UNIVERSIDADE ESTADUAL PAULISTA
FACULDADE DE MEDICINA VETERINÁRIA E ZOOTECNIA
Campus de Botucatu

APLICAÇÃO DE MODELOS LINEARES PARA ANÁLISE DE EXPRESSÃO GÊNICA EM EXPERIMENTOS DE MICROARRAYS

SAMIA RAMOS HADDAD

Dissertação apresentada ao Programa de Pós-Graduação em Zootecnia-Área de Concentração: Nutrição e Produção Animal, como parte das exigências para a obtenção do Título de Mestre em Zootecnia.

Botucatu – São Paulo
Janeiro - 2007

UNIVERSIDADE ESTADUAL PAULISTA
FACULDADE DE MEDICINA VETERINÁRIA E ZOOTECNIA
Campus de Botucatu

APLICAÇÃO DE MODELOS LINEARES PARA ANÁLISE DE EXPRESSÃO GÊNICA EM EXPERIMENTOS DE MICROARRAYS

SAMIA RAMOS HADDAD
Zootecnista

Orientador: Prof. Dr. Henrique Nunes de Oliveira

Dissertação apresentada ao Programa de Pós-Graduação em Zootecnia-Área de Concentração: Nutrição e Produção Animal, como parte das exigências para a obtenção do Título de Mestre em Zootecnia

Botucatu – São Paulo
Janeiro – 2007

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉCNICA DE AQUISIÇÃO E TRATAMENTO DA INFORMAÇÃO – SERVIÇO TÉCNICO DE BIBLIOTECA E DOCUMENTAÇÃO
UNESP - FCA - LAGEADO - BOTUCATU (SP)

Haddad, Samia Ramos, 1978-
H126a Aplicação de modelos lineares para análise de expressão gênica em experimentos de microarrays / Samia Ramos Haddad. – Botucatu : [s.n.], 2007.
iv, 49 f. : gráfs., tabs.

Dissertação (mestrado) -Universidade Estadual Paulista, Faculdade de Medicina Veterinária e Zootecnia, Botucatu, 2007
Orientador: Henrique Nunes de Oliveira
Inclui bibliografia

1. Genética molecular. 2. Modelos lineares (Estatística). 3. Acido desoxirribonucleico . 4. Análise de variância. I. Oliveira, Henrique Nunes. II. Universidade Estadual Pau- lista "Júlio de Mesquita Filho" (Campus de Botucatu). Fa- culdade de Medicina Veterinária e Zootecnia. III. Título.

Dedico...

Aos meus queridos e amados pais Neusa e Marum.

Aos meus amados irmãos Renata e Raphael e também aos irmãos de coração Wilber, Adriana Bonito e Adriana Abrahão

Ao André que com muito amor, atenção e paciência tornam os meus dias mais alegres e felizes.

Ao esperado sobrinho (a) que com certeza trará muita alegria e será muito querido (a) e amado (a).

Aos meus queridos tios e tias (Mauri, Eduarda, Madrinha, Padrinho , Zi e Suely).

Aos meus saudosos e amados avós: João e Júlia; Magid e Aurora.

Aos meus queridos primos: Jr., Alexandre, Grace, Dani, Reinaldo, Bá, Li e aos pequenos: Júlia, Melissa, Karina, Guilherme, Gustavo e Igor.

A minha linda e amada afilhada Carol.

Ao Sr.Nagib e D.Sonia pelo carinho e zelo.

Aos meus verdadeiros amigos: Carla Tini (Sarna); Carolina Bulhões (Tchonga); Maria Carolina (Cinira); Cíntia Nakayama (Pirralha); Gustavo Igaki (Tempurá) e Cristina; Willian (Ovo); Marcos Paulo Benedette (Baguá); Lili; Ana Carolina Lage; Dani Lara; Maurício Rocha; Maíra (Xinchila); Lílian Pulz; Karina Sanches; Vanessa (Xepa); Ana Paula Semenssato(USP); Adriana Almeida (USP); Clarissa (Novilha); Ana Cristina.

Aos amigos especiais: Cilamara, Sr. Ayrton, Luciana e Jr.; Tinha e Marlene;

Ao Dr. Paulo Marchiori que me acompanha com muito carinho transmitindo grandes ensinamentos e tornando minha vida melhor.

A Deus por me presentear diariamente com conquistas, alegrias, sabedorias e por me colocar em contato com muitas pessoas especiais

Agradecimentos

Ao Professor Henrique Nunes de Oliveira pela orientação, dedicação, paciência e pelo convívio de todos esses anos.

Ao Guilherme J. M. Rosa por todo o profissionalismo, amizade, incentivo à pesquisa e grande dedicação e preocupação pelo trabalho.

Ao Professor José Nicolau P. Puoli Filho pelas conversas, ensinamentos, experiência e dedicação.

Aos Professores do Departamento de Nutrição e Melhoramento Animal.

Aos Professores do Departamento de Produção Animal da FMVZ.

Aos Professores do Instituto de Biociência da Unesp de Botucatu.

Ao saudoso Dino pela amizade e profissionalismo.

