Mus musculus* CCAAT/enhancer binding protein (C/EBP), beta (Cebpb).
Myb	3.19	Myeloblastosis oncogene
Etv6	2.67	*Mus musculus* ets variant gene 6 (TEL oncogene) (Etv6).
PTPe	2.57	Mouse mRNA for protein tyrosine phosphatase epsilon, complete cds.
Nfkbia	2.35	*Mus musculus* nuclear factor of kappa light chain gene enhancer in B-cells inhibitor, alpha (Nfkbia).
Pscdbp	2.17	*Mus musculus* pleckstrin homology, Sec7 and coiled-coil domains, binding protein (Pscdbp).
Apaf1	2.06	*Mus musculus* apoptotic protease activating factor 1 (Apaf1).
Fli1	2.05	*Mus musculus* Friend leukemia integration 1 (Fli1).
Atm	2.01	*Mus musculus* ataxia telangiectasia mutated homolog (human) (Atm).
Rps6ka1	1.98	*Mus musculus* ribosomal protein S6 kinase polypeptide 1 (Rps6ka1).
Chic2	1.92	Cysteine-rich hydrophobic domain 2
Chk	1.88	Choline kinase
Ripk1	1.78	*Mus musculus* receptor (TNFRSF)-interacting serine-threonine kinase 1 (Ripk1).
Il15ra	1.57	*Mus musculus* interleukin 15 receptor alpha chain isoform 2D mRNA, complete cds, alternatively spliced.
Ncor1	1.49	*Mus musculus* nuclear receptor co-repressor 1 (Ncor1).
Nqo1	0.69	*Mus musculus* NAD(P)H dehydrogenase, quinone 1 (Nqo1).
Gzmb	0.64	*Mus musculus* granzyme B (Gzmb).
Cd68	0.60	*Mus musculus* CD68 antigen (Cd68).
Itga4	0.59	*Mus musculus* integrin alpha 4 (Itga4).
Lasp1	0.58	*Mus musculus* LIM and SH3 protein 1 (Lasp1).
Itgae	0.57	*Mus musculus* integrin, alpha E, epithelial-associated (Itgae).
Csf3	0.57	*Mus musculus* colony stimulating factor 3 (granulocyte) (Csf3).
Acp1	0.56	Acid phosphatase 1, soluble
Fyn	0.56	*Mus musculus* Fyn proto-oncogene (Fyn).
Hdac11	0.55	*Mus musculus* histone deacetylase 11 (Hdac11).
Igf1	0.54	*Mus musculus* insulin-like growth factor 1 (Igf1).
Tnfrsf25	0.53	*Mus musculus* tumor necrosis factor receptor superfamily, member 25 (Tnfrsf25).
Top2a	0.52	Topoisomerase (DNA) II alpha
Brca1	0.50	*Mus musculus* breast cancer 1 (Brca1).
Acsl6	0.49	*Mus musculus* acyl-CoA synthetase long-chain family member 6 (Acsl6).
Ssh1	0.48	Slingshot homolog 1 (Drosophila)
Adam10	0.48	*Mus musculus* a disintegrin and metalloprotease domain 10 (Adam10).
Igfbp3	0.47	*Mus musculus* insulin-like growth factor binding protein 3 (Igfbp3).
Ifnar2	0.46	*Mus musculus* interferon (alpha and beta) receptor 2 (Ifnar2).
Cat	0.46	*Mus musculus* catalase (Cat).
Ccnd1	0.45	*Mus musculus* cyclin D1 (Ccnd1).
Casp3	0.36	*Mus musculus* caspase 3, apoptosis-related cysteine protease (Casp3).

RAW (four genes) or RMA (five genes) but not by MBCB (Figure 5). Although there were major commonalities of the three methods as shown in Figure 5, the 44 genes that could not be detected by the RAW method represented 28% of all AML-related genes that changed and highlighted the fact that data compression can cause substantial loss of biologically significant information (Table 2). Included in the 44 genes not identified by the RAW/SAM combination were: *Aml1*, which is frequently found translocated and fused with the *Eto* gene to form the *Aml1-Eto* fusion protein in AML and which was upregulated by >5-fold in the AML samples; and *Nqo-1*, which is reported as suppressed in leukemia, was downregulated in the AML spleen samples. These data demonstrated again that the MBCB model is much more sensitive in detecting biologically relevant gene expression changes. Table 3 lists 12 genes that show significant expression differences identified by MBCB, but not by either the RAW or RMA methods. This list includes several important apoptosis-related genes such as *Tnfrsf6*/*Fas* and *Caspase 3*.

