Journal of Data Science **5**(2007), 425-439

Comparisons of Gene Expression Indexes for Oligonucleotide Arrays

Mounir Aout
Laboratoire Génétique des Maladies Multi-factorielles-CNRS UMR8090

Abstract: High density oligonucleotide arrays have become a standard research tool to monitor the expression of thousands of genes simultaneously. Affymetrix GeneChip arrays are the most popular. They use short oligonucleotides to probe for genes in an RNA sample. However, important challenges remain in estimating expression level from raw hybridization intensities on the array. In this paper, we deal with the problem of estimating gene expression based on a statistical model. The present method is like Li and Wong model (2001a), but assumes more generality. More precisely, we show how the model introduced by Li and Wong can be generalized to provide new measure of gene expression. Moreover, we provide a comparison between these two models.

Key words: Gene expression, model-based estimation, oligonucleotide arrays.

1. Introduction

High density oligonucleotide expression arrays are now widely used in many area of biomedical research for measurements of gene expression. In the Affymetrix system, an array contains several thousands of genes and ESTs. To probe genes, oligonucleotides of length 25 bp are used. Typically, a mRNA molecule of interest (usually related to a gene) is represented by a probe set. Every probe set consists of 10-20 probe pairs. Every probe pair is composed of a perfect match PM, a section of the mRNA molecule of interest and a mismatch MM, which is identical to the perfect match probe except for the base in the middle $(13th)$ position. After RNA samples are prepared, labeled and hybridized with arrays, these are scanned and images are produced and processed to obtain an intensity value for each probe. These intensities, PM_{ij} and MM_{ij}, represent the amount of hybridization for arrays $i = 1, ...I$ and probe pairs $j = 1, ..., J$ for any given probe set. There has been considerable discussion over the appropriate algorithm for constructing single expression estimates based on multiple-probe hybridization

data. At present, there are several analytical methods to measure such intensities. However, we will only discuss the Affymetrix Microarray Suite *MAS4.0* and *MAS5.0* (1999 and 2001) and the method of Li and Wong *LW* (2001a). The *MAS 4.0* uses an average over probe pairs $PM_{ij}-MM_{ij}, j = 1, ...J$ for each array $i = 1, ...I$. This average difference (AD) is motivated by underlying statistical model: $PM_{ij} - MM_{ij} = \theta_i + \epsilon_{ij}, j = 1...J$. The expression index on array i is represented with the θ_i. AD is an appropriate estimate of θ_i if the error term ϵ_{ij} has equal variance for $j = 1, ..., J$. However, the equal variance assumption does not hold for GeneChip probe level data, since probes with larger mean intensities have larger variances, see Irizarry et al. (2003c). The latest version of this software *MAS5.0* computes the anti-log of a robust average of $log_2(PM_{ij} - CT_{ij})$. A corresponding statistical model is $log(PM_{ij} - CT_{ij}) = log(\theta_{ij}) + \epsilon_{ij}, j = 1, ..., J$. The basic disadvantage for this method is that there is no learning about probe characteristics, based on the performance of each probe across chips. To account for probe affinity effect, *LW* method suggests that $PM_{ij} - MM_{ij} = \theta_i\phi_j + \epsilon_{ij}, i = 1, ...I, j = 1, ...J, \epsilon \cong N(0, \sigma^2)$. The probe affinity effect is represented by ϕ_j. The main object of this paper is to generalize this model by considering separate models for PM and MM and making general assumptions on the errors.

This paper is organized as follows: The next section deals with a general model based on Li and Wong's model. We make general assumptions on the empirical variance and correlation of and between PM and MM, and estimate the parameters using maximum likelihood. Based on our analysis, we will show that our model gives an unbiased estimate of the expression index with low variance. Section 3 is concerned by a special case using PM only with inconstant variance. In addition, we compare how well these methods perform using the spike-in experiment HGU95A described in more details in the same section.

