Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 3 2012 *Article* 8

Combining Multiple Laser Scans of Spotted Microarrays by Means of a Two-Way ANOVA Model

Jérôme Ambroise, *Université catholique de Louvain*
Bertrand Bearzatto, *Université catholique de Louvain*
Annie Robert, *Université catholique de Louvain*
Benoit Macq, *Uninversité catholique de Louvain*
Jean-Luc Gala, *Université catholique de Louvain*

Recommended Citation:
Ambroise, Jérôme; Bearzatto, Bertrand; Robert, Annie; Macq, Benoit; and Gala, Jean-Luc (2012) "Combining Multiple Laser Scans of Spotted Microarrays by Means of a Two-Way ANOVA Model," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 3, Article 8.

DOI: 10.1515/1544-6115.1738

Combining Multiple Laser Scans of Spotted Microarrays by Means of a Two-Way ANOVA Model

Jérôme Ambroise, Bertrand Bearzatto, Annie Robert, Benoit Macq, and Jean-Luc Gala

Abstract

Motivation: Assessment of gene expression on spotted microarrays is based on measurement of fluorescence intensity emitted by hybridized spots. Unfortunately, quantifying fluorescence intensity from hybridized spots does not always correctly reflect gene expression level. Low gene expression levels produce low fluorescence intensities which tend to be confounded with the local background while high gene expression levels produce high fluorescence intensities which rapidly reach the saturation level. Most algorithms that combine data acquired at different voltages of the photomultiplier tube (PMT) assume that a change in scanner setting transforms the intensity measurements by a multiplicative constant.

Methods and Results: In this paper we introduce a new model of spot foreground intensity which integrates a PMT voltage independent scanner optical bias. This new model is used to implement a ”Combining Multiple Scan using a Two-way ANOVA” (CMS2A) method, which is based on a maximum likelihood estimation of the scanner optical bias. After having computed scanner bias, coefficients of the two-way ANOVA model are used for correcting the saturated spots intensities obtained at high PMT voltage by using their counterpart values at lower PMT voltages. The method was compared to state-of-the-art multiple scan algorithms, using data generated from the MAQC study. CMS2A produced fold-changes that were highly correlated with qPCR fold-changes. As the scanner optical bias is accurately estimated within CMS2A, this method allows also avoiding fold-change compression biases whatever the value of this optical bias.

KEYWORDS: microarray, saturation, multiscan

Author Notes: Funding: J.A. is funded by Nanotic/Tsarine, a project of the Region Wallonne of Belgium (convention number: 516250). Ambroise Jérôme: Information and Communication Technologies, Electronics and Applied Mathematics, Université catholique de Louvain. Bertrand Bearzatto: Center for Applied Molecular Technologies, Institute of Experimental and Clinical Research, Université catholique de Louvain. Annie Robert: Epidemiology and Biostatistics, Institute of Experimental and Clinical Research, Université catholique de Louvain. Benoit Macq: Information and Communication Technologies, Electronics and Applied Mathematics, Uninversité catholique de Louvain. Jean-Luc Gala: Center for Applied Molecular Technologies, Institute of Experimental and Clinical Research, Université catholique de Louvain.

1 Introduction

Microarrays are a powerful technology used in functional genomics that allow investigators to study the expression levels of thousands of genes simultaneously in a single assay. Among the types of studies that can be conducted with spotted microarrays, class comparison aims to detect differentially expressed genes. If one-color spotted microarrays are used, each sample is hybridized on a separate array. Gene expression levels are then computed from the fluorescence intensities emitted by corresponding spots. The most common scanning configuration uses a laser to excite the fluorescent dye and a photomultiplier tube (PMT) to detect the fluorescence signal. The fluorescence intensity of each spot is extracted after image acquisition.

Unfortunately, quantification of the fluorescence intensity does not always perfectly reflect the level of gene expression. Low gene expression levels produce low fluorescence intensities which tend to be confounded with the local background because of the presence of an additive noise partially caused by quantization, i.e. discretization in the analog-to-digital converter. On the other hand, high gene expression levels produce high fluorescence intensities which rapidly reach the saturation level. Indeed, considering that pixel intensities are coded on a 16-bits gray scale, the highest value that can be recorded cannot exceed 65 535. When pixels intensities from a spot are above the upper threshold of detection, those intensity values are truncated to the upper threshold. Consequently, the linearity between fluorescence intensity and gene expression level is not achieved on the whole range of mRNA copy number which, in turn, leads to biased estimations of the genes expression fold-changes.

The existing relationship between gene expression level and fluorescence intensity, as given by the calibration curve, depends on the scanner settings and, more specifically, on the voltage of the photomultiplier tube (PMT) which is easily adjustable. Indeed, fluorescent photons produced by excited spots are amplified by PMT voltage (or PMT gain), which, in turn, strongly influences the calibration curve. The linear dynamic range is the whole range of target concentrations where a linear relationship is noted between dye concentrations, an indicator of gene expression level, and fluorescence intensity. This curve shifts toward low level of expression when PMT voltage is increased. To obtain non-biased estimations of genes expressions fold-changes, it is mandatory to work with fluorescence intensities distributed within the linear dynamic range. PMT voltage should therefore be selected according to the intensity of the fluorescent signal that needs to be quantified. However, among the whole panel of genes to be quantified, some have a very weak expression level while others disclose a much higher expression pattern. Fluorescent signals associated with spots of interest on spotted microarrays

therefore sometimes appear to be extremely variable in such a way that selecting a well defined PMT voltage can be suboptimal, if not impossible. Most of the protocols recommend therefore to scan the microarray at a PMT voltage that minimizes the number of spots showing a fully saturated fluorescence signal or that are confounded with local background. Using this methodology is, however, far from being optimal when the identification of truly differentially expressed genes is at stake. A way to solve this issue is to improve the estimations of the gene expression fold-changes by scanning the spots at different PMT voltages and by combining the information extracted from the acquired image successively at different PMT voltages.