Improved sensitivity for signaling pathway analysis with the MBCB method

IPA was used to study gene signaling pathway that was involved in biological processes of AML. Differentially expressed genes and their fold changes that were obtained by using the MBCB, RMA or RAW method, were uploaded into IPA for analysis. The results showed that in most gene groups, the MBCB method detected more genes than the RAW method (Figure 6A). One ontology category that showed the greatest difference in gene identified was the death receptor signaling pathway. With the MBCB method, *Tnfr2* and *NF-kappaB* mediated signaling pathway was clearly activated, while the RMA and RAW methods failed to allow the detection of this

Table 3. AML-related genes, and their expression ratios, identified through the MBCB methodology that were not detected when the RMA methodology was used

Gene	Ratio (AML versus Normal)	Description
Eif5a	2.40	Eukaryotic translation initiation factor 5A (Eif5a)
Nfkbia	2.35	Nuclear factor of κ-light chain gene enhancer in B-cells inhibitor, α (Nfkbia)
Tnfrsf6	1.94	Tumor necrosis factor receptor superfamily, member 6 (Tnfrsf6)
Ripk1	1.79	Receptor (TNFRSF)-interacting serine-threonine kinase 1 (Ripk1)
Cd68	0.60	CD68 antigen (Cd68)
Gzmb	0.60	Granzyme B (Gzmb)
Acp1	0.55	Acid phosphatase 1, soluble
Igf1	0.54	Insulin-like growth factor 1 (Igf1)
Igfbp3	0.47	Insulin-like growth factor binding protein 3 (Igfbp3)
Casp3	0.35	Caspase 3, apoptosis related cysteine protease (Casp3)
Lck	0.25	Lymphocyte protein tyrosine kinase (Lck)
Cxcl12	0.18	Chemokine (C-X-C motif) ligand 12 (Cxcl12), transcript variant 2