2. The Full Li and Wong Model

2.1 The full model: A simple case

Following Li and Wong, the PM and MM intensities are modeled as:

$$PM_{ij} = \nu_{ij} + \theta_i\alpha_j + \theta_i\phi_j + \epsilon_{ij}^P \tag{2.1}$$

$$MM_{ij} = \nu_{ij} + \theta_i\alpha_j + \epsilon_{ij}^M \tag{2.2}$$

where I denotes the number of samples and J denotes the number of probe pairs in a probe set. θ is the expression index, ν is a non-specific cross-hybridization term, α is the rate of increase of MM intensity and ϕ is the additional rate of increase of the PM intensity.

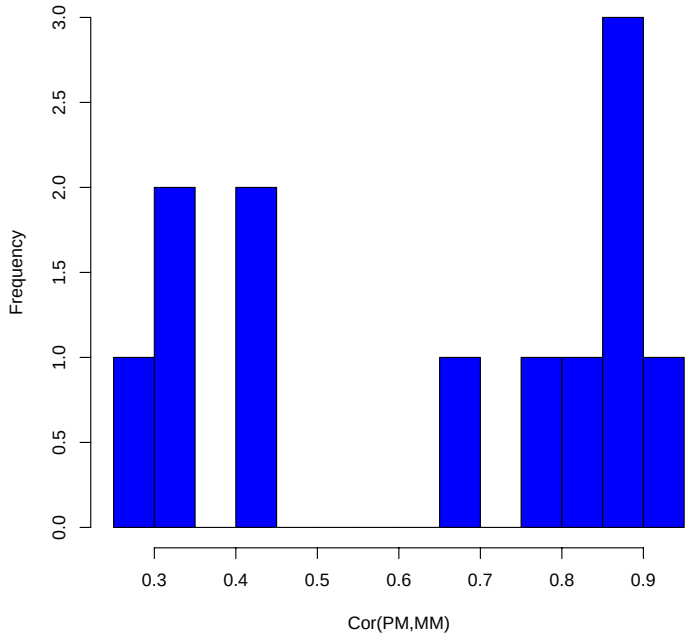

Figure 1: Correlation between PM and MM

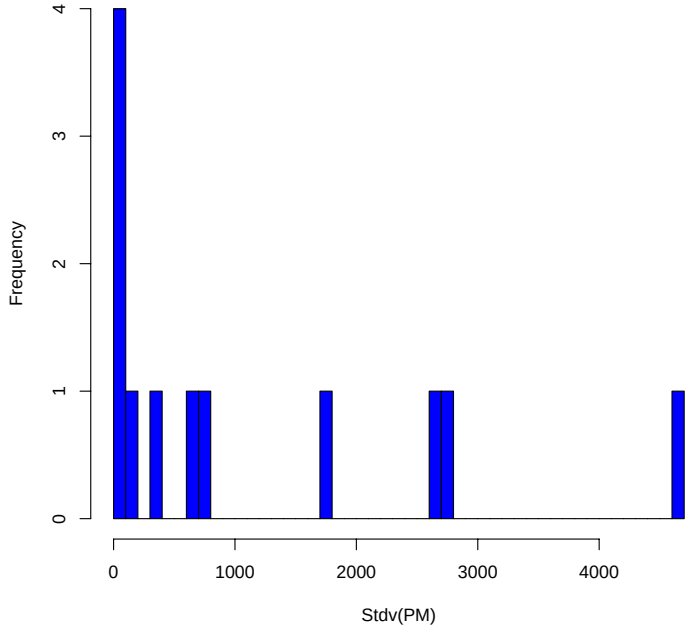

Figure 2: Standard deviation of PM

Although this model was introduced by Li and Wong, they have only treated the reduced case which we will call *RLW*:

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \epsilon_{ij}, \epsilon \cong N(0, \sigma^2)$$

Lemon et al.(2002) use the above equations, but assume that the PM and MM values are independent so their model describes the marginal distributions. Recently, Taib (2004) introduced a model in which it is assumed that the errors are correlated but with common variance and a constant correlation across samples. In general, these assumptions do not fit the observations as we will see later.

We propose then to augment the recent model to permit to the empirically observed correlation between PM and MM and the variances of PM and MM to change across the arrays as is shown in Figures 1-3. More precisely, we assume that the errors terms follow a bivariate normal distribution according to

$$\begin{pmatrix} \epsilon_{ij}^P \\ \epsilon_{ij}^M \end{pmatrix} \cong N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho_i \sigma_i^2 \\ \rho_i \sigma_i^2 & \sigma_i^2 \end{pmatrix}\right)$$

where σ_i^2 is the variance and ρ_i is the correlation coefficient. In the following this model will be called *FLW1*.