Some authors have developed approaches that use either pixel level data or spot summary data to calculate new values for the saturated spots. In that respect, Dudley, Aach, Steffen, and Church (2002) proposed to extrapolate, linearly and iteratively, the signal intensities of saturated spots using unsaturated data obtained at a lower PMT setting. Lyng, Badiee, Svendsrud, Hovig, Myklebost, and Stokke (2004) suggested correcting saturated spot intensities recorded at the highest PMT voltage by using corresponding intensities obtained at the lowest PMT voltage. Calculation of this correction factor assumes that a constant ratio exists between intensities of individual spots in both scans. de la Nava, Van Hijum, and Trelles (2004) presented two mathematical models based on linear and gamma curves in order to extend the dynamic range of gene expression data. Piepho, Keller, Hoecker, and Hochholdinger (2006) suggested a non-linear latent regression model to combine different signals from identical spots in order to correct for the biases caused by the saturation limit and to reduce technical errors for spots emitting low signals. Skibbe, Wang, Zhao, Borsuk, Nettleton, and Schnable (2006) compared Dudley's method with the multiple-scan method. In the multiple-scan method, the statistical analysis is performed separately at each PMT voltage, the smallest P-value being kept for each gene. Khondoker, Glasbey, and Worton (2006) proposed a statistical model based on a non-linear relationship with both additive and multiplicative error terms for estimating the "true intensities" of saturated spots. Gupta, Auvinen, Thomas, and Arjas (2006) suggested an approach based on a Bayesian latent intensity model in order to use several scans obtained at varying scanner sensitivities. Finally, Glasbey and Khondoker (2009) showed that combining multiple laser scans by maximizing the likelihood of a Gaussian structural regression model increases the power of detection of differential gene expression levels.

All the methods reported here above assume that a change in PMT voltage multiplies intensity measurements by a constant. However, Bengtsson, Jonsson, and Vallon-Christersson (2004) showed the presence of a channel specific bias present in two-color microarray data. They proposed a scanning protocol for two-color microarray and a constrained affine model that allow to identify and es-

timate the bias in each channel and to obtain a larger dynamic range as well as a greater signal-to-noise ratio. One advantage of the Bengtsson et al. method is that it takes into account the channel bias for computing the multiplicative effect of PMT voltage. Conversely, a drawback of this method is that the local background subtraction step is replaced by the subtraction of channel specific biases. Indeed, spatial heterogeneity of local background can induce a biased estimation of fold-changes, especially with one-color microarray data, where local background may differ markedly for each of the two control and reference spots referring to the same gene.

Accordingly, a first aim of this paper is to introduce a model for spot foreground intensity that integrates the PMT voltage multiplicative effect and a scanner optical bias which is independent of the PMT voltage. A second objective is to use this model to develop an algorithm combining data acquired at multiple PMT voltages while integrating the scanner optical bias. A third objective is to compare features of this algorithm with those of state-of-the-art methods. Such a comparison is performed by using datasets generated by the MicroArray Quality Control (MAQC) project with Dual chip Eppendorf arrays and with Taqman quantitative PCR (qPCR) (Shi, Reid, Jones, Shippy, Warrington, Baker, Collins, De Longueville, Kawasaki, Lee et al., 2006). Data from the MAQC project provide indeed a unique opportunity to compare advantages and disadvantages of various analysis methods and to try to reach a consensus on the most satisfactory way to analyze microarray data. To the best of our knowledge, the current analysis performed at different PMT voltages and comparison with the existing multiple scan algorithms is a new application of the MAQC dataset.

2 Model

In this paper, we introduce a new model for the measured spot foreground intensity which integrates the scanner-induced, PMT voltage-independent optical bias. The model integrates the signal resulting from hybridization, two background components and the PMT voltage related-effect. Each array scan generates for each spot i values, which represent the median of pixel intensities for both foreground and local background. These pixel intensities are dependent of the PMT voltage v (v = 300V, .. 700V), as defined for image acquisition. In the current work, these values named foreground intensities and local background intensities, are referred to as Y_{iv} and B_{iv}, respectively. As a first step, we assume a general measurement model with both additive and multiplicative errors.

$$Y_i = \alpha_i + \mu_i e^{\eta_i} + \varepsilon_i \tag{1}$$

Y_i is the measured intensity value, α_i is the mean background, μ_i is the true expression and η_i and ε_i are multiplicative and additive error terms considered to be independent and Gaussian-distributed. This widely used model in analytical chemistry (Rocke and Durbin, 2001) was also applied to gene expression microarray in numerous studies (Rocke and Durbin, 2001, Durbin, Hardin, Hawkins, and Rocke, 2002, Durbin and Rocke, 2003, Huber, Von Heydebreck, Sueltmann, Poustka, and Vingron, 2003, Rocke and Durbin, 2003, Cui, Kerr, and Churchill, 2003, Lin, Du, Huber, and Kibbe, 2008). In this paper, we suggest dividing the background intensity α_i into two distinct components.

$$\alpha_i = \delta + \lambda_i \tag{2}$$

The first component, δ, is the optical bias which is generated by the scanner and therefore independent of spot i. The second component, λ_i, is the background due to glass slide autofluorescence and to unspecific hybridization. The latter is dependent on spatial localization, hence dependent on spot i.