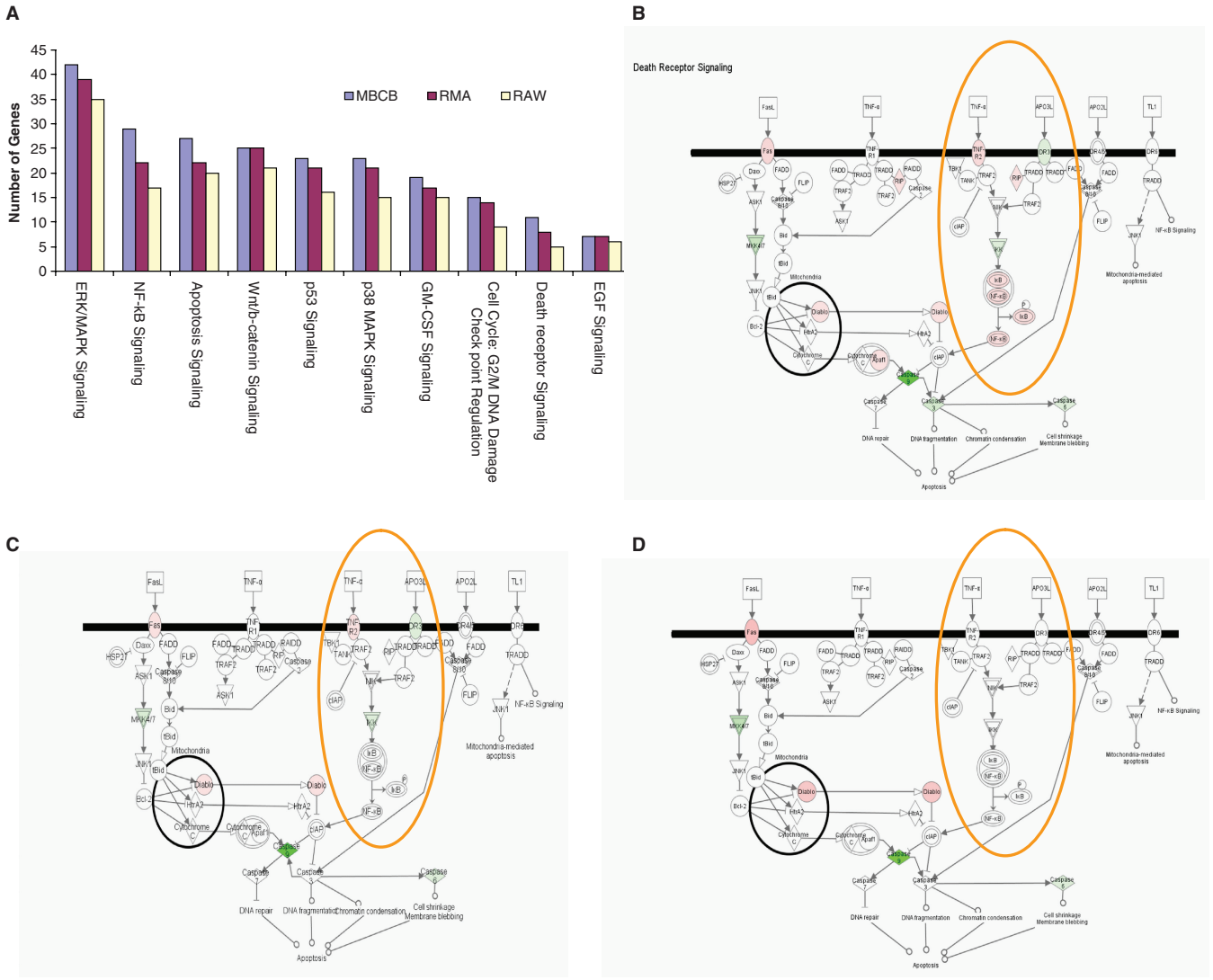


Figure 6. Comparison of pathway analysis on the data set using the MBCB, RMA or RAW background corrections. (**A**) IPA canonical pathway analysis depicting the top 9 signaling pathways that changed in AML tissues. The MBCB methodology identified more genes than either methodology in most categories. It also identified the death receptor pathway as the pathway that showed the most differences between the MBCB, RMA and RAW methods. (**B**) Scheme of death receptor pathway overlaid with the expression ratio of the MBCB data set. (**C**) Scheme of death receptor pathway overlaid with the expression ratio of the RMA data set. (**D**) Scheme of the death receptor pathway overlaid with the expression ratios from the RAW data set. Red, over-expressed in AML; green, under-expressed in AML; white, no significant change between AML and normal tissues. The cartoons show activation of *Tnfr2* and *NF-kappaB* signaling in the MBCB data set and the failure of detection of the same pathway by the RMA and RAW data sets.

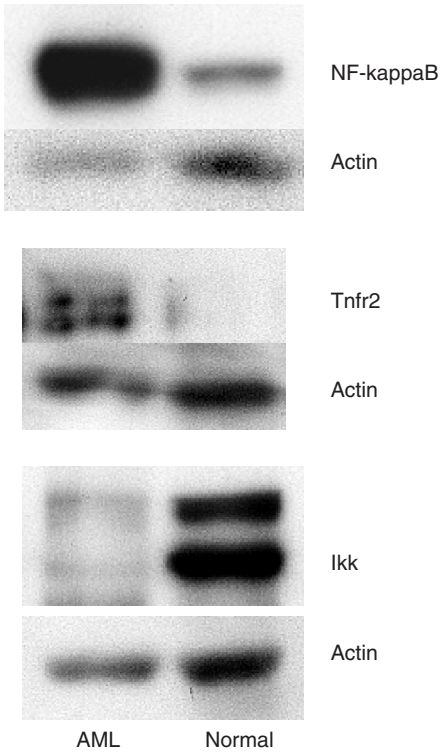


Figure 7. Western blot of proteins involved in the *Tnfr2* and *NF-kappaB* signaling pathway. The figures indicate upregulation of *Tnfr2* and *NF-kappaB*, and downregulation of *Ikk*. These results were consistent with gene-expression analysis using the MBCB data set and suggest active involvement of this pathway in mouse AML samples.

pathway (Figure 6B–D). Although the RMA method showed differential expression of several genes in this pathway, it failed to detect one of the key molecules in this pathway, that being *NF-kappaB* up-regulation (Figure 6C). The *NF-kappaB*, *Ikk* and *Tnfr2* expression changes were confirmed by western blot (Figure 7).

