CONTROLE DO ERRO TIPO I EM UM EXPERIMENTO DE MICROARRAYS COM EUCALIPTO

Renato Nunes PEREIRA[1]
Júlio Sílvio de Sousa BUENO FILHO[1]

▪ RESUMO: Experimentos com microarrays visam analisar milhares de testes de hipóteses simultâneos para diferentes variáveis-respostas na mesma estrutura de unidades experimentais.Diversos procedimentos têm sido adotados para o cálculo de níveis de significância corrigidos, para o controle da taxa de erro tipo I, sendo um dos principais a "False Discovery Rate" (Benjamini e Hochberg, 1995). No âmbito do projeto Genolyptus, que busca identificar aspectos da genômica funcional do eucalipto, foram realizados ensaios com 21413 sondas.Destas, 1113 não atendiam à pressuposição de normalidade, segundo o teste de Shapiro-Wilks, necessitando de uma transformação não linear de dados. A transformação de Box-Cox corrigiu o problema da não normalidade para 801 dessas sondas, tendo sido detectada significância nos níveis corrigidos pela FDR para o efeito de tratamento em 12 destas 801 análises. Para as demais sondas que não necessitaram de transformação de dados, 40 revelaram significância para o efeito de tratamento, perfazendo um total de 52 sondas a serem investigadas para os contrastes de maior interesse. O uso do procedimento FDR e a transformação Box-Cox mostraram-se úteis em reduzir o número de ocorrência de resultados falsos positivos, tornando possível montar estudos posteriores sobre o potencial de expressão diferencial.

▪ PALAVRAS-CHAVE: Eucalipto; FDR; microarrays; transformação Box-Cox.

1 Introdução

Experimentos para a detecção de genes com potencial de expressão diferencial entre tecidos e órgãos para variáveis de importância econômica ou fisiológica podem ser realizados com o auxílio de microarrays (microarranjos) de DNA. Nos experimentos de microarrays são realizados milhares de testes de hipóteses

[1]Departamento de Ciências Exatas, Universidade Federal de Lavras – UFLA, Caixa Postal 3037, CEP: 37200-000, Lavras, MG, Brasil. E-mail: *rnpmoc@gmail.com / jssbueno@ufla.br*

simultâneos para vários modelos de análise de variância (ANAVA) para variáveis dependentes distintas, medidas na mesma unidade experimental.

Na análise de qualquer situação experimental, dois erros podem ser cometidos. O erro tipo I, ou falso positivo, e o erro tipo II (falso negativo). Em geral controla-se em níveis arbitrários o erro tipo I e planeja-se o experimento com poder suficiente para controlar o erro tipo II. Quando muitas hipóteses estão sendo testadas, a chance de cometer o erro tipo I aumenta rapidamente com o número de hipóteses, necessitando assim, de um procedimento de testes múltiplos que controle a taxa do erro tipo I. Para contornar este problema, uma estratégia que tem sido encontrada na literatura é a utilização de um procedimento proposto por Benjamini e Hochberg(1995), que controla a taxa de falsos positivos (FDR, do inglês *False discovery rate*). A FDR é definida como sendo a proporção de hipóteses nulas H_0 verdadeiras, entre as hipóteses nulas rejeitadas, ou seja, a proporção de erros devido à rejeição errônea de H_0 verdadeiras, também chamada proporção de falsos positivos.

Para melhor compreensão desse procedimento, considere-se que sejam testadas m hipóteses (H_0), das quais um determinado número m_0 seja verdadeira e que R das m hipóteses foram rejeitadas. Os dados da Tabela 1 resumem a situação apresentada. R e m são variáveis observáveis e U, V, S, W e m_0 são variáveis aleatórias não observáveis. Em termos dessas variáveis aleatórias, o nível de significância global, ou FWER, ou *familywise error rate*, é $P(V \geq 1)$.

Tabela 1 - Número de erros cometidos ao se tratarem m hipóteses

	Não rejeitadas	**Rejeitadas**	**Total**
H_0 *verdadeira*	U	V	m_0
H_0 *Falsas*	W	S	$m - m_0$
Total	$m - R$	R	m

A proporção de erros devido à falsa rejeição é dada pela variável aleatória $Q = \frac{V}{V+S} = \frac{V}{R}$; naturalmente, define-se $Q = 0$ quando $R = 0$. Q é uma variável aleatória não observável. Assim, define-se a FDR, Q_e; como sendo a esperança matemática de Q, isto é, $Q_e = E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{R}\right]$.