When a spot is excited by the laser, photons are emitted and are transformed in an electron signal by striking the cathode. In the literature, it has been shown that the electron signal is proportional to the number of photons and to the PMT voltage (Khondoker et al., 2006, Gupta et al., 2006). Increasing the scanner PMT voltage multiplies each intensity measurement by a constant. Besides this observation, Bengtsson et al. (2004) demonstrated a channel-specific bias when studying two-color microarrays scanned at multiple PMT voltages. Decomposing the local background α_i, as in equation 2, enables us to consider the scanner optical bias (δ) as being independent of the PMT voltage, as given by the following formula, where β_v stands for the PMT multiplicative effect on recorded intensities:

$$Y_{iv} = \delta + \beta_v(\lambda_i + \mu_i)e^{\eta_{iv}} + \varepsilon_{iv} \tag{3}$$

Introducing a scanner optical bias when correcting the saturated spots helps avoid underestimation of correction factors. Model 3 remains valid up to the saturation threshold. As it is reasonable to assume that background intensity induced by the glass slide auto-fluorescence and the unspecific hybridization (λ_i) equally affects foreground and local background areas of a spot, local background intensity can be modeled as follows:

$$B_{iv} = \delta + (\beta_v(\lambda_i))\, e^{\eta_{iv}} + \varepsilon_{iv} \tag{4}$$

Random error terms η_{iv} and ε_{iv} follow a normal distribution with a 0 mean and variances of σ_η^2 and σ_ε^2, respectively. For any spot without specific hybridization ($\mu_i = 0$), the probability of obtaining a local background higher than foreground ($B_{iv} > Y_{iv}$) is equal to 0.5, due to both random error terms. This probability decreases when specific hybridization intensity (μ_i) and PMT voltage (β_v) increase. This is the reason why the number of spots whose foreground intensity is lower than local background decreases at a higher PMT voltage (Table 1).

3 Methods

3.1 Data

MAQC project data generated on Eppendorf Dualchip (Shi et al., 2006) were used in this study. Eppendorf Dualchip analyzes the expression level of 294 genes from four distinct samples assessed in five replicates on three different sites. As data from site 2 were affected by annotation problems, we only used data from sites 1 and 3. Regarding the samples assessed in the MAQC study, sample A corresponds to the Universal Human Reference RNA (UHRR) from Stratagene while sample B corresponds to the Human Brain Reference RNA (HBRR) from Ambion. Samples C and D are a mixture from the original samples. In order to maximize the range of resulting fold-changes, it was decided to use samples A and B rather than C or D. Each array was scanned at three PMT voltages different from the PMT used in each site, namely 50, 70 and 100 for site 1, and 450, 700 and 800 for site 3. Raw data were downloaded from the Gene Expression Omnibus (GEO) repository (GEO accession: GSE5350) (Edgar, Domrachev, and Lash, 2002).

On Eppendorf Dualchip platforms, each gene expression level is assessed on three replicated spots. In addition to these spots, Eppendorf arrays contain internal standard spots as well as negative and positive hybridization control spots, so that a total of 1089 spots are present on each array (3 * 363). Accordingly, each site provides data for 10,890 spots (2 samples * 5 biological replicates * 1089 spots). In the current study, foreground intensities above 50 000 were considered as partially saturated. Lyng et al. (2004) have indeed shown that expression ratios are independent from PMT voltage within the intensity range from 200 to 50 000. As displayed in Table 1, data acquired at low PMT voltage contain no saturated spot but several spots whose foreground intensities are lower than local background. Conversely, data acquired at high PMT voltage contain numerous saturated spots but very few spots that have a foreground intensity lower than the local background. This ob-

servation illustrates the increase in signal-to-noise ratio with the PMT voltage and prompts the use of methods combining data acquired at multiple PMT voltages.

Normalized qPCR data, downloaded from the GEO repository, and linear models from the *limma* package, were used to calculate gene expression fold-changes between samples A and B. In the current study, these values are referred to as gold-standard fold-changes. All data from the MAQC study used in this study are summarized in Table 2.

Table 1: Number of spot whose foreground intensity is lower than local background intensity ($Y_{iv} \leq B_{iv}$) and whose fluorescence signal is saturated ($Y_{iv} > 50\,000$) from Eppendorf platform in sites 1 and 3.

PMT voltage	Site 1		Site3	
	$Y_{iv} \leq B_{iv}$	Sat.	$Y_{iv} \leq B_{iv}$	Sat.
Low	670	0	1316	0
Medium	409	261	459	135
High	350	1161	392	382

Table 2: Data generated from MAQC study which are used in this study

technology	Platform and site	Sample A	Sample B
		n = number of replicates	
Spotted cDNA	Eppendorf: Site 1	5	5
Spotted cDNA	Eppendorf: Site 3	5	5
qPCR	Taqman	4	4

3.2 CMS2A

A second objective of this paper is to present CMS2A, a new algorithm based on model 3 for combining data acquired at different PMT voltages. At first, the CMS2A method estimates the scanner optical bias δ by using a maximum likelihood approach. As model 3 is only valid for unsaturated spots, all saturated spots ($Y_{iv} > 50\,000$) are first filtered out. The lowest foreground intensities ($Y_{iv} < B_{iv} + 20$) are also discarded in order to make additive error ε insignificant. When the scanner optical bias is subtracted from foreground intensities and a log2 transformation is applied, the following model is obtained:

$$\log 2(Y_{iv} - \delta) = log2(\beta_v) + log2(\lambda_i + \mu_i) + \eta_{iv} \quad (5)$$

Model 5 can be seen as a two-way ANOVA model. The approach consists in finding the value which maximizes the likelihood of the estimated model in a grid of possible δ values. Once the δ value is estimated, the last step is to calculate saturated values obtained at high PMT voltage using data acquired at lower PMT voltage. If data have been acquired at three voltages (1=low, 2=medium, 3=high), the following correction is applied on saturated spots acquired at high PMT voltage:

$$\hat{Y}_{i3} = ((Y_{i2} - \hat{\delta}) * \hat{F}_{cor}) + \hat{\delta} \quad (6)$$

Where $\hat{Y}_{i3}$ is the corrected intensity of spot i acquired at high PMT voltage, Y_{i2} is the intensity of spot i acquired at medium PMT voltage and $\hat{F}_{cor}$ is the estimated correction factor. This correction factor is calculated based on the difference of the PMT voltage effects in the log2 scale, as estimated in the two-way ANOVA model.