Aos colegas da Pós-Graduação da FMVZ/UNESP-Botucatu.

Ao André Rodrigues Abrahão pela experiência, paciência e dedicação.

Ao Professor Raysildo Lôbo e toda a equipe do bloco C da Genética da USP pelo acolhimento e dedicação na fase final.

A equipe do Centro de Química de Proteína (Hemocentro- USP- Ribeirão Preto) por todo o apoio e pelos valiosos ensinamentos.

À Seila e Carmem por todo profissionalismo, amizade e dedicação.

A todos os funcionários da FMVZ/UNESP-Botucatu.

Ao CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pela concessão da bolsa de estudo.

SUMÁRIO

CAPÍTULO 1
CONSIDERAÇÕES INICIAIS

A área de pesquisa conhecida atualmente como biologia molecular se iniciou a partir do modelo estrutural do DNA proposto por Francis Harry Compton Crick e por James Watson em 1953.

Até pouco tempo, a análise da expressão gênica era feita com metodologias que avaliavam poucos genes de cada vez: *Northern Blots*, *dot blots*, *RT-PCR*, entre outros. O crescimento exponencial do número de genes descobertos nos projetos de genomas e o desenvolvimento de arranjos de DNA propiciaram uma nova abordagem nos estudos da regulação gênica, tornando possível o monitoramento simultâneo de níveis de transcritos de um grande número de genes (Felix *et al*., 2002).

A análise da Expressão Gênica por *Microarray* foi apontada como uma técnica de biologia molecular com grandes promessas para auxiliar geneticistas a explorarem e entenderem o genoma (Kerr and Churchill, 2001b) e por isso se tornou uma ferramenta muito comum nos laboratórios (Wolfinger *et al*., 2001). O princípio da técnica de *Microarray* baseia-se na habilidade da molécula de RNA mensageiro (mRNA) ligar-se especificamente ou hibridizar ao DNA molde do qual foi originado. Com um arranjo (*array)* contendo milhares de amostras de DNA, pode-se determinar, em um único experimento, o nível de expressão de milhares de genes, medindo a quantidade de mRNA ligado em cada sítio do *array*. A quantidade de mRNA ligado aos *spots* (estruturas que contém a seqüência complementar para a característica) é diretamente proporcional à expressão gênica e por isso, quantificando-se o mRNA, estima-se a expressão gênica e assim, é gerado um perfil de expressão gênica do tecido estudado.

Esta técnica permite a comparação de expressão gênica entre diferentes tecidos em um mesmo organismo (cérebro vs. fígado); mesmo tecido em um mesmo organismo (tratamento vs. controle, tumor vs. não tumor); mesmo tecido em organismos diferentes (selvagem vs. mutante); experimentos temporais (diferentes estágios de desenvolvimentos) entre outros. Essa versatilidade gerada possibilita uma importante ampliação do conhecimento básico da função gênica (Villa, 2004).

Os arranjos *Affymetrix* empregam seqüências curtas de DNA (cadeia curta) que são sintetizadas na própria superfície sólida em que deverá ocorrer a hibridação, utilizando um processo de síntese química em áreas fotoativadas. No caso dos arranjos de cadeia longa de DNA, as seqüências sintetizadas em laboratório são transferidas para uma pequena superfície (*slide*) em um arranjo de sítios de teste miniaturizados, em que é realizada a hibridação com as amostras de RNA extraído dos tecidos de interesse. Duas amostras são analisadas por vez sendo, cada uma, corada com um corante fluorescente vermelho ou verde (Cy3 ou Cy5).

1.1 Delineamento de experimentos de microarray

Para que o objetivo de um experimento de *microarray* seja atingido é necessário que seja realizado um delineamento experimental adequado e de acordo com o interesse do pesquisador. O delineamento adotado deve eliminar o confundimento nas fontes de variação e evidenciar a variabilidade biológica (ROSA *et al.*, 2005b). A importância do delineamento está também diretamente relacionada com os altos custos dos ensaios, sendo fundamental que seja feito uso eficiente dos recursos disponíveis.

Quando se estudam dois grupos experimentais, uma alternativa natural para a distribuição das amostras nos slides é ter uma amostra de cada grupo representado em cada *slide*. Algumas variações adicionais devem ser consideradas como, a troca das colorações (vermelho e verde) e a união das amostras.

Para comparar duas amostras, Kerr e Churchill (2001a) esquematizaram dois exemplos hipotéticos mostrando métodos de comparação direta e indireta. A comparação indireta ocorre quando se usa amostras referências, em que as variáveis de interesse são comparadas com essa referência. Esse delineamento é dito prorrogável, pois, uma nova variável de interesse pode ser adicionada ao decorrer do experimento, por outro lado, a maior quantidade de dados é coletada da amostra referência do que propriamente a de interesse, além disso, a variável é completamente confundida com o *dye,* uma vez que cada amostra só precisa ser marcada com uma coloração. A comparação

direta segundo os autores fornece mais informações dos contrastes específicos entre duas amostras. As estimativas da variância para um contraste específico resultam a partir da combinação de todas as comparações diretas e indiretas ligadas às duas amostras no experimento (Rosa *et al.*, 2005a). É importante ressaltar que a comparação direta entre duas amostras é possível somente se as variedades estiverem representadas dentro do mesmo bloco, e a indireta pode ser comparada usando a referência, mesmo estando em blocos diferentes.