O procedimento para determinar o ponto de corte em testes múltiplos, controlando a FDR, pode ser realizado do seguinte modo (Benjamini e Hochberg, 1995): para cada uma das hipóteses a serem testadas $H_{0_1}, H_{0_2}, ..., H_{0_m}$, obter o valor da estatística teste e o correspondente *valor p* (probabilidade sob a hipótese H_0, de obter um valor ou igual ou superior ao obtido para estatística teste), $P_1, P_2, ..., P_m$. Em seguida, ordenar os valores P_i. Seja $P_{(1)}, P_{(2)}, ..., P_{(m)}$ os valores de P ordenados e H_{0_i} a hipótese correspondente. Definindo $q = \frac{mP_{(i)}}{i}$, a FDR pode ser controlada em um nível q^*, determinando-se o maior i, para o qual:

$$q^* \geq \frac{mP_{(i)}}{i}. \tag{1}$$

O procedimento FDR tem sido aplicado à análise de microarrays, para encontrar genes co-expressados (Tusher et al., 2001; Efron et al., 2001; Dudoit et al.,2003), além de outras aplicações (por exemplo, Keselman et al., 1999). Os microarrays são utilizados em experimentos para a detecção de genes com potencial de expressão diferencial entre tecidos e órgãos para variáveis de importância econômica ou fisiológica.

Em técnicas de análise em grande escala, como microarrays, em que a quantidade de dados gerada é muito grande, o tratamento e a análise dos dados são etapas muito trabalhosas e sujeitas a erros. Em ensaios de microarrays, a variável que se deseja analisar é a razão da hibridização entre duas amostras de cDNA que competem por um mesmo sítio, o que é obtido pela intensidade de fluorescência emitida por cada uma das amostras. O que torna a análise ainda mais complexa é a comparação de diversas amostras entre si, uma vez que os ensaios, neste caso, são feitos aos pares. Para tanto, há que se fazer um plano experimental cuidadoso, levando-se em consideração a disponibilidade de material e a precisão do resultado final.

A análise da variância (ANAVA), desenvolvida por Fisher entre 1920 e 1930 para análise e interpretação de dados experimentais, tem sido largamente utilizada para a identificação inicial de genes diferencialmente expressos (Speed, 2003). Para a utilização dessa técnica, algumas pressuposições básicas devem ser satisfeitas, entre elas a independência e a normalidade dos erros. O uso de tranformações não lineares é uma das possíveis formas de contornar o problema de dados que não obedecem aos pressupostos da análise de variância (Cox e Reid, 2000). Diversos têm sido os procedimentos de normalização adotados para microarrays, com especial atenção para procedimentos de normalização intra-slide (Speed, 2003). No presente trabalho o foco será a utilização de procedimentos básicos de normalização dos resíduos da ANAVA, com atenção especial na transformação de Box e Cox (1964), sem entrar na polêmica sobre a normalização intra-slides.

O exemplo em estudo é derivado do projeto Genolyptus, que busca identificar aspectos da genômica funcional do eucalipto. Neste projeto foi realizado um experimento de microarrays para detectar espécies, tecidos e órgãos com expressão diferencial para o teor de lignina, que é um caráter de importância tecnológica no aproveitamento do eucalipto, seja para a diminuição dos teores na polpa de celulose para a indústria de papel ou para o aumento dos teores em madeira de corte e energia.

O objetivo do presente trabalho foi verificar em que a medida transformação de dados e o uso da FDR permitem melhor interpretar os resultados deste experimento de microarrays acima citado.

2 Material e métodos

Os dados utilizados neste trabalho referem-se a estudos iniciais do genoma funcional do eucalipto (Genolyptus). As sondas de 21.413 cDNAs produzidas no

contexto do projeto foram montadas utilizando 10 microarrays monocromáticos da plataforma Nimblegen, com duas medidas na mesma lâmina.

O presente estudo de expressão gênica enfocou as diferenças entre espécies, entre clones dentro de espécies e diferenças entre indivíduos no mesmo tratamento.

Cada dado observado provém de nove réplicas da sonda dentro de um bloco, dentro do slide. O quadrado médio da fonte de variação slides dentro de tratamentos (t/s) foi considerado como a estimativa do erro experimental para a presença do efeito de tratamento, garantindo que as réplicas técnicas não inflam o erro, pois o erro experimental considerado foi o entre parcelas.