$$\hat{F}_{cor} = 2^{(log2(\hat{\beta}_3) - log2(\hat{\beta}_2))} \quad (7)$$

If the intensity acquired at medium PMT voltage Y_{i2} is also saturated, the corrected value ($\hat{Y}_{i3}$) is obtained by using data acquired at low PMT voltage (Y_{i1}). In this case, $\hat{F}_{cor}$ is calculated based on the difference between $log2(\hat{\beta}_3)$ and $log2(\hat{\beta}_1)$. The principle of the CMS2A method is therefore to keep all non-saturated signals obtained at the highest PMT voltage and to calculate all saturated signals from data acquired at lower PMT voltage. While this procedure seems similar to the one reported by Lyng et al. and Dudley et al., the main difference lies in the multiplicative effect of PMT voltage, which is calculated after subtraction of the PMT voltage-independent scanner optical bias.

3.3 Comparison with state-of-the-art methods

A third objective of this study is to compare features of CMS2A with those of state-of-the-art multiple scan algorithms. Implementation details of methods tested in this study appear below:

Khondoker: Khondoker et al.'s algorithm is implemented in the "multiscan" method from the "multiscan" Bioconductor package. As background subtraction is not recommended in this method, two versions were applied: the first

version (Khondoker 1) consists in applying the multiscan method on foreground intensities without subtracting local background. The second version (Khondoker 2) consists in subtracting the local background intensities from the foreground, in order to filter out the negative values that can not be handled by the multiscan function and to apply the multiscan function on the remaining corrected intensities.

Bengtsson: Bengtsson et al.'s algorithm is implemented in the "calibrateMultiscan" function from the "aroma.light" Bioconductor package. In this method, all foreground intensities are calculated from intensities scanned at different PMT voltage. The authors recommend working with foreground intensities without subtracting local background noise in order to avoid negative values. However, a bias equivalent to the channel specific bias is calculated with a constrained affine model and subtracted from foreground intensities.

Lyng: Lyng et al.'s algorithm is based on the requirement of a constant ratio between individual spot intensities in two separate scans and was implemented in R. At first, a correction factor is calculated by using spots with intensities that lie within the range of 20.000 to 30.000 pixels at the highest PMT voltage. A correction of saturated spots (foreground intensities above 50 000) at high PMT voltage is performed by calculating the product of intensities obtained at lower PMT voltage and the correction factor. This method is very similar to the method proposed by Dudley et al. (2002).

Low: In this method, only data acquired at a single low PMT voltage are used. Spots of the weakly expressed genes tend to be confounded with the local background whereas spots of the highly expressed genes are not saturated.

High: In contrast with the above method, only data acquired at a single high PMT voltage are used. Spots of the weakly expressed genes tend to be above the local background noise whereas spots of highly expressed tend to reach the saturation threshold.

Medium: This is an alternative to both previous methods. In this case, data are only acquired at a single medium PMT voltage.

After having combined all data scanned at different PMT voltages, the 'Edwards' background correction method was applied. The 'Edwards' background correction method minimizes the fold-change compression and optimizes the correlation between microarray fold-changes and qPCR fold-changes (Ambroise, Bearzatto, Robert, Govaerts, Macq, and Gala, 2011), while avoiding the missing values caused by the log2 transformation. After background correction and log2 transformation, data were normalized using internal standards as recommended by the manufacturer. Considering the relatively low number of biological replicates, the *eBayes* algorithm (Smyth, 2004) from the *limma* Bioconducor package was then used to calculate the log2 fold-changes (Allison, Cui, Page, and Sabripour, 2006).

4 Results and Discussion

4.1 CMS2A

CMS2A was applied to data from site 1 and site 3 to combine intensities acquired at low, medium and high PMT voltages. The scanner optical bias δ of each array was estimated separately (Table 3) by using the methodology described in the method section. Effects of the CMS2A method on the first arrays of sites 1 and 3 are illustrated in figures 1 and 2, respectively. The data representation used to monitor and detect the presence of a scanner bias (figures 1-A and 2-A) consists of 3 sets of points, each forming a line. The top line corresponds to all foreground log-intensities scanned at high voltage as a function of foreground log-intensities scanned at high voltage whose saturated spots were corrected. This line is therefore the identity line except for saturated intensities (log2 intensities higher than $log2(62\,635)$). The lowest line corresponds to all foreground log-intensities scanned at low PMT voltage as a function of foreground log-intensities scanned at high voltage whose saturated spots were corrected. The intermediate line corresponds to data acquired at a medium PMT voltage. As the lines are not parallel, it can be stressed that changing the PMT voltage has no perfectly constant multiplicative effect on recorded intensities because of the presence of a PMT voltage-independent scanner optical bias.

In figures 1-B and 2-B, the log-likelihood of ANOVA model 5 is plotted as a function of the value of the scanner optical bias value (δ). The maximum of the log-likelihood was observed for a value of δ equal to 58.7 and 22.8 for Site 1 and Site 3, respectively. The δ bias value was subtracted from the foreground intensities, the effect of this correction being shown in figures 1-C and 2-C. The lines corresponding to each of the three PMT voltages are now strictly parallel. These results confirm that a change in PMT voltage has a multiplicative effect on the foreground intensities that is constant after subtracting the PMT voltage-independent scanner optical bias. Finally, the correction of the saturated spots was performed using equation 6 (figures 1-D and 2-D).