Yang e Speed (2002) representaram graficamente os experimentos de *microarray,* ilustrando bem um delineamento referência com duas amostras; delineamento em *"loop"* com três amostras e delineamentos com replicações biológicas. Steibel *et al* (2005) também ilustraram os delineamentos de experimento de *microarray* de duas cores com a troca de coloração (*Dye-swap*) e replicação biológica.

1.2 Análise Estatística de *Microarray*

A decisão de analisar os dados na escala logaritima é baseada em muitas considerações. Esse recurso permite que os dados continuem sendo analisados por meio da ANOVA (Análise de Variância), cujas exigências são que tenham distribuição normal. A variável transformada por logaritmo é a maneira natural para análise dos dados com modelos aditivos quando se acredita que os efeitos nos dados sejam multiplicativos. Assim, as comparações dos dados não transformados com outras metodologias de transformações, demonstram que a logarítmica apresenta melhor resposta (Sapir e Churchill, 2000).

Em geral, a análise quantitativa de dados de *microarray* engloba dois estágios. O primeiro é denominado *normalização*, a qual se refere a um pré-processamento dos dados para a minimização de variações sistemáticas nas medidas de expressão gênica. O segundo estágio refere-se à *análise estatística* propriamente dita, com os testes de significância. Vários procedimentos têm sido propostos para a normalização dos dados e para a determinação de diferenças significativas entre medidas de expressão gênica obtidas de diferentes grupos experimentais.

1.2.1 Normalização dos dados

Os dados a serem analisados nos experimentos de *microarray* são as medidas da intensidade dos sinais das imagens dos slides hibridizados, que são realizadas utilizando-se softwares especializados. As intensidades dos sinais representam a quantidade de DNA fluorescente ligado ao slide, e apresentam, em cada *spot,* variações decorrentes de uma série de fatores não sistemáticos, tais como os defeitos de fabricação que afetam a qualidade do slide. As fontes de variação sistemática intrínsecas à amostra ou não, afetam as medidas do nível de expressão gênica em cada slide (Yang *et al*. 2002). No caso dos experimentos de DNA de cadeia longa, a remoção das fontes de variação que causam diferenças no nível geral de expressão entre os dois corantes (*dyes*) é denominada normalização. Segundo Yang et al. (2002) mesmo quando se utiliza em um slide, a mesma amostra de RNA para os dois corantes, a média da medida da expressão dos dois corantes será diferente. As causas destas diferenças incluem as propriedades físicas dos corantes, eficiência nas técnicas de coloração e processamento da amostra, diferenças na potência dos *lasers* utilizados na leitura dos resultados no slide; diferenças na quantidade de RNA, na distribuição das amostras no slide, entre outras (Dudoit *et al.*, 2002; Guérette, 2001).

Assim, utiliza-se a normalização para balancear as intensidades de fluorescência para as duas colorações (Cy3 (verde) e Cy5 (vermelho)), o que permitirá uma comparação do nível de expressão dos genes entre os slides de um mesmo experimento (normalização de locação). Além disto, pode haver diferenças nos níveis de expressão entre os diferentes slides. Estas diferenças refletiriam na variação do logaritmo da razão das medidas de expressão dos dois corantes (log-ratio) e os resultados de um slide poderiam influir decisivamente no resultado de todo o experimento. Para corrigir esta variação é necessária uma normalização de escala (Yang et al. 2002). Utilizando-se da normalização, as diferenças biológicas e a comparação dos níveis de expressão por meio dos slides de *microarray*, podem ser distinguidas com maior facilidade.

O processo de normalização pode ser realizado tomando-se por base todos os genes dos slides, ou apenas genes que (se acredita) tenham expressão constante, ou ainda genes controle (genes de espécies diferentes da estudada e cujo mRNA é incluído em quantidades iguais nas duas amostras). A escolha do conjunto de genes utilizado para normalização depende do tipo de experimento. Quando se espera um grande número de genes com expressão diferencial, então a escolha deve recair sobre a segunda ou terceira opções. No presente trabalho, apenas a primeira abordagem será considerada.

Segundo Guérette (2001) e Yang *et al.* (2002) a normalização pode ser feita de três maneiras, dependendo do tipo de experimento:

1) Normalização dentro do slide
2) Normalização em slides pareados
3) Normalização entre slides.

A Normalização dentro do slide é feita em cada slide em separado e pode envolver todos os genes do slide, ou ser realizada por regiões do slide. Pode ser ainda para correção de locação ou escala.