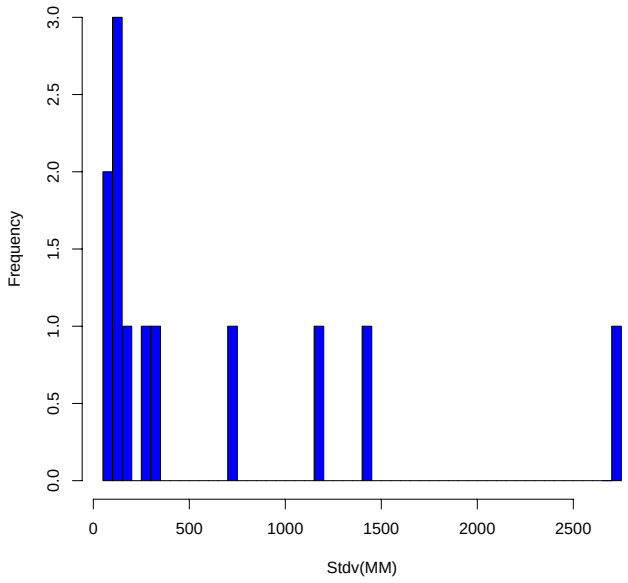

Figure 3: Standard deviation of MM

2.2 The estimates

Given data (PM_{ij}, MM_{ij}) we can estimate the parameters of our model using the maximum likelihood.
It is known that the likelihood function of the bivariate normal distribution can be expressed as:

$$\begin{aligned} L &= \prod_{i,j} L(PM_{ij}, MM_{ij}, \theta_i, \alpha_j, \phi_j, \nu_{ij}, \sigma_i, \rho_i) \\ &= \prod_{i,j} K_i \exp \frac{-1}{2\sigma_i^2(1-\rho_i^2)} \left[X_1^2 - 2\rho_i X_1 X_2 + X_2^2\right] \end{aligned}$$

where $X_1 = PM_{ij} - \nu_{ij} - \theta_i\alpha_j - \theta_i\phi_j$ and $X_2 = MM_{ij} - \nu_{ij} - \theta_i\alpha_j$.
The corresponding log likelihood function is

$$l = \sum_{i,j} log(K_i) - \sum_{i,j} \frac{1}{2\sigma_i^2(1-\rho_i^2)} \left[X_1^2 - 2\rho_i X_1 X_2 + X_2^2\right]$$

To get the estimates of the parameters we take the partial derivatives with respect to the corresponding parameters and we set the resulting expression equal to zero. Hence, we obtain:

$$\begin{aligned} \hat{\phi_j} &= \frac{\sum_i \frac{\theta_i}{\sigma_i^2(1-\rho_i^2)} \left[(PM_{ij} - \rho_i MM_{ij}) - (1-\rho_i)(\nu_{ij} + \theta_i\alpha_j)\right]}{\sum_i \frac{\theta_i^2}{\sigma_i^2(1-\rho_i^2)}} \\ \hat{\alpha_j} &= \frac{\sum_i \frac{\theta_i}{\sigma_i^2(1+\rho_i)} \left[PM_{ij} + MM_{ij} - 2\nu_{ij} - \theta_i\phi_j\right]}{\sum_i \frac{2\theta_i^2}{\sigma_i^2(1+\rho_i)}} \\ \hat{\nu_{ij}} &= \frac{(PM_{ij} - \theta_i\alpha_j - \theta_i\phi_j) + (MM_{ij} - \theta_i\alpha_j)}{2} \\ \hat{\theta_i} &= \frac{A_i + B_i}{\sum_j \phi_j^2 + 2(1-\rho_i)\alpha_j^2 + 2(1-\rho_i)\alpha_j\phi_j} \\ \hat{\sigma_i^2} &= \frac{\sum_j (X_1^2 - 2\rho_i X_1 X_2 + X_2^2)}{2J(1-\rho_i^2)} \\ \hat{\rho_i} &= \frac{\sum_j X_1 X_2}{J\sigma_i^2}, \end{aligned}$$

