Two-color microarray data analysis taking into account probe-level inaccuracies

IME-USP

Adèle Helena Ribeiro, Roberto Hirata Júnior, Júlia Maria Pavan Soler
adele@ime.usp.br, hirata@ime.usp.br, pavan@ime.usp.br

Introduction

Most analyses of two-color microarray data are based on point estimation of the log-ratio of the two channels intensities. These estimates, commonly named M values, are conventionally obtained from some location measure of the pixel intensities of each channel, ignoring any imprecision. It is well known that the microarray technology is associated with many noise sources, and it has been shown that improved inferences can be obtained by including the inaccuracies involved and propagating them to downstream analysis. Using multivariate delta method, we propose new estimators for the mean and the variance of the M values, which take into account the probe-level inaccuracies in the analysis.

Improved Estimation

Classical estimators for the mean of the M and A values of the i-th probe of the j-th slide:

$$\hat{M}_{ij} \doteq \log_2(\mathbb{E}(R_{ij})) - \log_2(\mathbb{E}(G_{ij}));$$

$$\hat{A}_{ij} \doteq \frac{\log_2(\mathbb{E}(R_{ij})) + \log_2(\mathbb{E}(G_{ij}))}{2}.$$

Improved point-estimator for the mean of the M and A values of the i-th probe of the j-th slide, which are derived not only from the classical mean but also from the variance of the pixel intensities of both channels:

$$\overline{M}_{ij} \doteq \hat{M}_{ij} + \frac{1}{2\ln(2)}\left(\frac{\mathrm{Var}(G_{ij})}{\mathbb{E}^2(G_{ij})} - \frac{\mathrm{Var}(R_{ij})}{\mathbb{E}^2(R_{ij})}\right);$$

$$\overline{A}_{ij} \doteq \hat{A}_{ij} - \frac{1}{4\ln(2)}\left(\frac{\mathrm{Var}(G_{ij})}{\mathbb{E}^2(G_{ij})} + \frac{\mathrm{Var}(R_{ij})}{\mathbb{E}^2(R_{ij})}\right).$$

Interval Data Approach

We derived $\hat{\mathrm{Var}}(M_{ij})$, an estimator for the i-th probe log-ratio variance, in the j-th slide:

$$\frac{1}{\ln^2(2)}\left(\frac{\mathrm{Var}(R_{ij})}{\mathbb{E}^2(R_{ij})} + \frac{\mathrm{Var}(G_{ij})}{\mathbb{E}^2(G_{ij})} - \frac{2\,\mathrm{Cov}(R_{ij}, G_{ij})}{\mathbb{E}(R_{ij})\mathbb{E}(G_{ij})}\right)$$

From this estimator, we can determine the $100(1-\alpha)$ confidence interval for $\overline{M}_{ij}$ and use it to summarize the relative gene expression combined with its imprecision.