Normalização Global

Este tipo de normalização é realizada para correção de locação. Neste caso, supõe-se que as intensidades das medidas dos corantes vermelho e verde têm uma relação constante (Cy5=k.Cy3). Assim, a correção é feita subtraindo-se do logaritmo da razão das medidas de expressão dos dois corantes (log-ratio) de cada spot, a média destes para todo o slide:

$$Log_2(Cy5/Cy3)_i = Log_2(Cy5/Cy3)_i - \mu = Log_2(Cy5/k.Cy3)_i$$ (Yang et al. (2002))

Desta forma, o centro da distribuição de todos os logaritmos da razão das amostras é zero. Este método foi o primeiro a ser utilizado para normalização, e tem a desvantagem de não considerar os efeitos da distribuição espacial dos spots nos slides na expressão dos genes e por considerar que as diferenças são independentes da intensidade de expressão.

Normalização dependente da intensidade de expressão

Na maioria dos experimentos de microarray, aparecem vícios dependentes da intensidade de expressão na distribuição dos log-ratios. Isto pode ser notado quando se observa o gráfico da distribuição dos log-ratio em função da intensidade da expressão. Estes gráficos são conhecidos como gráficos de M por A ou simplesmente MA. Nesta sigla, M representa o log-ratio (M=*Log_2 (Cy5/Cy3)*) e A representa a média da medida da expressão dos dois corantes $\left(A = Log_2\sqrt{Cy5.Cy3} = (Log_2(Cy5) + Log_2(cy3))/2\right)$.

Havendo esta dependência, a normalização dependente da intensidade de expressão que é indicada. Para realizar esta normalização utiliza-se um método não paramétrico para regressão de M em A. Este método, conhecido como Lo(w)ess (Local-Weighted Regression and Smoothing Scatterplot) adapta uma curva linear de regressão ponderada, considerando para cada ponto, apenas os dados na vizinhança do mesmo. A ponderação é dada em função da distância de cada dado em relação ao ponto. A partir da regressão, a correção é feita como:

$Log_2(Cy5/Cy3)_i = Log_2(Cy5/Cy3)_i - b(A) = Log_2(Cy5/k(A).Cy3)_i$ (Yang *et al*., 2002)

E os valores de Cy5 e Cy3 ajustados podem ser obtidos como:

$Log_2(Cy5)_i = A + Log_2(Cy5/Cy3)_i / 2$ e $= Log_2(Cy3)_i = A - Log_2(Cy5/Cy3)_i$

Normalização dentro de grupo de impressão

O processo de depositar os oligonucleotídeos de cDNA nos respectivos pontos (*spots*) nas placas (*slides*) é chamado de impressão. O processo mecânico é realizado por um robô que utiliza um conjunto de agulhas para depositar os oligonucleotídeos de acordo com os grupos de impressão. Este conjunto de agulhas pode ser uma fonte de variação sistemática de variação no nível de expressão medido. Quando isto ocorre, a normalização pode ser realizada dentro de grupo de impressão. Este tipo de normalização é semelhante ao anterior sendo, no entanto, realizado por grupo de impressão.

Normalização dentro de slide para escala

Uma vez que os dados estejam normalizados quanto à locação os log-ratios de cada grupo de impressão, esses estarão centrados em zero, mas é possível que a variação dos log-ratios dentro de cada grupo de impressão seja diferente. Nestes casos é necessária uma correção da escala. Uma maneira de realizar esta normalização, segundo Yang et al. (2002), é admitir a média como valor zero e a variância $o_i^2\sigma^2$, em que σ^2 é a variância da razão logarítmica e o_i^2 é o fator da escala para o *i*-éssimo grupo. Uma estimativa de o_i^2 é necessária para proceder a normalização. Com a restrição $\sum_{i=1}^{I} \log o_i^2 = 0$; sendo *I* o total de grupos de impressão no *array*, a estimativa de máxima verossimilhança é dada por:

$$\hat{o}_i = \frac{\sum_{j=1}^{ni} M_{ij}^2}{\sqrt[1]{II_{k=1}^{I} \sum_{j=1}^{ni} M_{kj}^2}}$$

Em que M_{ij} denota a o j-éssimo log-ratio do i-éssimo grupo de impressão, j=1,...., n_i. Uma alternativa robusta é estimar:

$$\hat{o}_i = \frac{MADi}{\sqrt[1]{II_{i=1}^{I} MAD_i}}$$

Em que a mediana absoluta da variação (*MAD*) é definida por :

$$MAD_i = mediana_j \mid M_{ij} - mediana_j(M_{ij}) \mid$$

Normalização em slides pareados

Este tipo de normalização é usado quando se invertem os corantes (*dyes)* e se repetem os slides *(dye-swap*).

A relação logarítmica para o primeiro slide é definida por $\log_2$ (Cy5/Cy3)-*c* e para o segundo slide definida por $\log_2$ (Cy5'/Cy3')-*c'*, em que *c* e *c'* denotam as funções de normalização para os dois slides. Uma vez que os corantes estão trocados nos dois slides, espera-se que os log-ratios dos dois slides sejam de magnitude semelhante e sinais opostos.

$$\frac{1}{2}\left[\log_2(Cy5/Cy3) - c - \left(\log_2(Cy5'/Cy3' - c')\right)\right] \approx$$

$$\frac{1}{2}\left[\log_2(Cy5/Cy3) + \log_2(Cy5'/Cy3')\right] = \frac{1}{2}\log_2\left(\frac{Cy5Cy3'}{Cy3Cy5'}\right) = \frac{1}{2}\left(M - M'\right)$$

A principal suposição é que *c* é semelhante a *c'* e esse método pode ser aplicado para todos os genes, até mesmo se eles estiverem sendo expressos diferentemente. Na prática, utiliza-se a regressão não paramétrica (*loess*) fazendo-se $\frac{1}{2}(M - M')$ *vs* $\frac{1}{2}(A - A')$ em lugar de M vs A.