DISCUSSION

Traditional methodologies involving subtraction of local background or nonspecific control spots often generate negative intensities. These signals are often filtered out and subsequently result in missing values (15,16). There have been studies suggesting that performing a background correction should be avoided (17,18), and this holds good for data generated by Illumina arrays as suggested by a previous study (12). On the other hand, not performing background subtraction can exacerbate the problem of data compression, which can then result in an underestimation of signal. There are numerous reports where <50% of genes on an array have detectable signal (19–21). And while this could very well be true, it is likely that data compression had led to an increase in false negative genes which are then ignored in subsequent analyses such as predicting group membership or signal pathway and regulatory network discoveries.

Illumina BeadArrays have more than 50 000 beads linked with nonspecific oligonucleotides, which constitute a large population of true negative controls for hybridization. These beads are the same carriers of the gene-specific oligo probes. It is likely that a background correction method using these controls will outperform methods that subtract background from the localized array attachment substrate, be it glass, silica or other (7). There are two models available that consider the nonspecific probes as background controls aside from the local substrate. One is the model described by Li and Wong (13) and the other is RMA, both of which are widely used for Affymetrix GeneChip analysis. The Li and Wong model-based expression analysis approach in dCHIP is not directly applicable to Illumina arrays because there are no sensitivity differences amongst beads within the same bead type. Therefore, we modified the RMA method and adapted it for use on the Illumina platform. However, the RMA model is limited because it completely ignores information from nonspecific control beads. The data from this experiment and simulation results (22) suggested that RMA overestimated background values. This resulted in excess background subtraction and increased variances in arrays.

To overcome these limitations, we have described a model-based background correction method for Illumina BeadArrays. Our model was motivated by the different distributions of the observed intensity level for both genes and negative controls. We verified that this difference of distributions is typical for Illumina expression arrays as we could observe it in all hybridizations (data not shown). Our model makes use of the negative control beads of the arrays and the parameter estimation was based on the Markov chain Monte Carlo simulations (23). This method provides the positive estimation of background corrected expression level for each gene. We demonstrated that our MBCB model had a better correlation with QPCR results.

Our results clearly showed that both RMA and MBCB methods out-performed a no background correction approach. Although adopting RMA to Illumina background correction is still relatively new, MBCB incorporates the negative control beads, which is not available for the RMA method. MBCB improves the parameter estimation compared to the RMA method and we therefore, recommend the MBCB method for Illumina background correction.

We applied two popular significance analysis methods, SAM and BRB ArrayTools, to further validate the performance of our model. Both of the methods resulted in detecting the most differentially expressed genes when using the MBCB model when comparing with RMA and RAW methods, demonstrating improved sensitivity for gene discovery through the use of the MBCB model.

Through a literature search and with help of IPA software, we showed that the MBCB background correction model detected 44 more AML-associated genes considered to be differentially regulated than were found when the RAW methodology was used. As shown in Table 2, *Aml1* is a gene that has been shown frequently translocated and forms fusion protein with *Eto* (24,25). In our MBCB model, *Aml1* was overexpressed >5-fold in the *AML* samples when compared to nonleukemic samples, whereas in RAW data set, there was no significant