where $A_i = \sum_j \phi_j \left[PM_{ij} - \rho_i MM_{ij} - (1-\rho_i)\nu_{ij}\right]$, $B_i = (1-\rho_i) \sum_j \alpha_j \left[PM_{ij} + MM_{ij} - 2\nu_{ij}\right.$. The last two equations can be written as:

$$\hat{\sigma_i}^2 = \frac{\sum_j (X_1^2 + X_2^2)}{2J}$$

$$\hat{\rho_i} = \frac{2\sum_j X_1 X_2}{\sum_j (X_1^2 + X_2^2)}$$

These formulas have to be understood as steps in an iterative procedure that will lead to final estimates. In this case we will not be concerned by solving these equations. However, they are useful when it comes to deriving various properties. If we assume the other parameters to be known, It will be easy to see that $\hat{\theta_i}$ is an unbiased estimate of θ_i since $E\left[\hat{\theta_i}\right] = \theta_i$. For the variance, we get:

$$Var(\hat{\theta_i}) = \frac{\sigma_i^2(1-\rho_i^2)}{\sum_j \phi_j^2 + 2(1-\rho_i)\alpha_j^2 + 2(1-\rho_i)\alpha_j\phi_j} \tag{2.3}$$

2.3 Comparisons between *FLW1* and *RLW*

In this section, we will give a brief description of the reduced Li and Wong model and make a comparison between the estimates obtained in each model in terms of accuracy (bias) and precision (variance) .
For the *RLW model*, we recall that:

$$Y_{ij} := PM_{ij} - MM_{ij} = \theta_i\phi_j + \epsilon_{ij}, \sum_j \phi_j^2 = J, \epsilon_{ij} \cong N(0, \sigma^2)$$

The estimated expression index $\hat{\theta_i}$ can be obtained using the maximum likelihood or the least squares. Hence

$$\hat{\theta_i} = \frac{\sum_j Y_{ij}\phi_j}{\sum_j \phi_j^2}$$

The variance of the estimate, based on the assumptions of *RLW* model is

$$Var(\hat{\theta_i}) = \frac{2\sigma^2}{J}$$

But, based on the *FLW1* assumptions, on can easily show that

$$Var(\hat{\theta_i}) = \frac{2\sigma_i^2(1-\rho_i)}{\sum_j \phi_j^2} \tag{2.4}$$

and it is easy to see that (2.3) $\leq$ (2.4).

Given the Li and Wong Model, one could choose a suitable model based on the distribution of the errors. Another important point for the selection of the convenient estimate is the unbiasedness and low variance. Since we have shown that the corresponding $\hat{\theta_i}$ for our model is an unbiased estimate with low variance,

and according to the comparison above, we see that the full model should be a good choice.

2.4 The full model: A general case

In this section section, we propose to augment the last model to take into account the difference of the empirically observed variances between PM and MM as is shown in Figure 4.

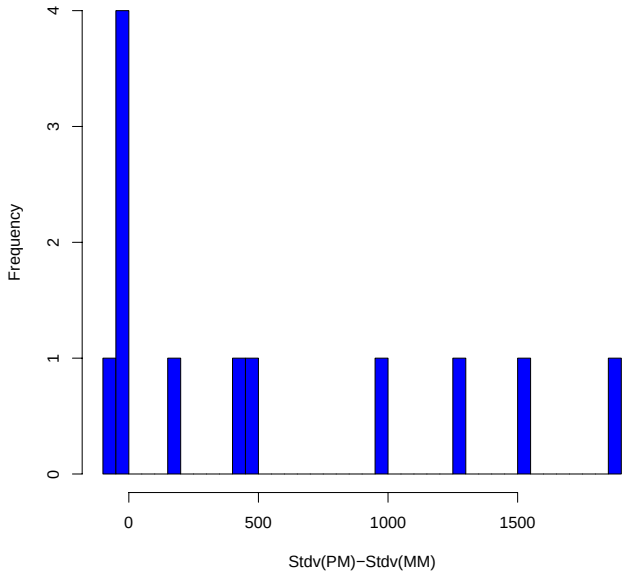

Figure 4: Difference between standard deviation of PM and MM

We will then assume that the error terms in 2.1 and 2.2 are distributed according to