Normalização em múltiplos slides

Esta normalização é utilizada para permitir a comparação entre os slides do experimento. É aplicada após a normalização dentro do slide para correção dos efeitos de escala entre slides. O mesmo método utilizado para a correção de escala dentro de slide, pode ser diretamente estendido para esta situação. .

1.2.2 Aplicação de Modelos lineares para análise de experimentos de microarray

Após a normalização, os dados resultantes podem ser analisados com o objetivo de identificar aqueles que se expressam de maneira diferencial com relação às amostras testadas. Para que possamos chegar a este resultado é necessário inicialmente que sejam identificadas as fontes de variação que possam interferir nas medidas de expressão gênica após a normalização. Num experimento típico de *microarray*, as análises precisam considerar os efeitos de Corante (*Dye* - Vermelho/Verde); Slide (*Array)*; Gene e Interação entre estes fatores além da Variabilidade Biológica das amostras que seria o fator de

interesse principal. Esses múltiplos fatores foram considerados simultaneamente nas análises realizadas nos trabalhos iniciais, seguindo o modelo proposto por Kerr e Churchill (2000):

$$y_{ijkgr} = \mu + \mathrm{A}_i + D_j + (AD)_{ij} + G_g + (AG)_{ig} + (DG)_{ig} + T_k + (TG)_{kg} + \varepsilon_{ijkgr} ,$$

em que:

y_{ijkg} representa a expressão da intensidade em escala logarítmica , μ é a média, A_i representa o efeito do i-éssimo array, D_j representa o efeito j-éssimo *Dye* (corante), $(AD)_{ij}$ representa a variação total nos *arrays* e coloração (*dyes*), G_g representa o efeito g-éssimo Gene, $(AG)_{ig}$ representa a interação do *array i* e o gene *g,* $(DG)_{ig}$; representa os efeitos de coloração específicos de genes, T_k representa o efeito de tratamento ou variedade; $(TG)_{kg}$ representa a interação do tratamento ou variedade *t* e o gene *g.* Supôs-se que os erros ε_{ijkg} seriam independentes e identicamente distribuídos com média zero e variância σ^2 .

Os efeitos do *array* A_i foram estimados para levar em consideração possíveis diferenças entre as médias de todos os *spots*, *dyes* e os tratamentos presentes nos mesmos, podendo ocorrer devido às condições particulares de cada slide na confecção ou no processo de hibridação. Similares, os efeitos do D_j representam diferenças entre as médias devido às diferenças causadas pelos efeitos dos corantes. O efeito $(AD)_{ij}$ considera os efeitos particulares dos corantes em cada slide, ou seja, as diferenças entre as médias dos corantes em todos os genes daquele slide. Embora a normalização corrija para estes efeitos dentro de slide, quando são considerados conjuntamente com outros efeitos, os corantes podem continuar agindo como fonte de variação. O efeito de tratamento (T_k) considera as diferenças gerais entre as amostras nos dois corantes e em todos os genes e slides do experimento. As diferenças podem aparecer se algumas variedades tiverem mais atividade de transcrição, ou, simplesmente se o diferencial de concentração de mRNA é maior em uma das amostras. A variável G_g considera a média dos efeitos de cada gene em todos os slides do experimento.

A interação $(AG)_{ig}$ (efeito do *spot)* considera o efeito da média do *gene g* no *array i*. O efeito pode aparecer porque não há controle completo sobre a quantidade e concentração de cDNA imobilizado no *spot*. Todos esses efeitos não são geralmente de interesse, mas precisam ser considerados como fonte

de variação. O efeito de interesse nesse modelo linear com efeitos fixos, é a interação entre a variedade e o gene, $(TG)_{kg}$, ou seja, saber quanto cada tratamento influencia na expressão de cada gene em particular.

O maior problema com esta maneira de analisar os resultados é a suposição de homocedasticidade implícita no modelo de análise. Como o nível de expressão pode variar bastante entre os genes, já que supostamente um slide com milhares de genes deve conter uma boa parcela deles que absolutamente não se expressariam em nenhuma das amostras e outros em que a atividade de transcrição deveria estar em níveis elevados, torna-se difícil aceitar que a variância seja igual para todos os genes.

Posteriormente, Wolfinger *et al*.,(2001) propuseram a realização de análises em dois estágios como forma para contornar o problema da heterogeneidade de variância dos genes. Essa metodologia permitiu análises individuais para cada gene com sua respectiva variância.