A estrutura experimental para cada uma das variáveis resposta é descrita na Tabela 2

Tabela 2 - Estrutura de fatores para o experimento com microarray

Bloco	Tratamento	Indivíduo	"Slide"	Espécie	Clone	"Ramet"	Tecido
1	1	1	1	*grandis*	1	1	Folha
1	1	1	2	*grandis*	1	2	Folha
1	2	1	3	*grandis*	1	1	Tronco
1	2	1	4	*grandis*	1	2	Tronco
1	3	2	5	*grandis*	2	3	Tronco
1	3	2	6	*grandis*	2	4	Tronco
1	4	3	7	*globulos*	3	5	Tronco
1	4	3	8	*globulos*	3	6	Tronco
1	5	4	9	*globulos*	4	7	Tronco
1	5	4	10	*globulos*	4	8	Tronco
2	1	1	1	*grandis*	1	1	Folha
2	1	1	2	*grandis*	1	2	Folha
2	2	1	3	*grandis*	1	1	Tronco
2	2	1	4	*grandis*	1	2	Tronco
2	3	2	5	*grandis*	2	3	Tronco
2	3	2	6	*grandis*	2	4	Tronco
2	4	3	7	*globulos*	3	5	Tronco
2	4	3	8	*globulos*	3	6	Tronco
2	5	4	9	*globulos*	4	7	Tronco
2	5	4	10	*globulos*	4	8	Tronco