$$\begin{pmatrix} \epsilon_{ij}^{P} \\ \epsilon_{ij}^{M} \end{pmatrix} \cong N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{1,i}^2 & \rho_i \sigma_{1,i}\sigma_{2,i} \\ \rho_i \sigma_{1,i}\sigma_{2,i} & \sigma_{2,i}^2 \end{pmatrix}\right)$$

where $\sigma_{1,i}^2$ and $\sigma_{2,i}^2$ are the variances and ρ_i is the corresponding correlation coefficient. From now on, we will call this model the *FLW2* model.

In this case, the likelihood function has the form

$$\begin{aligned} L &= \prod_{i,j} L(PM_{ij}, MM_{ij}, \theta_i, \alpha_j, \phi_j, \nu_{ij}, \sigma_{1,i}, \sigma_{2,i}, \rho_i) \\ &= \prod_{i,j} K_i \exp \frac{-1}{2(1-\rho_i^2)} \left[\frac{X_1^2}{\sigma_{1,i}^2} - 2\rho_i \frac{X_1 X_2}{\sigma_{1,i}\sigma_{2,i}} + \frac{X_2^2}{\sigma_{2,i}^2} \right] \end{aligned}$$

The same computations as above lead to the maximum likelihood estimates of the parameters:

$$\begin{aligned} \hat{\phi_j} &= \frac{\sum_i \frac{\theta_i}{\sigma_{1,i}^2(1-\rho_i^2)} \left[(PM_{ij} - \rho_i \frac{\sigma_{1,i}}{\sigma_{2,i}} MM_{ij}) - (1 - \rho_i \frac{\sigma_{1,i}}{\sigma_{2,i}})(\nu_{ij} + \theta_i \alpha_j) \right]}{\sum_i \frac{\theta_i^2}{\sigma_{1,i}^2(1-\rho_i^2)}} \\ \hat{\alpha_j} &= \frac{\sum_i \frac{\theta_i}{1-\rho_i^2} \left[a_i PM_{ij} + b_i MM_{ij} - \nu_{ij}(a_i + b_i) - a_i \theta_i \phi_j \right]}{\sum_i \frac{\theta_i^2}{1-\rho_i^2}(a_i + b_i)} \\ \hat{\nu_{ij}} &= \frac{a_i(PM_{ij} - \theta_i \alpha_j - \theta_i \phi_j) + b_i(MM_{ij} - \theta_i \alpha_j)}{a_i + b_i} \\ \hat{\theta}_i &= \frac{A_i + B_i}{\sum_j \frac{\phi_j^2}{\sigma_{1,i}^2} + (a_i + b_i)\alpha_j^2 + 2a_i \alpha_j \phi_j} \\ \hat{\sigma}_{1,i}^2 &= \frac{\sum_j X_1^2}{J} \\ \hat{\sigma}_{2,i}^2 &= \frac{\sum_j X_2^2}{J} \\ \hat{\rho}_i &= \frac{\sum_j X_1 X_2}{\sqrt{(\sum_j X_1^2)}\sqrt{(\sum_j X_2^2)}} \end{aligned}$$

where

$$\begin{aligned} A_i &= \sum_j \phi_j \left[\frac{1}{\sigma_{1,i}^2} PM_{ij} - \frac{\rho_i}{\sigma_{1,i}\sigma_{2,i}} MM_{ij} - a_i \nu_{ij} \right] \\ B_i &= \sum_j \alpha_j \left[a_i PM_{ij} + b_i MM_{ij} - (a_i + b_i)\nu_{ij} \right] \\ a_i &= \frac{1}{\sigma_{1,i}^2}(1 - \rho_i \frac{\sigma_{1,i}}{\sigma_{2,i}}) \qquad \text{and} \\ b_i &= \frac{1}{\sigma_{2,i}^2}(1 - \rho_i \frac{\sigma_{2,i}}{\sigma_{1,i}}) \end{aligned}$$