change of *Aml1*. Aberrations in cell cycle regulation and p53-dependent apoptosis have been frequently found in hematological malignancies and we found in the MBCB generated data set that several cell cycle and apoptosis-related genes, such as *Cdkn1a*, *Cdkn2a*, *Brca1* and *Nfil3*, were significantly changed in leukemic mice. These genes were not detected in the RAW data set. *Brca1* has been reported to be hypermethylated in therapy-related AML (26) and our results suggest that BRCA1 expression was suppressed in radiation-induced mouse AML model. On the other hand, lack of methylation was found for *p21* in AML (27), and our data showed an increased expression for this gene. *Myb* is associated with differentiation and proliferation in leukemia cell lines (28) and our data indicate a 3-fold overexpression in the leukemic samples. Also on the list, *Nqo1* a member of the NAD(P)H dehydrogenase (quinone) family, which encodes a cytoplasmic 2-electron reductase, was underexpressed in the leukemic samples in line with the notion that no or low *Nqo1* activity is associated with an increased risk of *de novo* acute leukemia in humans (29). The data above clearly indicate that if one chose not to perform a background subtraction there is a risk for significant loss of biological information. By using the RMA methodology, 32 of the 44 genes identified by the MBCB method were detected. Within the remaining 12 genes were several important apoptosis-associated molecules, such as *Tnfrsf6*/*Fas*, *Nfkbia*, *Casp3* and *Ripk1*. In fairness, there were six genes that showed significant change in either the RAW (four genes) or RMA methods (five genes) that were not detected when the MBCB method was used. In summary, our data supported the conclusion that MBCB discovered more biologically relevant findings than the other two methods.

Our data demonstrated that the new MBCB background correction model enhanced sensitivity and in the mean time, remained highly specific during the gene discovery process. This is evident by the fact that more differentially expressed genes were detected and had better correlations with QPCR data. We also demonstrated an enhanced ability to perform functional pathway analysis due to the higher sensitivity. By using the MBCB model, we were able to discover that *Tnfr2* and *NF-kappaB* mediated death receptor pathway was activated in the radiation-induced AML samples. Activation of the *NF-kappaB* pathway has been reported in human leukemia to support cancerous proliferation, resistance to apoptosis and sustain angiogenesis (30). Targeting *NF-kappaB* pathway activation via pharmacological inhibition induced apoptosis in human AML cells (31–33). It has been shown in human leukemia cells that the TNF signal machinery is necessary to switch cells between a proliferative versus an apoptotic phenotype (34,35). Sustained *Tnf*/*Tnfr2*-induced *NF-kappaB* signaling and transcription allow the cells to survive and proliferate. On the other hand, switching the *NF-kappaB* pathway off results in *Tnf*/*Tnfr1*-driven stimulation of proapoptotic pathways, such as sustained *Jnk* and *p38* MAPK activity. In our mouse AML expression profile generated without background subtraction (RAW), there was no suggestion of any *NF-kappaB* pathway activity. The RMA method indicated partial activation but failed to detect *NF-kappaB* upregulation. Using the MBCB model, it was clearly shown that the *Tnfr2* and *NF-kappaB* death receptor pathway was activated given the differential expression of *Tnfr2*, *Ikk* and *NF-kappaB*. Since the activation of this pathway has not been systemically confirmed in mouse AML models, western blotting was performed to validate the protein expression of these genes. The results were consistent with the gene-expression profile.

In summary, we have demonstrated that appropriate background correction (MBCB) will lead to better detection of differentially expressed genes, which then results in an improved sensitivity for performing gene function and pathway analysis. The MBCB is a robust model that avoids the problem of generating negative values after background subtraction, and substantially reduces data compression that occurs when using RAW data without background correction. Our data indicate that the large population of negative control features found on Illumina arrays are useful for estimating noise value and should be considered in array designs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Shane Scoggin of the Simmons Cancer Center Genomics Core, UT Southwestern Medical Center, for processing all samples for microarray analysis. This study was funded by grants from NASA NSCORS NAG9-1569, NNJ05HD36G, NCI CA06294 and NIH UL1RR024982. Funding to pay the Open Access publication charges for this article was provided by NASA.

Conflict of interest statement. None declared.