This procedure was used to estimate the optical bias δ from the data of each array and each site (Table 3). Ten estimations of δ were therefore obtained for each scanner (each site), with a reliability coefficient of 0.998. This high reliability coefficient demonstrates the presence of the scanner-specific optical bias δ as well as the high stability of CMS2A for estimating δ. This bias is probably caused by a dark current generated inside the PMT, i.e., the current produced in absence of photon (Van Wijk, Kobayashi, and Van Wijk, 2006). Electrons do indeed undergo thermal excitation in the vacuum of the tube, hence producing a bias on scanner-recorded intensities. If this hypothesis is true, variability of δ estimations should, at least

partly, be related to the reading conditions variability which occurs during the scanning procedure. The dark current intensity is indeed dependent on the temperature inside the PMT.

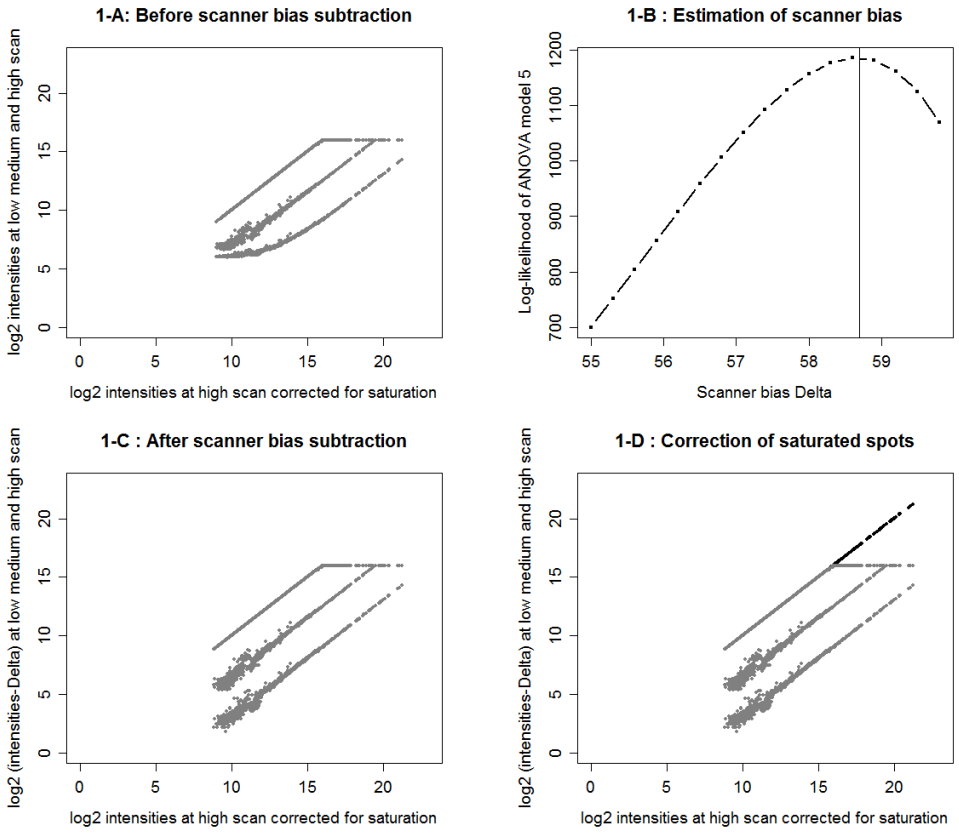

Figure 1: Effect of the CMS2A method on Eppendorf data from Site 1. 1-A: Scatter plot of foreground log2-intensities scanned at three PMT voltages (low, medium and high) as a function of foreground intensities scanned at high PMT voltage whose saturated spots were corrected with CMS2A. 1-B: Log-Likelihood of ANOVA model 5 as a function of the value of the PMT voltage-independent optical bias (δ). The maximum likelihood is obtained with a value of 58.7. 1-C: Scatter plot constructed as in 3-A but on foreground intensities whose the scanner optical bias (δ=58.7) was subtracted from. 1-D: Correction of saturated spots obtained at high PMT Voltage. Original data are plotted in gray while corrected data are plotted in black.

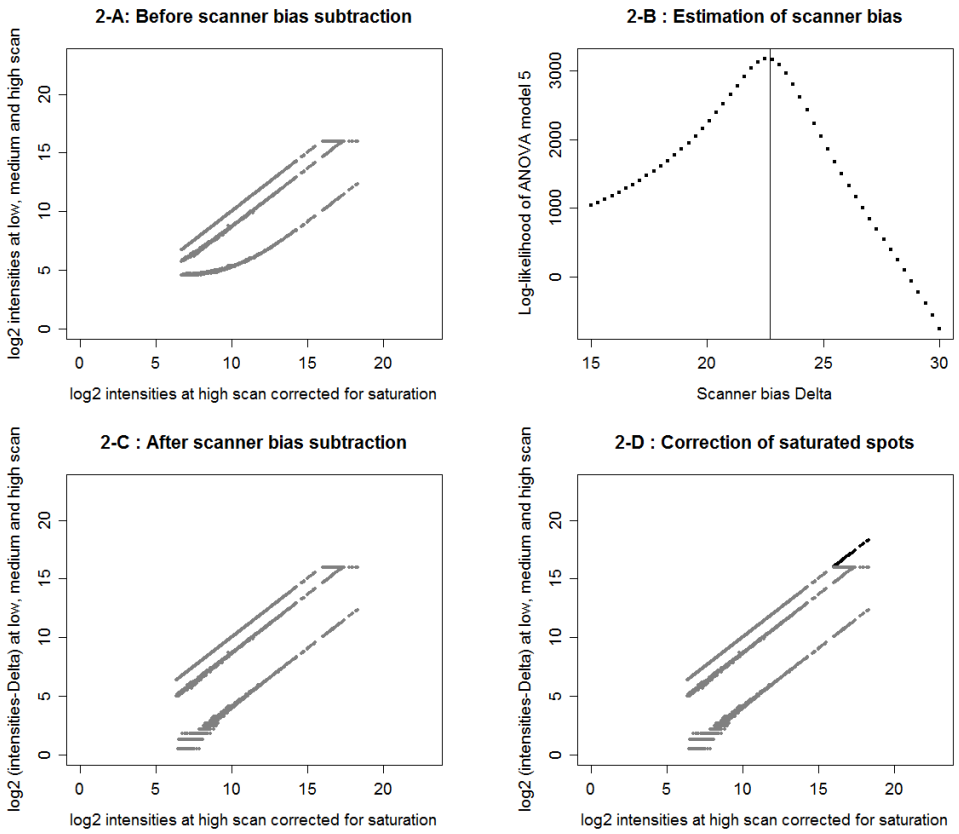

Figure 2: Effect of the CMS2A method on Eppendorf data from Site 3. The maximum likelihood is obtained for an optical bias (δ) value of 22.8.