Given the other parameters, it is thus easy to see that the estimate $\hat{\theta}_i$ of the expression index is unbiased. For the variance we get

$$Var(\hat{\theta}_i) = \frac{1-\rho_i^2}{\sum_j \frac{\phi_j^2}{\sigma_{1,i}^2} + (a_i+b_i)\alpha_j^2 + 2a_i\alpha_j\phi_j} \tag{2.5}$$

On the other hand the variance of $\hat{\theta}_i$ based on the *RLW* is

$$Var(\hat{\theta}_i) = \frac{\sigma_{1,i}^2 + \sigma_{2,i}^2 - 2\rho_i\sigma_{1,i}\sigma_{2,i}}{\sum_j \phi_j^2} \tag{2.6}$$

and it is not easy to compare these variances. For example when $a_i \geq 0$ we have (2.5) $\leq$ (2.6). In general, we use data from the spike-in studies *HGU95A* and *HGU133* to make this comparison (see Figures 5-6 and we see that (2.5) $\leq$ (2.6) for almost all data (99 per cent of data)

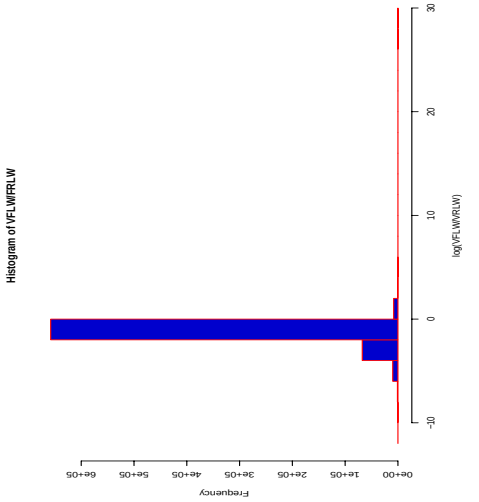

Figure 5: Ratio of log-variance between FLW and RLW- HGU133

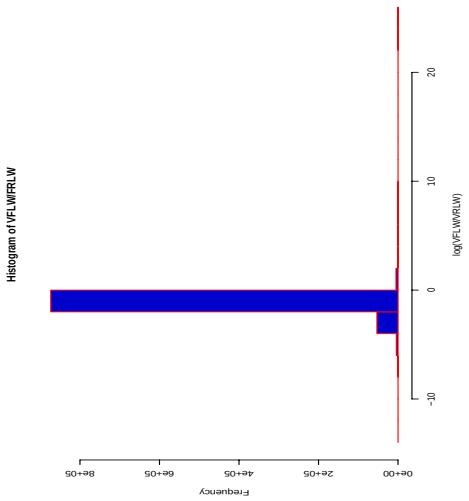

Figure 6: Ratio of log-variance between FLW and RLW- HGU95A

3. Numerical Results and Conclusions

3.1 The model based on PM only

It has been observed that some MM probes may respond poorly to the changes in the expression level of the target gene as discussed in Li and Wong (2001b). This phenomenon raised questions on the efficiency of using MM

probes, and led some investigators to calculate fold changes using only PM probes. To investigate the relative performance of PM-only using RLW and FLW, we modified the FLW model to estimate gene expression levels using only PM probes, and compared it to RLW. The modified FLW model becomes

$$PM_{ij} = \nu_j + \theta_i \phi_j + \epsilon_{ij}$$

where $\epsilon_{ij} \cong N(0, \sigma_i^2)$ The same procedure as above gives:

$$\hat{\phi_j} = \frac{\sum_i \frac{\theta_i}{\sigma_i^2}(PM_{ij} - \nu_{ij})}{\sum_i \frac{\theta_i^2}{\sigma_i^2}}$$

$$\hat{\nu_j} = \frac{\sum_i \frac{1}{\sigma_i^2}(PM_{ij} - \theta_i \phi_j)}{\sum_i \frac{1}{\sigma_i^2}}$$

$$\hat{\theta_i} = \frac{\sum_j \phi_j (PM_{ij} - \nu_j)}{\sum_j \phi_j^2}$$

$$\hat{\sigma}_i^2 = \frac{\sum_j (PM_{ij} - \theta_i \phi_j - \nu_j)^2}{J}$$

To evaluate how this model performs, we use a spike-in study *HGU95A* designed by Affymetrix.