REFERENCES

1. Canales,R.D., Luo,Y., Willey,J.C., Austermiller,B., Barbacioru,C.C., Boysen,C., Hunkapiller,K., Jensen,R.V., Knight,C.R., Lee,K.Y. *et al.* (2006) Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.*, **24**, 1115–1122.
2. Guo,L., Lobenhofer,E.K., Wang,C., Shippy,R., Harris,S.C., Zhang,L., Mei,N., Chen,T., Herman,D., Goodsaid,F.M. *et al.* (2006) Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.*, **24**, 1162–1169.
3. Patterson,T.A., Lobenhofer,E.K., Fulmer-Smentek,S.B., Collins,P.J., Chu,T.M., Bao,W., Fang,H., Kawasaki,E.S., Hager,J., Tikhonova,I.R., *et al.* (2006) Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nat. Biotechnol.*, 24, 1140–1150.
4. Shi,L., Reid,L.H., Jones,W.D., Shippy,R., Warrington,J.A., Baker,S.C., Collins,P.J., de Longueville,F., Kawasaki,E.S., Lee,K.Y. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.
5. Shippy,R., Fulmer-Smentek,S., Jensen,R.V., Jones,W.D., Wolber,P.K., Johnson,C.D., Pine,P.S., Boysen,C., Guo,X., Chudin,E. *et al.* (2006) Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat. Biotechnol.*, **24**, 1123–1131.
6. Tong,W., Lucas,A.B., Shippy,R., Fan,X., Fang,H., Hong,H., Orr,M.S., Chu,T.M., Guo,X., Collins,P.J. *et al.* (2006) Evaluation