Table 3: Estimations of the scanner optical bias (δ) for both sites

Replicate	Estimations in site 1		Estimations in site 3	
	Array A	Array B	Array A	Array B
Sample 1	58.7	58.5	22.8	22.2
Sample 2	58.1	57.4	19.6	19.6
Sample 3	57.9	56.9	22.2	20.2
Sample 4	57.9	56.6	19.6	20.2
Sample 5	56.6	58.2	20.8	20.2

4.2 Comparison with state-of-the-art methods

Features of CMS2A method were compared with those of state-of-the art methods. A first feature is the estimation of the correction factor between scans obtained at

low and high PMT voltages. A second feature is the correlation between qPCR fold-changes and microarray fold-changes.

4.2.1 Multiplicative factor

For all multiple scan methods, the multiplicative effects of the PMT voltages on recorded intensities are calculated. The multiplicative effects obtained with each of the methods assessed in this study are displayed in Table 4. Regarding site 1, the high scanner optical bias value ($\delta \approx 58$) induced high differences between multiplicative effects obtained by the different methods. Multiplicative effect between low and high voltage is close to 120 for the methods which take into account the optical scanner bias (CMS2A, Bengtsson, Khondoker 2) while it is close to 90 for those that did not. We consider here that the Khondoker 2 method takes into consideration the scanner optical bias when calculating the multiplicative effect. With the latter method, local background intensities B_{iv}, which include the scanner optical bias (see equation 4), are indeed subtracted from foreground intensities before applying the "multiscan" function. Such an amendment of the first version of the Khondoker method therefore leads to less variable estimations of the multiplicative effect, an observation in line with the existence of a scanner optical bias. In site 3, the scanner bias had a lower value ($\delta \approx 21$), bringing the multiplicative effect estimations of each method closer to each other.

Table 4: Correction factor between scans obtained at low and high PMT voltages

Method	site1				site3			
	sample A		sample B		sample A		sample B	
	Fcor	sd	Fcor	sd	Fcor	sd	Fcor	sd
CMS2A	122.4	1.6	122.2	1.1	62.3	1.3	62.5	2.3
Lyng	95.1	0.9	95.3	1.0	58.3	1.1	56.4	3.6
Bengtsson	122.0	1.4	122.7	1.9	52.2	3.4	50.9	8.7
Khondoker 1	90.2	4.5	44.3	2.8	44.8	3.0	18.7	4.1
Khondoker 2	123.6	1.1	124.6	1.6	61.1	1.2	60.5	2.3

4.2.2 Correlation between qPCR fold-changes and microarray fold-changes

Intraclass correlation (ICC) and product-moment correlation (r) coefficients were calculated between log2 qPCR fold-changes and log2 Eppendorf fold-changes from both sites (Site 1 and 3) after each multiple scan method (Table 5). The intraclass correlation coefficient is designed to assess consistency or conformity between two

or more quantitative measurements (Muller and Buttner, 1994) while the product-moment correlation coefficient is a measure of the strength of linear dependence between two variables. Correlation coefficients were calculated using the 133 genes for which expression level is commonly quantified on the Eppendorf Dualchip and by qPCR. However, for the 'High' method in site 1, fold-changes were only obtained for 124 of the 133 commonly measured genes because 9 of these genes analyzed at high PMT-voltage in this site had saturated spots in all replicate biological samples.

Table 5: Intraclass correlation (ICC) and product-moment correlation (r) between log2 fold-changes obtained with qPCR and with microarrays.

Method	Site 1		Site 3	
	ICC	r	ICC	r
Low	0.723	0.839	0.679	0.817
Medium	0.718	0.844	0.694	0.824
High	0.667	0.825	0.693	0.826
CMS2A	0.722	0.849	0.693	0.826
Lyng	0.718	0.851	0.692	0.826
Bengtsson	0.595	0.814	0.524	0.778
Khondoker 1	0.511	0.802	0.558	0.786
Khondoker 2	0.716	0.843	0.692	0.824

The highest correlation coefficients were obtained for CMS2A and Lyng. While these two methods produced similar correlation coefficients, the corrected intensities obtained with each method can differ markedly (Figure 3). The differences between corrected values were more important in Site 1, where a lot of saturated spots were observed (N SAT = 1161) at high PMT voltage and where a high scanner optical bias ($\delta \approx 58$) was calculated using CMS2A. Conversely, the corrected values were close to each other in site 3, where only few saturated (N SAT = 382) spots were observed and where a low scanner optical bias ($\delta \approx 21$) was calculated using CMS2A. The differences between results produced with CMS2A and Lyng increases therefore with the proportion of saturated spots and with the value of the scanner optical bias.