grandis: Eucalyptus grandis
globulos: Eucalyptus globulos

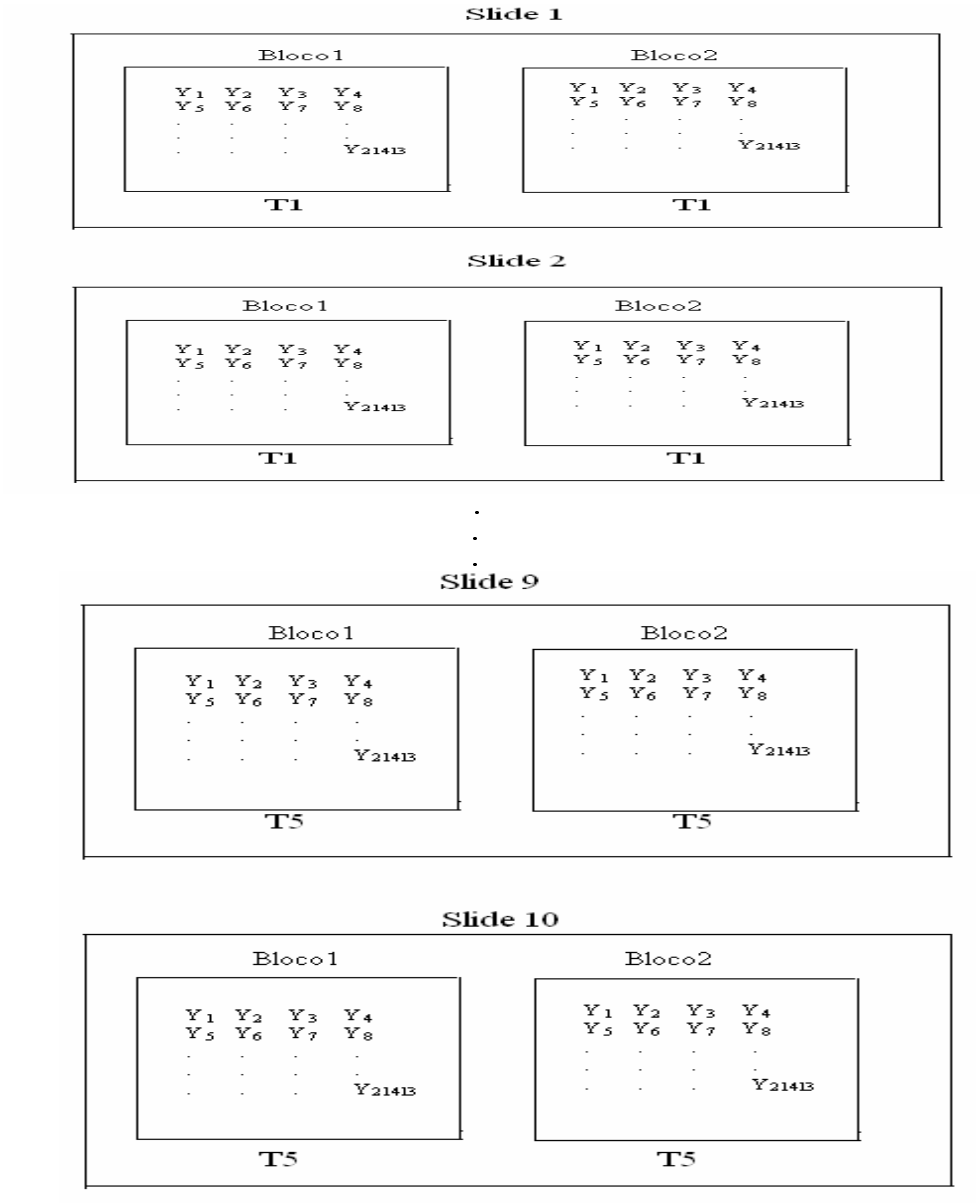

Figura 1 - Croqui da estrutura experimental.

2.1 Modelo de análise

Inicialmente foram feitas análises de variância para cada sonda, em que a hipótese de nulidade é que não havia efeitos diferenciais de tratamentos e posteriormente serão investigados alguns contrastes de interesse, que não são ortogonais. Os cinco tratamentos utilizados foram:
t_1: *Eucaliptus grandis*, clone 1, tecido de folha;
t_2: *Eucaliptus grandis*, clone 1, tecido de tronco;
t_3: *Eucaliptus grandis*, clone 2, tecido de tronco;
t_4: *Eucaliptus globulos*, clone 3,tecido de tronco;
t_5: *Eucaliptus globulos*, clone 4, tecido de tronco.

O modelo estatístico, para cada variável resposta, ou "sonda" (l) é:

$$y_{ijk(l)} = \mu(l) + t_{i(l)} + s_{j(il)} + e_{ijkl}, \tag{2}$$

em que, $y_{ijk(l)}$ representa as intensidades de expressão, na escala logarítmica.
μ é uma constante inerente a todas as observações;
$t_{i(l)}$ com i = 1,...,5 é o efeito fixo de tratamentos;
$s_{j(il)}$ com j = 1,2 é o efeito do slide j dentro do tratamento i;
e_{ijkl} com k = 1,2 é o erro experimental.

A Tabela 3 traz o esquema do quadro resumo da ANAVA e implica em que o teste F correto para a presença de efeito de tratamento é dado pela divisão do quadrado médio de tratamentos pelo quadrado médio de slides dentro de tratamentos (que é a estimativa do erro experimental no nível de slides).

Para as sondas que resultaram significativas para efeito de tratamento, foram investigados os seguintes contrastes de interesse:
Contraste1: Folha versus Tronco em *E.grandis*
$\lambda_1 = 2t_1 - t_2 - t_3$
Contraste2: *E.grandis* versus *E.globulos*.
$\lambda_2 = 2t_1 + 2t_2 + 2t_3 - 3t_4 - 3t_5$
Contraste3: Clones dentro de espécies
$\lambda_3 = t_1 + t_2 - 2t_3$
$\lambda_4 = t_4 - t_5$

Tabela 3 - Esquema da análise de variância com as componentes de variância

FV	G.L.	QM	E(QM)	F
T	I	Q_1	$(\sigma^2 + \mathrm{K}\sigma_s^2) + \mathrm{JK}\sigma_t^2$	Q_1/Q_2
T/S	I(J-1)	Q_2	$(\sigma^2 + \mathrm{K}\sigma_s^2)$	
Resíduo	IJ(k-1)	Q_3	σ^2	

Uma análise inicial foi feita para a resposta em cada sonda, considerando apenas a normalização original, que é o logaritmo da intensidade luminosa, para verificar possíveis violações nas pressuposições da ANAVA (ver por exemplo,

Montgomery, 1991). Sondas que falharam no teste de Shapiro e Wilk (1965) para a normalidade dos resíduos sofreram a transformação de Box e Cox (1964) e em seguida uma nova análise foi realizada. Para a transformação de Box e Cox (1964), considerou-se que λ pode variar no intervalo de $[-3, 3]$. Para automatizar a aproximação do valor de λ, este intervalo foi dividido em 100 partes iguais.