3.2 Data

*HGU95A*GeneChip is a subset of the data used to develop and validate the *MAS5.0* algorithm. Human cRNA fragments matching 16 probe-sets on the *HGU95A* GeneChip were added to the hybridization mixture of the arrays at concentrations ranging from 0 to 1024 picoMolar. The same hybridization mixture, obtained from a common tissue source, was used for all arrays. The cRNAs were spiked-in at a different concentration on each array (apart from replicates) arranged in a cyclic Latin square design with each concentration appearing once in each row and column. Within each experiment, only the spike-in concentrations are varied, background is the same for all arrays. Fold change calculations are always made within experiment to ensure that only spiked-in genes will be differentially expressed. For more details see(http://www.affymetrix.com/analysis/downloadcenter2.affx).

3.3 Numerical results

This section is concerned by evaluating how the *FLW* based on PM-only performs. Actually we present a numerical comparison between *FLW* and *RLW* using the spike-in study *HGU95A* GeneChip.
we computed our estimates using the **R** environment see Ihaka and Gentleman (1996), which can be freely obtained from (http://cran.r-project.org) and the methods for Affymetrix Oligonucleotide Arrays R package described in Irrizary et al. (2003a), which is freely available as part of the Bioconductor project http://www.bioconductor.org. We then use a benchmark for Affymetrix GeneChip expression measures developed by Cope et al. (2003) which aims to evaluate and compare summaries of Affymetrix probe level data. We submitted our data to the corresponding webtool which is available at (http://affycomp.biostat.jhsph.edu). The results obtained are summarized in the table below (see Tables 1-2). We got results for RLW from (http://affycomp.biostat.jhsph.edu/AFFY2/rafajhu.edu/030519.1451/complete-assessment.pdf) and results corresponding to FLW are given in the Affycomp-webtool report. The score components for Table NR1 are as follows:

1. Signal detect slope: Slope obtained from regressing expression values on nominal concentrations in the spike-in data.
2. Signal detect R2: R-squared obtained from regressing expression values on nominal concentrations in the spike-in data.
3. AUC ($FP < 100$): Area under the ROC curve up to 100 false positives.
4. AFP, call if $fc > 2$: Average false positives if we use fold-change > 2 as a cut-off.
5. ATP, call if $fc > 2$: Average true positives if we use fold-change > 2 as a cut-off.
6. IQR: Interquartile range of log ratios among genes not differentially expressed.
7. Obs intended-fc slope: Slope obtained from regressing observed log-fold-changes against nominal log-fold-changes.
8. Obs (low)int-fc slope: Slope obtained from regressing observed log-fold-changes against nominal log-fold-changes for genes with nominal concentrations less than or equal to 2.
9. $FC = 2$, AUC ($FP < 100$): Area under the ROC curve up to 100 false positives when comparing arrays with nominal fold changes of 2.

10. $FC = 2$, AFP, call if $fc > 2$: Average false positives if we use fold-change> 2 as a cut-off when comparing arrays where nominal fold-changes are 2.

11. $FC = 2$, ATP, call if $fc > 2$: Average true positives if we use fold-change > 2 as a cut-off when comparing arrays where nominal fold-changes are 2.

and for Table 2:

1. Median SD: Median SD across replicates.

2. null log-fc IQR: Inter-quartile range of the log-fold-changes from genes that should not change.

3. null log-fc 99.9%: 99.9% percentile of the log-fold-changes if from the genes that should not change.

4. Signal detect R2: R-squared obtained from regressing expression values on nominal concentrations in the spike-in data.

5. Signal detect slope: Slope obtained from regressing expression values on nominal concentrations in the spike-in data.

6. low.slope: Slope from regression of observed log concentration versus nominal log concentration for genes with low intensities.

7. med.slope: As above but for genes with medium intensities.

8. high.slope: As above but for genes with high intensities.

9. Obs-intended-fc slope: Slope obtained from regressing observed log-fold-changes against nominal log-fold-changes.

10. Obs-(low)int-fc slope: Slope obtained from regressing observed log-fold-changes against nominal log-fold-changes for genes with nominal concentrations less than or equal to 2.