Modelos mistos

A primeira se refere a “Normalização Global”, em que o modelo é aplicado em todo o conjunto de dados.

$$y_{ijkgr} = \mu + \mathrm{A}_i + D_j + \left(AD\right)_{ij} + e_{ijkgr} ,$$

Sendo que:

y_{ijkg} representa a expressão da intensidade em escala logarítmica, μ é a média, A_i representa o efeito i-éssimo *array*, D_j representa o efeito j-éssimo *Dye* (corante), $(AD)_{ij}$ representa a variação total nos *arrays* e coloração (*dyes*). A partir da normalização global, a variação dos resíduos é analisada em um segundo modelo, em cada gene:

$$\hat{e}_{ijkgr} = \mu_k + A_{ik} + D_{jk} + T_{gk} + \varepsilon_{ijkgr} \quad ,$$

Sendo:

μ_k a média constante do gene específico, A_{ik} o efeito do *array*, D_{jk} efeito do *dye* e T_{gk} efeito do tratamento ou variedade. É importante ressaltar que a análise dos dados não deve contemplar apenas a replicação biológica e efeitos aleatórios no modelo, mas tem grande importância a definição da unidade

experimental e a escolha de um modelo condizente com o objetivo do trabalho e que permita a comparação adequada dos tratamentos propostos (Wernish *et al.,* 2003). Em experimentos em que cada gene é representado em muitos *spots* em cada *array*, um termo adicional é necessário para modelar esse nível extra de replicação, que é, na realidade, uma replicação técnica e não biológica.

Os modelos estatísticos para dados de *microarray* devem ser coerentes com o delineamento experimental. Rosa *et al*. (2005a) chamam a atenção para o fato de que várias publicações recentes fazem uso de modelos de análise de variância para comparação de dados de expressão gênica em experimentos de *microarray*, sem, entretanto, fazer distinção, de forma adequada entre os diferentes tipos de replicação nestes experimentos. Este fato pode levar a erros no cálculo do tamanho das amostras necessárias para análise e no poder dos testes estatísticos aplicados para comparação das médias dos grupos. Os autores consideram que a definição adequada da unidade experimental e a correta distinção entre replicações técnicas e biológicas são cruciais para a validade das inferências do experimento. O Ideal é que se tenham múltiplas amostras biológicas independentes representando cada condição analisada, mas se houver apenas replicações técnicas, ainda será possível realizar testes estatísticos, embora a interpretação seja limitada (Cui e Churchill, 2003).

Entretanto, qualquer que seja o delineamento experimental e o modelo de análise utilizado em um estudo deste tipo, existe ainda uma série de problemas a serem resolvidos. Testes estatísticos para comparação conjunta de médias em milhares de genes, a falta de independência entre os genes testados, e o fato de as medidas de expressão gênica não seguirem necessariamente uma distribuição normal, mesmo depois de transformadas, são algumas das dificuldades existentes.

Considerando o problema dos testes múltiplos, uma das dificuldades das análises assim realizadas é controlar o nível de significância nos testes de média, para detectar diferenças na expressão dos genes associados a uma ou mais variáveis independentes e as possíveis ocorrências de erros *tipo I e II* (Allison e Coffey, 2002). É natural que em testes de hipóteses ocorram erros do *tipo I* (rejeitar H_0 quando esta é verdadeira) ou erro *tipo II* (aceitar H_0 quando esta é falsa). Cada tipo de erro tem uma probabilidade de ocorrer, sendo α para

o erro tipo I e β para o erro tipo II. Com a alteração das probabilidades de erros, o pesquisador tem grande chance de tomar decisões erradas. (Silva e Ferreira, 2002).

À medida que aumenta o número de testes realizados, a probabilidade de ocorrência de erros do tipo I aumenta. Está bem estabelecido que é preciso controlar este erro de alguma maneira. Uma opção é controlar a taxa global de erro, utilizando a correção de Bonferroni. Essa metodologia consiste em determinar o nível de significância (α*) de todo conjunto de genes analisados a partir do nível de significância individual (α). Para a determinação desses α* é necessário realizar testes "*t*" independentes (Silva e Vencovsky, 2002). Bland e Altman (1995) relataram um exemplo que explica didaticamente a metodologia.

Numa comparação de dois tratamentos dentro de cinco subgrupos de pacientes num ensaio clínico, os tratamentos serão significativamente diferentes ao nível de 0,05 se houver um valor de "P" menor que 0,01 dentro de qualquer um dos subgrupos. Embora esta metodologia seja útil para controlar a ocorrência de erros do tipo I, os níveis de erros do tipo II ficam elevados, diminuindo o poder do teste e dificultando a detecção de diferenças na expressão dos genes. Benjamin & Hoccheberg (1995) tentando amenizar este e outros problemas encontrados nos testes de comparações múltiplas propuseram a razão de falsas descobertas (FDR) também chamada de proporção de falsos positivos.

O FDR permite controlar a ocorrência de erros do tipo I, sem aumentar muito a probabilidade de erros do tipo II. Essa metodologia pode ser aplicada aos valores de "P" (probabilidade de erro) gerados nos testes individuais das hipóteses dos dados de expressão gênica num experimento de *microarray* (Benjamini *et. a*l; 2005).