of external RNA controls for the assessment of microarray performance. *Nat. Biotechnol.*, **24**, 1132–1139.
7. Ritchie,M.E., Silver,J., Oshlack,A., Holmes,M., Diyagama,D., Holloway,A. and Smyth,G.K. (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics.*, **23**, 2700–2707.
8. Irizarry,R.A., Hobbs,B., Colin,F., Beazer-Barclay,Y.D., Antonellis,K., Scherf,U. and Speed,T.P. (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
9. McGee,M. and Chen,Z. (2006) Parameter estimation for the exponential-normal convolution model for background correction of Affymetrix GeneChip data. *Stat. Appl. Genet. Mol. Biol.*, **5**, Article 24.
10. Kuhn,K., Baker,S.C., Chudin,E., Lieu,M.H., Oeser,S., Bennett,H., Rigault,P., Barker,D., McDaniel,T.K. and Chee,M.S. (2004) A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res.*, **14**, 2347–2356.
11. Gunderson,K.L., Kruglyak,S., Graige,M.S., Garcia,F., Kermani,B.G., Zhao,C., Che,D., Dickinson,T., Wickham,E., Bierle,J. *et al.* (2004) Decoding randomly ordered DNA arrays. *Genome Res.*, **14**, 870–877.
12. Barnes,M., Freudenberg,J., Thompson,S., Aronow,B. and Pavlidis,P. (2005) Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.*, **33**, 5914–5923.
13. Li,C. and Hung Wong,W. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.*, **2**, RESEARCH0032.
14. 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.
15. Beissbarth,T., Fellenberg,K., Brors,B., Arribas-Prat,R., Boer,J., Hauser,N.C., Scheideler,M., Hoheisel,J.D., Schutz,G., Poustka,A. *et al.* (2000) Processing and quality control of DNA array hybridization data. *Bioinformatics*, **16**, 1014–1022.
16. Kooperberg,C., Fazzio,T.G., Delrow,J.J. and Tsukiyama,T. (2002) Improved background correction for spotted DNA microarrays. *J. Comput. Biol.*, **9**, 55–66.
17. Qin,L.X. and Kerr,K.F. (2004) Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res.*, **32**, 5471–5479.
18. Tran,P.H., Peiffer,D.A., Shin,Y., Meek,L.M., Brody,J.P. and Cho,K.W. (2002) Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res.*, **30**, e54.
19. Yuen,T., Wurmbach,E., Pfeffer,R.L., Ebersole,B.J. and Sealfon,S.C. (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.*, **30**, e48.
20. Vlachou,D., Schlegelmilch,T., Christophides,G.K. and Kafatos,F.C. (2005) Functional genomic analysis of midgut epithelial responses in Anopheles during Plasmodium invasion. *Curr. Biol.*, **15**, 1185–1195.
21. Arlinde,C., Sommer,W., Bjork,K., Reimers,M., Hyytia,P., Kiianmaa,K. and Heilig,M. (2004) A cluster of differentially expressed signal transduction genes identified by microarray analysis in a rat genetic model of alcoholism. *Pharmacogenomics J.*, **4**, 208–218.
22. Xie,Y., Ding,L., Xiao,G., Allen,J. and Story,M.D. (2007) Model based background correction for Illumina beadarray data. *2007 Proceedings of the Biometrics Section, American Statistical Association*, 338–342.
23. Gelman,A., Carlin,J., Stern,H. and Rubin,D. (2003) *Bayesian Data Analysis*, 2nd edn. London, CRC Press.
24. Mikhail,F.M., Sinha,K.K., Saunthararajah,Y. and Nucifora,G. (2006) Normal and transforming functions of RUNX1: a perspective. *J. Cell. Physiol.*, **207**, 582–593.
25. Blyth,K., Cameron,E.R. and Neil,J.C. (2005) The RUNX genes: gain or loss of function in cancer. *Nat. Rev. Cancer*, **5**, 376–387.
26. Scardocci,A., Guidi,F., D'Alo,F., Gumiero,D., Fabiani,E., Diruscio,A., Martini,M., Larocca,L.M., Zollino,M., Hohaus,S. *et al.* (2006) Reduced BRCA1 expression due to promoter hypermethylation in therapy-related acute myeloid leukaemia. *Br. J. Cancer*, **95**, 1108–1113.
27. Brakensiek,K., Langer,F., Kreipe,H. and Lehmann,U. (2005) Absence of p21(CIP 1), p27(KIP 1) and p 57(KIP 2) methylation in MDS and AML. *Leuk. Res.*, **29**, 1357–1360.
28. Corradini,F., Cesi,V., Bartella,V., Pani,E., Bussolari,R., Candini,O. and Calabretta,B. (2005) Enhanced proliferative potential of hematopoietic cells expressing degradation-resistant c-Myb mutants. *J. Biol. Chem.*, **280**, 30254–30262.
29. Long,D.J. II, Gaikwad,A., Multani,A., Pathak,S., Montgomery,C.A., Gonzalez,F.J. and Jaiswal,A.K. (2002) Disruption of the NAD(P)H:quinone oxidoreductase 1 (NQO1) gene in mice causes myelogenous hyperplasia. *Cancer Res.*, **62**, 3030–3036.
30. Karin,M. (2006) Nuclear factor-kappaB in cancer development and progression. *Nature*, **441**, 431–436.
31. Ohsugi,T., Horie,R., Kumasaka,T., Ishida,A., Ishida,T., Yamaguchi,K., Watanabe,T., Umezawa,K. and Urano,T. (2005) In vivo antitumor activity of the NF-kappaB inhibitor dehydroxymethylepoxyquinomicin in a mouse model of adult T-cell leukemia. *Carcinogenesis*, **26**, 1382–1388.
32. Griessinger,E., Imbert,V., Lagadec,P., Gonthier,N., Dubreuil,P., Romanelli,A., Dreano,M. and Peyron,J.F. (2007) AS602868, a dual inhibitor of IKK2 and FLT3 to target AML cells. *Leukemia*, **21**, 877–885.
33. Fabre,C., Carvalho,G., Tasdemir,E., Braun,T., Ades,L., Grosjean,J., Boehrer,S., Metivier,D., Souquere,S., Pierron,G. *et al.* (2007) NF-kappaB inhibition sensitizes to starvation-induced cell death in high-risk myelodysplastic syndrome and acute myeloid leukemia. *Oncogene*, **26**, 4071–4083.
34. Tucker,S.J., Rae,C., Littlejohn,A.F., Paul,A. and MacEwan,D.J. (2004) Switching leukemia cell phenotype between life and death. *Proc. Natl Acad. Sci. USA*, **101**, 12940–12945.
35. Choi,C.H., Xu,H., Bark,H., Lee,T.B., Yun,J., Kang,S.I. and Oh,Y.K. (2007) Balance of NF-kappaB and p38 MAPK is a determinant of radiosensitivity of the AML-2 and its doxorubicin-resistant cell lines. *Leuk. Res.*, **31**, 1267–1276.