11. low AUC: Area under the ROC curve (up to 100 false positives) for genes with low intensity standardized so that optimum is 1.

12. med AUC: As above but for genes with medium intensities.

13. high AUC: As above but for genes with high intensities.

14. weighted avg AUC: A weighted average of the previous 3 ROC curves with weights related to amount of data in each class (low,medium,high).

For more details we refer to Irizarry et al. (2003c).

Table 1: Comparison results 1

	FLW-PMonly	RLW-PMonly	Perfection
Signal detect slope	0.480	0.533	1
Signal detect R2	0.852	0.846	1
AUC ($FP < 100$)	0.783	0.674	1
AFP, call if $fc > 2$	7.331	36.907	0
ATP, call if $fc > 2$	10.728	11.427	16
IQR	0.211	0.446	0
Obsintendedfc slope	0.471	0.523	1
Obs(low) intfc slope	0.204	0.317	1
FC =2, AUC ($FP < 100$)	0.460	0.167	1
FC=2, AFP, call if $fc > 2$	6.821	28.642	0
FC=2, ATP, call if $fc > 2$	1.000	1.250	16

Table 2: Comparison results 1

	FLW-PMonly	RLW-PMonly	Perfection
Median SD	0.066	0.132	0
null log-fc IQR	0.105	0.204	0
null log-fc IQR %99.9	0.656	1.437	0
Signal detect R2	0.852	0.846	1
Signal detect slope	0.480	0.533	1
low.slope	0.138	0.249	1
med.slope	0.547	0.641	1
high.slope	0.404	0.390	1
Obs-intended-fc slope	0.471	0.523	1
Obs-(low) int-fc slope	0.204	0.317	1
low AUC	0.295	0.041	1
med AUC	0.831	0.202	1
high AUC	0.612	0.011	1
weighted average AUC	0.427	0.079	1

4. Conclusions

We have presented a comparison between the reduced and full form of Li and Wong models using either the full bivariate or PM-only models. To understand the difference in the performance of calls generated by these two models, we

used both theoretical and numerical criteria. To make a decision as a choice of a model, one can make comparison in terms of accuracy(unbiased or low bias) and precision (low variance). We have shown that *FLW1* has a less variance than *RLW*. Furthermore, using the Spikein study, it seems clear that *FLW2* has considerably less variance than *RLW*. We also see that the PM-only model provides important improvements in various aspects compared to the same model based on *RLW*.

References

Affycomp-webtool (2005). Bioconductor expression assessment tool for affymetrix oligonucleotide arrays (affycomp). *Report*.

Affymetrix (1999). Microarray Suite User Guide, Version 4.

Affymetrix (2001). Microarray Suite User Guide, Version 5.

Cope, L. M., Irizarry, R. A., Jaffee, H., Wu, Z. and Speed, T. P. (2003). A benchmark for affymetrix geneChip expression measures. *Bioinformatics* **20**, 323-331.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299-314.

Irizarry, R., Gautier, L. and Cope, L. (2003a). An R package for analyses of Affymetrix oligonucleotide arrays. In *The Analysis of Gene Expression Data: Methods and Software* (Edited by Parmigiani, G., Garrett, E. S.,Irizarry, R. A. and Zeger, S. L.), 313-341. Springer.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U. and Speed, T. (2003c). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264.

Lemon, W. J., Palatini, J. J. T., Krahe, R. and Wright, F. A. (2002). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays.*Bioinformatics* **18**,1470-6.

Li, C. and Wong, W. H. (2001a). Model based analysis of oligonucleotide arrays:Expression index computation and outliers detection. *Proc. Natioanl Academy of Science* **98**, 31-36.

Li, C. and Wong, W. H. (2001b). Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biology* **2**, research0032.1-0032.11.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675-1680.

Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. John Wiley.

Taib, Z. (2004). Statistical analysis of oligonucleotide microarray data. *Comptes Rendus de l'Acadmie des Sciences* **237**, 175-180.

Received January 3, 2006; accepted April 23, 2006.

Mounir Aout
Department of Statistics and Data Processing
IUT de Caen (Lisieux)
11 Bd Jules Ferry
14100 Lisieux France
m.aout@lisieux.iutcaen.unicaen.fr