Miller *et. al*. (2001) propôs uma metodologia que controla os erros tipo I com um procedimento de dois estágios (*two-stage*). Em um ensaio em que se testam *n* hipóteses, um número de genes *m* terá efeitos significantes ($0 \leq m \leq n$) (1° estágio). A partir desse fato, um segundo conjunto de dados independentes é reunido e somente os *m* genes que foram significativos no 1° estágio são testados com um outro nível de significância α (2° estágio). Os testes desse segundo estágio teriam uma maior sensibilidade.

Cui e Churchill (2003) realizaram uma extensa revisão sobre a utilização de testes estatísticos para determinar diferenças na expressão gênica em experimentos de microarray, indo desde o puro e simples estabelecimento de um ponto arbitrário a partir do qual as diferenças são consideradas importantes até testes “F” usando estimativas combinadas de erro. Os autores consideram que um dos maiores problemas na aplicação dos testes estatísticos é a instabilidade da estimação da variância residual, uma vez que se, ao acaso, a estimativa da variância residual é pequena, as chances dos testes revelarem-se significativos é bem mais alta. Várias alternativas foram propostas para resolver este problema, desde a soma de uma constante de pequeno valor à variância residual até a combinação de informações da variância de cada gene com a informação da variância de todos os genes. Além disto, os autores propuseram um teste estatístico em que a média da variância residual calculada para todos os genes e a variância usada para cada gene individual era utilizada para cada gene.

Com relação ao problema da diferença de variabilidade entre os genes, Smyth (2004) propôs uma abordagem bayesiana empírica que seria equivalente a uma regressão da variância residual de cada gene, no sentido de uma média conjunta das variâncias de todos os genes, o que segundo o autor resultaria em valores mais estáveis de variância quando o número de slides é pequeno.

Recentemente Cui *et al*. (2005) propôs uma metodologia que também regride as estimativas das variâncias residuais em relação à média, utilizando um estimador baseado em Lindley (1962). Neste caso, não é feita nenhuma suposição a priori para a distribuição da variância entre os genes. Assumindo que todas as variâncias são iguais e usando-se um estimador comum, é possível aumentar substancialmente o poder para detecção da diferença de expressão gênica.

As fórmulas abaixo demonstram e ilustram como se estabelece o estimador da variância conjunta (5).

$$X_g \sim \sigma_g^2 \chi_v^2 \qquad (1)$$

$$\ln \frac{X_g}{v} \sim \ln \sigma_g^2 + \ln \frac{X_v^2}{v} \qquad (2)$$

$$X'_g \sim \ln \sigma_g^2 + \in'_g \qquad (3)$$

$$\overline{X}' + \left(1 - \frac{(G-3)v}{\sum (X'_g - \overline{X}')^2}\right)_+ x(X'_g - \overline{X}') \qquad (4)$$

$$\tilde{\sigma}_g^2 = \left(\prod_{g=1}^{G} (X_g / v)^{1/G}\right) Bx \exp\left[\left(1 - \frac{(G-3)V}{\sum (\ln X_g - \overline{\ln X_g})^2}\right)_+ x(\ln X_g - \overline{\ln X_g})\right], \qquad (5)$$

em que,

Xg é a soma dos quadrados dos resíduos; σ^2_g é a verdadeira variância de cada gene e supõe-se que Xg/σ^2_g são independentes e cada um apresenta uma distribuição de Qui-quadrado com *v* graus de liberdade (1). Em (2) obtém-se a transformação em logaritmo natural; em (5) a transformação voltou para a escala original e obtém-se o estimador para σ^2_g.

Abaixo estão demonstradas as fórmulas para a construção do teste F por meio do estimador da variância conjunta. O modelo (6) é o modelo misto geral, já o (9) é um modelo misto para dados de *microarray* que é ajustado para cada gene e a matriz representada em (7) é a mesma para todos os genes. Em (10) Δ_g representa a soma de quadrados de tratamentos, sendo o F_S o teste proposto.

$$Y = X\beta + Zu + \in \qquad (6) \qquad \hat{C} = X \begin{bmatrix} X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\ Z'\hat{R}^{-1}X & Z'\hat{R}^{-1}Z + \hat{G}^{-1} \end{bmatrix}^{-} \qquad (7)$$

$$F = \frac{\hat{\beta}'L'(L'\hat{C}L)^{-1}L\hat{\beta}}{rank(L)} \qquad (8) \qquad Y_g = X\beta_g + Zu_g + \in_g \qquad (9)$$

$$F_S = \Delta_g \Big/ \tilde{\sigma}_g^2 \qquad (10)$$

REFERÊNCIAS BIBLIOGRÁFICAS

ALLISON D.B.; COFFEY C.S. Two-Stage Testing in Microarray Analysis: What is Gained? J**. GERNTOL. A BIOL. SCI. MED. SCI.**, MAY 1, v.57 (5), p. B189-192; 2002.

BENJAMINI, Y; KENIGSBERG, E.; REINER, A.; YEKUTIELI, D. FDR adjustments of Microarray Experiments; 2005.

BENJAMINI, Y;Hocheberg, Y. Controling the False Discovery Rate: a Pratical and Powerful Approach to multipleTesting. **Journal of the Royal Statistics Society B,** v. 57, p.289-300, 1995

BLAND, J.M.; ALTMAN D.G. Multiple significance tests: the Bonferroni method. **BMJ**, v.310, p.170; 1995.