2.2 Aplicação do critério FDR

Para controlar os falsos positivos encontrados numa abordagem como esta, que envolve testes simultâneos em 21.413 sondas, foi feita a aplicação do procedimento FDR. Foi utilizado o critério FDR para conservar os níveis globais de significância 0.05 e 0.01.

O mesmo critério foi aplicado aos contrastes de interesse, nos casos em que foram constatadas diferenças significativas entre os tratamentos e para obter o valor p corrigido para cada sonda, utiliza-se a equação 1.

Os cálculos foram feitos utilizando-se o software R v.2.8.0 (R Development Core Team, 2008).

3 Resultados e discussão

Um dos resultados esperados com o uso da FDR é a redução no número de ocorrência de falsos positivos. Isto torna a interpretação mais compreensível e confiável, pois apenas serão realizados experimentos com sondas em que os indícios de expressão diferencial forem muito fortes. Para as sondas que mostraram significância para os efeitos diferenciais, foram analisados os contrastes de interesse. Um quadro geral do número de hipóteses rejeitadas está na tabela 4, que tem como objetivo mostrar que a análise trivial sem checar a pressuposição de normalidade leva a falsos positivos.

Tabela 4 - Número de hipóteses rejeitadas dentre as 21.413 hipóteses de nulidade testadas para os dados transformados e não transformados, considerando diferentes níveis de significância, com e sem a aplicação do FDR

	Dados Não Transformados		Dados Transformados	
	5%	1%	5%	1%
Sem FDR	6.060	2.547	5.518	2.627
Com FDR	973	59	1.529	79

Para os dados não transformados observa-se pela Tabela 4 que das 21.413 hipóteses de nulidade testadas, 6.060 foram rejeitadas, considerando um nível de significância $\alpha = 5\%$. Com a aplicação do procedimento FDR, esse número foi reduzido para 973. Como era de se esperar, houve um excessivo número de falsos positivos (84%).

Mesmo com a aplicação do procedimento FDR, o número de hipóteses rejeitadas continua alto para estudos mais específicos para cada sonda. Sendo

assim, foi aplicado novamente o procedimento, mas considerando-se um nível de significância conjunto $\alpha^* = 1\%$. Com esta nova consideração, para α^*, foram detectadas 59 sondas com efeito de tratamento. No entanto, 19 destas sondas, necessitaram de transformação de dados.

Para os dados transformados, observa-se que das 21.413 hipóteses de nulidade testadas para o efeito de tratamentos, 5.518 foram rejeitadas, a 5% de significância. Já para $\alpha = 1\%$, houve 2.627 sondas com resultado significativo. Com a aplicação do procedimento FDR, houve 1.529 rejeições com ($\alpha = 5\%$), apenas 79 foram também significativas com $\alpha = 1\%$.

3.1 Número de sondas combinando as duas análises

A análise completa para cada sonda seria inexequível de forma rotineira, por isto foi proposta a comparação das análises combinando dados não transformados e transformados para todas as sondas (o trabalho de interpretação é bem menor, embora o tempo computacional seja ligeiramente maior). Assim, a transformação foi aplicada por conveniência a todo o conjunto de dados e não apenas às variáveis que realmente precisavam, mas para concluir a respeito das sondas que interferem na expressão diferencial entre os tratamentos, foram combinadas as 40 sondas da análise sem transformação com aquelas dentre as 79 que na análise inicial não atendiam à pressuposição de normalidade.

Os resíduos da ANAVA de 1.113 sondas não atendiam à pressuposição de normalidade. A transformação de Box-Cox corrigiu o problema para 801 destas sondas, tendo em apenas 12 das 801 análises sido detectada significância para o efeito de tratamentos.

Em resumo, combinando as duas análise, tem-se um total de 52 sondas com expressão diferencial para tratamentos. O número de contrastes significativos para as 52 sondas que resultaram significativas para efeito de tratamento, são apresentados na Tabela 5.

Tabela 5 - Número de sondas com contrastes significativos, a 1% de significância, dentre as 52 com efeito de tratamento significativo

Contrastes	FDR(BH)
Folha x Tronco em *E. grandis*	33
E. grandis x E. globulos	43
Clones dentro de *E. grandis*	36
Clones dentro de *E. globulos*	13

Das 52 sondas que interferem na expressão diferencial entre tratamentos, 33 estão relacionadas a diferenças entre folhas e tronco, 43 à diferença interespecífica, 36 entre clones de *E. grandis* e 13 entre clones de *E. globulos*.