5 Conclusion

In this paper we introduced a new model of spot foreground intensity including the effect of PMT voltage as well as that of optical scanner bias δ, which is independent of the PMT voltage. The presence of this optical bias δ was confirmed on

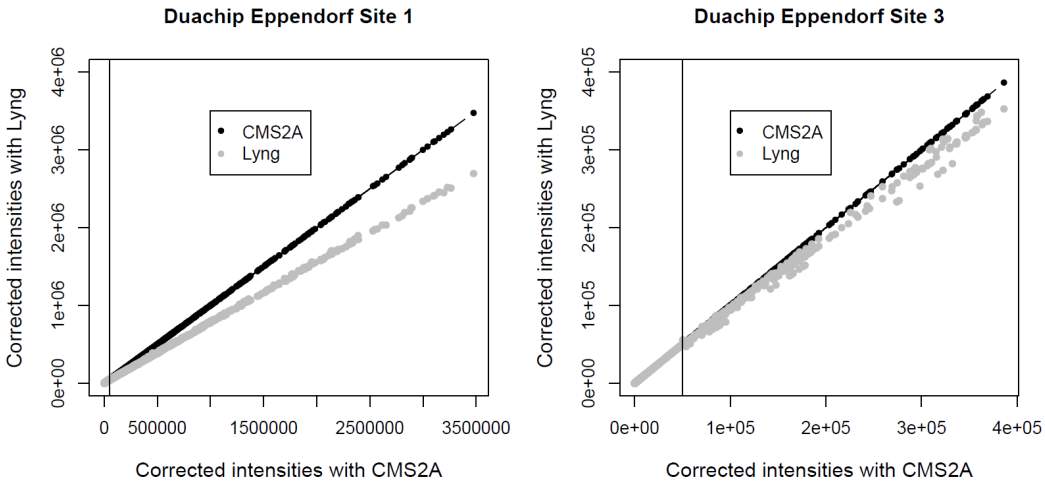

Figure 3: Comparison of the corrected values obtained with the Lyng et al. method and with CMS2A. All points that lie on the right side of the vertical line (saturation threshold of 50 000) are corrected intensities

Eppendorf dualchip data generated by the MAQC project by plotting the intensities acquired at different PMT voltages in the log2 scale (figures 1-A and 2-A). Despite the smaller value of the scanner bias δ in site 3 (21 compared to 58 in site 1), its effect was obvious, presumably because the spot segmentation was performed at a single PMT voltage and was used to extract foreground and background intensities at the three PMT voltages. This procedure ensures that differences observed between the spots intensities scanned at different PMT voltages are only caused by the change of this setting and not by an alteration of the pixels sets belonging to the spots. Conversely, if spot segmentation is performed on the images acquired at each PMT voltage, fluorescent intensities corresponding to the different PMT voltages tend to have a higher dispersion, as observed in data from site 1 and in the Khondoker et al.'s study .

Eppendorf data acquired at high PMT voltage contained a lot of saturated spots but few spots confounded with the local background. Conversely, data acquired at low PMT voltage contained a substantial number of foreground spots that did not stand out from local background but had no saturated spot. While Williams and Thomson (2010) showed that the signal-to-noise ratio decreases with PMT voltage, the opposite effect of PMT voltage on signal-to-noise ratio is commonly accepted in the bioinformatics community. Higher signal-to-noise ratio at higher PMT voltage explains the higher proportion of spot foregrounds above local background observed with the data from the MAQC study at high PMT voltage (see

Table 1). This observation prompts the use of methods combining data acquired at multiple PMT voltages.

The model proposed in the first part of the current study was used to develop CMS2A, a new algorithm that aims to correct saturated spots acquired at high PMT voltage by exploiting data acquired at lower PMT voltage. CMS2A is based on a maximum likelihood estimation of the scanner optical bias δ and on correction factors which reflect the multiplicative effects of PMT voltage on bias-corrected foreground intensities. The CMS2A was applied to Eppendorf data from two distinct sites enrolled in the MAQC project. Ten estimations of δ were calculated from data of each site (i.e., from each scanner from these sites) with a reliability coefficient of 0.998. Results confirm the presence of the scanner-specific optical bias δ and demonstrate the high stability of CMS2A to estimate it.

Characteristics of the new algorithm were compared with those of state-of-the-art methods. Among the latter, the method of Bengtsson et al. and both versions of Khondoker et al. calculated all intensities by taking into consideration data acquired at different PMT voltages. On the other hand, Lyng et al.'s method and CMS2A only take the intensities acquired at high PMT voltage for weakly expressed genes and correct the saturated spot obtained for highly expressed gene by multiplying their corresponding intensities acquired at lower PMT voltage by a correction factor. This strategy is motivated by the higher signal-to-noise ratio observed at high PMT Voltage. The difference between CMS2A and Lyng's methods lies in the correction factor computation, with the scanner optical bias being accounted for only in CMS2A.

A comparison was performed between the correction factors obtained with all multiple scan methods. In Site 1 (where $\delta \approx 58$), higher correction factors were obtained with the methods that consider scanner optical bias (CMS2A, Bengtsson, Khondoker 2) than with the methods that neglect it (Lyng, Khondoker 1). For both sites, product-moment and intraclass correlation coefficients were calculated between the qPCR fold-changes and the microarray fold-changes obtained after the different multiple scan methods. The lowest correlation coefficients were obtained with Bengtsson and Khondoker-1 methods which do not perform local background subtraction. Furthermore, current results are in line with data obtained in our previous study (Ambroise et al., 2011), showing that local background subtraction is an essential preprocessing step when processing one-color microarray data. With this type of array, the local background affecting each of the two spots measuring to the expression level of the same gene in two different conditions can indeed be different. In comparison, the Khondoker 2 method produced higher correlation coefficients. However, this method produced some missing values. Before applying this method, local background was indeed subtracted from foreground and negative values were filtered because the Khondoker method cannot consider them.