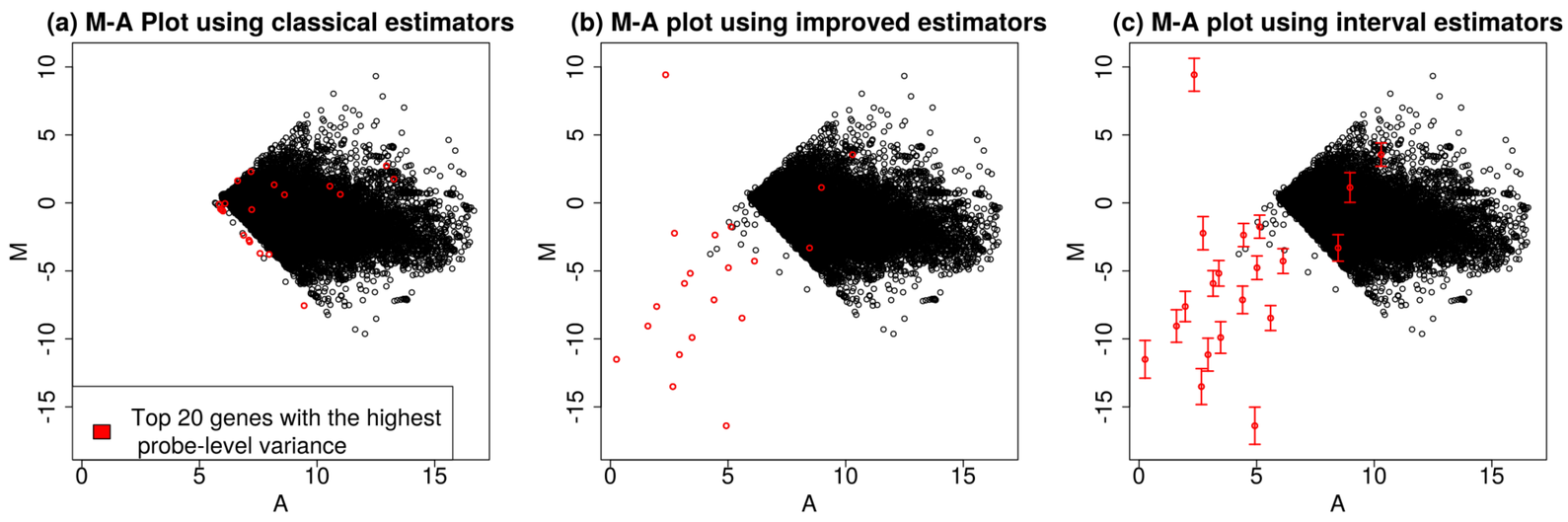
$$I_{M_{ij}} \doteq \left(\overline{M}_{ij} - \varepsilon_{ij}, \overline{M}_{ij} + \varepsilon_{ij}\right), \text{ where,}$$

- $\varepsilon_{ij} = t_{1-\alpha/2,N-1}\sqrt{\frac{\hat{\mathrm{Var}}(M_{ij})}{N}}$;
- $t_{1-\alpha/2,N-1}$ is the critical point of the t distribution with $N-1$ degrees of freedom; and
- N is the number of pixels of the probe.