Combinando-se as diversas análises, observa-se, tanto nos dados transformados como nos não transformados, que, com a aplicação do procedimento FDR para corrigir a significância nos testes múltiplos, houve grande redução no número de

sondas em que se detecta diferença significativa para o efeito de tratamento. A redução do número de sondas mencionada anteriormente reforça a necessidade do uso do procedimento FDR, confirmando a presença de falsos positivos além de apresentar ordenadamente as sondas com indícios de expressão diferencial para que sejam investigadas em um próximo estágio da pesquisa.

Conclusões

O uso do critério FDR reduziu o número de variáveis em estudo, com segurança de se concentrar atenção nas 52 sondas com expressão diferencial mais relevante. O uso da transformação de Box e Cox (1964) auxiliou conferir validade para efeitos diferenciais de 12 dentre estas sondas.

Agradecimentos

À CAPES pela bolsa de mestrado para o primeiro autor e ao CNPq pela bolsa de produtividade em pesquisa para o segundo autor.

Ao professor Alexandre Siqueira Guedes Coelho da Universidade Federal de Goiás-UFG, pela gentileza de ceder os dados.

PEREIRA, R. N; BUENO FILHO, J. S. S. Type I error rate control in a microarrays experiment with eucalyptus. *Rev. Bras. Biom.*, São Paulo, v.27, n.4, p.350-359, 2009.

- *ABSTRACT: Microarray experiments deals with testing thousands of simultaneous hypotesys in the same experiment. To control experimentwise type one error rate many corrections on significance levels has been proposed, as is "False Discovery Rate" (Benjamini e Hochberg, 1995). In Genolyptus project, that aims for identifying aspects of the functional genomics of Eucalyptus, 21413 probes were tested in microarrays. Response from 1113 do not verify normality of residuals (according to Shapiro-Wilks test) and went on non-linear transformation to be analysed. Box-Cox corrected the problem for 801 of these probes and 12 out of it have shown significance for treatment effects. From the remaining probes that were not transformed, 40 has shown significance for treatment effects. Contrasts among five treatments for these 52 probes were investigated. FDR and Box-Cox transformation has shown usefull to reduce the number of treatments and made possible to devise further study on differential expression.*

- *KEYWORDS: Box-Cox transformation, eucalyptus, FDR, microarrays.*

Referências

BENJAMINI, Y.; HOCHBERG,Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, London, v.57, n.1, p.289-300, 1995.

BOX, G. E. P.; COX, D. R. An analysis of transformations. *J. R. Stat. Soc.*, London, v.26, n.2, p.42, 1964.

COX, D. R.; REID, N. *The theory of the design os experiments*. Boca Raton, FL: Chapman & Hall/CRC, 2000. p.323. (Monographs on statistics and applied probability, 86).

DUDOIT, S.; SHAFFER, J. C. Multiple hypothesis testing in microarrays experiments. *Stat. Sci.*, Hayward, v.18, n.1, p.71-103, 2003.

EFRON, B.; STOREY, J. D.; TIBSHIRANI, R. *Microarrays, empirical bayes methods, and false discovery rates*. Stanford: Stanford University, 2001, p.2001-2023. (Technical Report).

KESELMAN , H. J.; CRIBBIE ,R.; HOLLAND B. The pairwise multiple comparison multiplicity problem: an alternative approach to familywise and comparisonwise type I error control. *Psychol. Methods*, Washington, v.4, n.1, p.58-59, 1999.

MONTGOMERY, D. C. *Design and analysis of experiments*. 3.ed. New York : Wiley & Sons, 1991. 649p.

R DEVELOPMENT CORE TEAM. **R:** a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Disponível em: $< http://www.R-project.org >$. Acesso em: 20 dez. 2008.

SHAPIRO, S, S.; WILK, M.B. An analysis of variance for normality: complete sample. *Biometrika*, London, v.52, n.3/4, p.549-611, 1965.

SPEED, T. P. *Statistical analysis of gene expression microarray data*. Chapman & Hall/CRC. 2003. p.222.

TUSHER, V. G.; TIBSHIRANI, R.; CHU, G. Significance analysis of microarray applied to the ionizing radiation response.*Proc. Nat. Acad. Sci.*. Washington, v.98, n.18, p.5116-5121, 2001.

Recebido em 27.03.2009.

Aprovado após revisão em 01.10.2009.