The highest correlation coefficients between microarray fold-changes and qPCR fold-changes were obtained with CMS2A and Lyng's method. Regarding the data from Site 3, correlation coefficients obtained with CMS2A and Lyng's method were very similar because the number of saturated spots at high PMT voltage was relatively small. Significant differences between corrected intensities were however obtained with both methods on data from Site 1 where a higher number of saturated spots and a higher optical bias ($\delta \approx 58$) were recorded. Results show that CMS2A and Lyng's method tend to produce similar results when the scanner optical bias is relatively low. In contrast, an increase of the scanner optical bias and of the proportion of saturated spots significantly impacts on the results. Overlooking the scanner optical bias in these cases leads indeed to a significant underestimation of the multiplicative effect of PMT Voltage, hence to an important underestimation of the corrected intensities. Using CMS2A therefore avoides underestimation biases whatever the value of the scanner optical bias may be. In the present study, CMS2A was applied on two microarray datasets generated in a class comparison application. In such microarray applications, underestimation biases can produce fold-change compressions. Using CMS2A therefore limits the fold-change compression effect, which is an important drawback of the microarray technology (Ritchie, Silver, Oshlack, Holmes, Diyagama, Holloway, and Smyth, 2007, Tarca, Romero, and Draghici, 2006, Ambroise et al., 2011). Moreover, the use of CMS2A is not limited to class comparison application. In fact, the method can also be applied to microarray data from a class prediction study (Tarca, Romero, and Draghici, 2006) in order to improve classification accuracy.

References

Allison, D., X. Cui, P. Page, and M. Sabripour (2006): "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, 7, 55.

Ambroise, J., B. Bearzatto, A. Robert, B. Govaerts, B. Macq, and J. Gala (2011): "Impact of the spotted microarray preprocessing method on fold-change compression and variance stability," *BMC Bioinformatics*, 12, 413.

Bengtsson, H., G. Jonsson, and J. Vallon-Christersson (2004): "Calibration and assessment of channel-specific biases in microarray data with extended dynamical range," *BMC bioinformatics*, 5, 177.

Cui, X., M. Kerr, and G. Churchill (2003): "Transformations for cDNA microarray data," *Statistical applications in genetics and molecular biology*, 2, 4.

de la Nava, G., S. Van Hijum, and O. Trelles (2004): "Saturation and quantization reduction in microarray experiments using two scans at different sensitivities," *Statistical Applications in Genetics and Molecular Biology*, 3, 1057.

Dudley, A., J. Aach, M. Steffen, and G. Church (2002): "Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range," *Proceedings of the National Academy of Sciences of the United States of America*, 99, 7554.

Durbin, B., J. Hardin, D. Hawkins, and D. Rocke (2002): "A variance-stabilizing transformation for gene-expression microarray data," *Bioinformatics*, 18, S105.

Durbin, B. and D. Rocke (2003): "Estimation of transformation parameters for microarray data," *Bioinformatics*, 19, 1360.

Edgar, R., M. Domrachev, and A. Lash (2002): "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic acids research*, 30, 207.

Glasbey, C. and M. Khondoker (2009): "Efficiency of functional regression estimators for combining multiple laser scans of cDNA microarrays," *Biometrical journal. Biometrische Zeitschrift*, 51, 45.

Gupta, R., P. Auvinen, A. Thomas, and E. Arjas (2006): "Bayesian hierarchical model for correcting signal saturation in microarrays using pixel intensities," *Statistical applications in genetics and molecular biology*, 5, 20.

Huber, W., A. Von Heydebreck, H. Sueltmann, A. Poustka, and M. Vingron (2003): "Parameter estimation for the calibration and variance stabilization of microarray data," *Statistical Applications in Genetics and Molecular Biology*, 2, 3.

Khondoker, M., C. Glasbey, and B. Worton (2006): "Statistical estimation of gene expression using multiple laser scans of microarrays," *Bioinformatics*, 22, 215.

Lin, S., P. Du, W. Huber, and W. Kibbe (2008): "Model-based variance-stabilizing transformation for Illumina microarray data," *Nucleic Acids Research*, 36, e11.

Lyng, H., A. Badiee, D. Svendsrud, E. Hovig, O. Myklebost, and T. Stokke (2004): "Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction," *BMC genomics*, 5, 10.

Muller, R. and P. Buttner (1994): "A critical discussion of intraclass correlation coefficients," *Statistics in Medicine*, 13, 2465–2476.

Piepho, H., B. Keller, N. Hoecker, and F. Hochholdinger (2006): "Combining signals from spotted cDNA microarrays obtained at different scanning intensities," *Bioinformatics*, 22, 802–807.

Ritchie, M., J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G. Smyth (2007): "A comparison of background correction methods for two-colour microarrays," *BIOINFORMATICS-OXFORD-*, 23, 2700.

Rocke, D. and B. Durbin (2001): "A model for measurement error for gene expression arrays," *Journal of Computational Biology*, 8, 557–569.

Rocke, D. and B. Durbin (2003): "Approximate variance-stabilizing transformations for gene-expression microarray data," *Bioinformatics*, 19, 966.

Shi, L., L. Reid, W. Jones, R. Shippy, J. Warrington, S. Baker, P. Collins, F. De Longueville, E. Kawasaki, K. Lee, et al. (2006): "The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements," *Nature biotechnology*, 24, 1151–1161.

Skibbe, D., X. Wang, X. Zhao, L. Borsuk, D. Nettleton, and P. Schnable (2006): "Scanning microarrays at multiple intensities enhances discovery of differentially expressed genes," *Bioinformatics*, 22, 1863.

Smyth, G. (2004): "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical applications in genetics and molecular biology*, 3, 1027.

Tarca, A., R. Romero, and S. Draghici (2006): "Analysis of microarray experiments of gene expression profiling," *American journal of obstetrics and gynecology*, 195, 373–388.

Van Wijk, R., M. Kobayashi, and E. Van Wijk (2006): "Anatomic characterization of human ultra-weak photon emission with a moveable photomultiplier and CCD imaging," *Journal of Photochemistry & Photobiology, B: Biology*, 83, 69–76.

Williams, A. and E. Thomson (2010): "Effects of scanning sensitivity and multiple scan algorithms on microarray data quality," *BMC bioinformatics*, 11, 127.