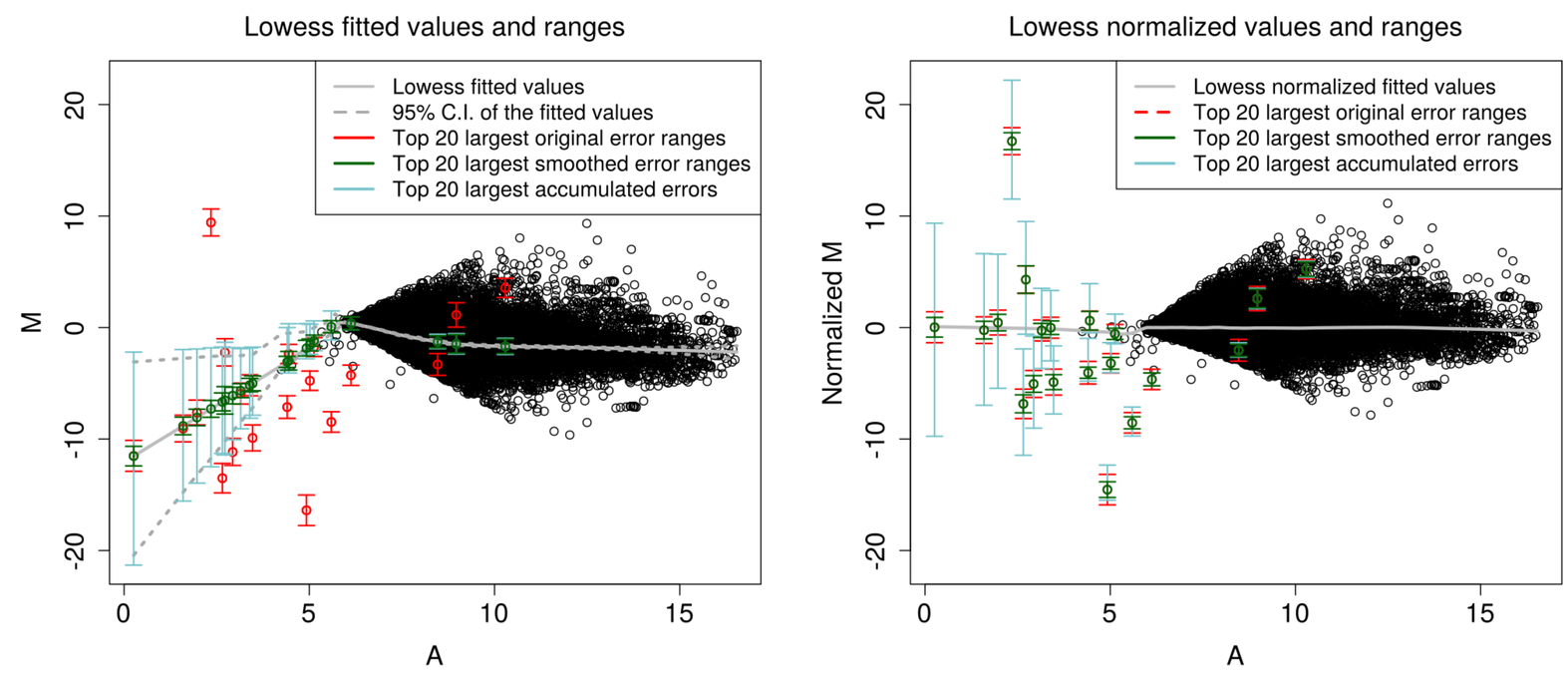
Application

We analyzed data from 90 two-color microarray slides, in which the same channel has been used for the common reference and the test samples are from 35 subjects with type II intestinal metaplasia and 55 healthy subjects. These experiments are part of the project [1].

Preprocessing Results

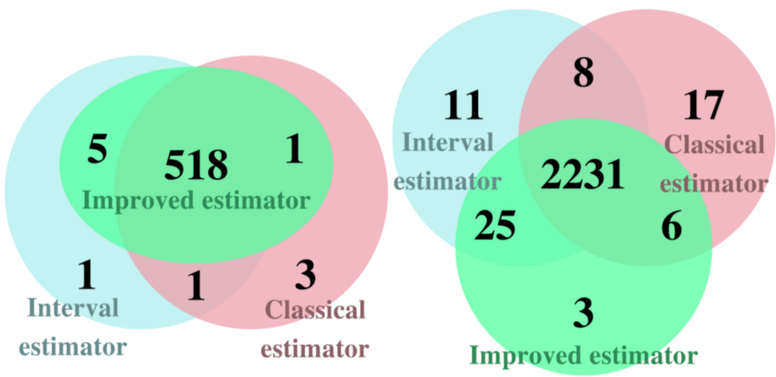


In order to extend the conventional microarray data analysis to deal with interval-data, we propose to adapt the LOWESS method with the techniques of [2] for regression on interval data in the normalization process.



Comparisons

For identifying differentially expressed genes, we conducted an ANOVA F-test after fit a linear model to the data. In the data interval approach, as the gene sample consists of the corresponding interval values of each slide, we fit a linear mixed-effects model using the slide as a random variable and heteroskedastic within-group errors. The variance function used maps different values for each slide. The replicates were obtained from the discretized intervals assuming the normal distribution.



FDR corrected p-value ≤ 0.001

FDR corrected p-value ≤ 0.05

Results

Identified differentially expressed genes which are in the 'Pathways in Cancer':

Interval Estimator	Improved Estimator	Classical Estimator
A_23_P155335 (PLD1)	A_23_P155335 (PLD1)	A_23_P155335 (PLD1)
p-valor: 6.504e-05 FC: 1.01	p-valor: 6.522e-05 FC: 1.01	p-valor: 6.416e-05 FC: 1.01
A_23_P1594 (VEGFB)	A_23_P1594 (VEGFB)	A_23_P1594 (VEGFB)
p-valor: 0.0002885 FC: -0.905	p-valor: 0.0002708 FC: -0.905	p-valor: 0.0002307 FC: -0.92
A_23_P502464 (NOS2)	A_23_P502464 (NOS2)	A_23_P502464 (NOS2)
p-valor: 0.0004816 FC: 1.39	p-valor: 0.0004949 FC: 1.39	p-valor: 0.0004867 FC: 1.39
A_24_P360674 (CDKN2B)	A_24_P360674 (CDKN2B)	A_24_P360674 (CDKN2B)
p-valor: 0.0005366 FC: 0.999	p-valor: 0.0005359 FC: 0.999	p-valor: 0.0005276 FC: 1

Identified differentially expressed genes which are associated with intestinal metaplasia of the stomach in the literature data:

Interval Estimator	Improved Estimator	Classical Estimator
A_23_P71017 (CLDN3)	A_23_P71017 (CLDN3)	A_23_P71017 (CLDN3)
p-valor: 3.311e-08 FC: 2.79	p-valor: 3.36e-08 FC: 2.79	p-valor: 3.402e-08 FC: 2.79
A_23_P256784 (MUC2)	A_23_P256784 (MUC2)	A_23_P256784 (MUC2)
p-valor: 9.754e-08 FC: 1.78	p-valor: 1.012e-07 FC: 1.78	p-valor: 9.624e-08 FC: 1.78
A_23_P58788 (CDX1)	A_23_P58788 (CDX1)	A_23_P58788 (CDX1)
p-valor: 5.456e-07 FC: 2.19	p-valor: 5.373e-07 FC: 2.19	p-valor: 5.404e-07 FC: 2.19
A_23_P76312 (GUCY2C)	A_23_P76312 (GUCY2C)	A_23_P76312 (GUCY2C)
p-valor: 1.774e-06 FC: 2.34	p-valor: 1.776e-06 FC: 2.34	p-valor: 1.79e-06 FC: 2.33
A_23_P111947 (CDH17)	A_23_P111947 (CDH17)	A_23_P111947 (CDH17)
p-valor: 2.25e-06 FC: 2.71	p-valor: 2.219e-06 FC: 2.71	p-valor: 2.231e-06 FC: 2.71
A_23_P76654 (CDX2)	A_23_P76654 (CDX2)	A_23_P76654 (CDX2)
p-valor: 2.593e-06 FC: 1.03	p-valor: 2.577e-06 FC: 1.03	p-valor: 2.663e-06 FC: 1.02
A_23_P112086 (DEFA5)	A_23_P112086 (DEFA5)	A_23_P112086 (DEFA5)
p-valor: 2.408e-05 FC: 3.35	p-valor: 2.411e-05 FC: 3.35	p-valor: 2.428e-05 FC: 3.35
A_24_P115183 (CLDN4)	A_24_P115183 (CLDN4)	A_24_P115183 (CLDN4)
p-valor: 0.0001195 FC: 1.22	p-valor: 0.0001182 FC: 1.22	p-valor: 0.0001189 FC: 1.22
A_24_P58673 (REG4)	A_24_P58673 (REG4)	A_24_P58673 (REG4)
p-valor: 0.0002818 FC: 2.53	p-valor: 0.0002792 FC: 2.53	p-valor: 0.0002759 FC: 2.53

Conclusion

We propose two improved estimation procedures for the relative expression of genes which consider the probe-level variances and covariances. The results showed that by including additional information from the imprecisions in the analysis it is possible to improve inferences and prevent misleading conclusions. Besides, the results were consistent with the literature data.

References

[1] Reis, Luiz Fernando Lima Expressão gênica em tumores do estômago e do esôfago: da biologia ao diagnóstico. 2008–2012. *Projeto Fapesp*.

[2] de Carvalho, Francisco de AT and Neto, Eufrasio de A Lima and Tenorio, Camilo P A new method to fit a linear regression model for interval-valued data. 2004 Springer

The source code and other related works are available at `http://goo.gl/7soGTR`