CUI, X.; HWANG,J.T.G.; QIU, J.; BLADES, N.J.; CHURCHILL, G.A. Improved statistical tests for differential gene expression by shrinking variance components estimates. **Biostatistics**, v.6, 1, pp.59-75, 2005.

CUI, X. e CHURCHILL, G. A. Statistical tests for differential expression in cDNA microarray experiments. **Genome Biology.** V.4. (4):210, 2003.

DUDOIT, S.; YANG, Y.H.; CALLOW, M. J.; SPEED, T.P.. Statistical methods for identifying differentially expressed gene in replicated cDNA microarray experiments. **Statistica Sinica**, v. 12, n° 1, p.111-139, 2002.

FELIX, J.M.; DRUMMOND, R.D.; NOGUEIRA, F.T.S.; ROSA, V.E.; JORGE, R.A.; ARRUDA, P.; MONOSSI, M. Genoma funcional: Uso de arranjos de DNA em náilon para a análise da expressão gênica em larga escala**. Biotecnologia Ciência & Desenvolvimento**, nº 24, p. 60-67, janeiro/fevereiro 2002**.**

GUÉRETTE, A. Normalization Techniques for cDNA Microarray Data. Disponível em: http://www.cs.mcgill.ca/~aguere/308-761B/norm/summary_normalization.htm. Acessado em: 17/03/2004.

KERR, M. K. e CHURCHILL, G. A. Experimental design for gene expression *microarrays.* **Biostatistics,** v. 2; p. 183-201. 2001a.

KERR, M. K. e CHURCHILL, G. A. Statistical design and the analysis of gene expression *microarray* data*.* **Genet. Res.**, v. 77; p. 123-128. 2001b.

KERR, M. K.; MARTIN, M.; CHURCHILL, G. A. Analysis of variance for gene expression *microarray* data. **J. Comp. Biol.,** v. 7; p. 819-837, 2000.

LINDLEY, D. V. Discussion of Professor Stein's paper, 'Confidence sets for the mean of a multivariate normal distribution'. ***Journal of the Royal Statistical Society Series B* 24**, 265–296, 1962.

MILLER RA, Galecki A, Shmookler-Reis RJ. Interpretation, design, and analysis of gene array expression experiments. **J Gerontol A-Biol**; 56: B52-B57, 2001.

ROSA, G.J.M.; STEIBEL, J.P. ; TEMPELMAN, R. J. Reassessing design and analysis of two-color microarray experiments using mixed effects models. **Comparative and Functional Genomics**, v. 6; p.123-131, 2005 a.

ROSA, F. H. F. P. ; SOLER, J.M. . Perspectivas em análise de dados de microarrays. In: I Simpósio de Iniciação Científica e Pós-Graduação do IME-USP, 2005b, São Paulo. Anais do I Simpósio de Iniciação Científica e Pós-Graduação do IME-USP. São Paulo : Instituto de Matemática e Estatística, v.1., 2005b.

SAPIR, M.; CHURCHILL, G.A. Estimating the Posterior Probability of Gene Expression from Microarray Data, 2000. Disponível em: http://www.jax.org/research/churchill/pubs/marina.pdf.

SILVA, B.V.; FERREIRA, D.F.Comparação da robustez de alternativas do teste de igualdade de duas médias populacionais sob não normalidade por simulação Monte Carlo. **Acta Scientiarum**, v.24, n.6,p.1771-1776, 2002.

SILVA H.D. ; VENCOVSKY R. Poder de detecção de "Quantitative Trait Loci", da análise de marcas simples e da regressão linear múltipla. **Scientia agric.** (Piracicaba, Braz) v.59 n.4, p.755-762, 2002.

SMYTH, G.K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. **Statistical Applications in Genetics and Molecular Biology**. V.3. Issue 1.Article 3. 2004.

STEIBEL, J. P. ; TEMPELMAN R. J.; ROSA, G.J.M. Power and sample size determinations for two color microarray experiments based on different levels of replication. *Submetido à publicação.*

VILLA, L.L. Biologia molecular: conceitos e princípios básicos. Disponível em: http://www.cervicolp.com.br/atualizacao/biologiaLuisa.htm. Acessado em: 20/04/2004.

WERNISH, L.; KENDALL, S. L.; SONEJI, S.; WIETZORREK, A.; PARRISH, T.; HINDS, J.; BUTCHER, P. D.; STOCKER, N. G. Analysis of whole-genome *microarray* replicates using mixed models. **Bioinformatics,** v. 19; p. 53-61, 2003.

WOLFINGER, R. D.; GIBSON, G.; WOLFINGER, E. D.; BENNETT, L.; HAMADEH, H.; BUSHEL, P.; AFSHARI, C.; PAULES, R. S. Assessing gene significance from cDNA *microarray* expression data via mixed models. **J. Comp. Biol**.,v. 8; p. 625-637, 2001.

YANG, Y. H. e SPEED, T. Design issue for cDNA *microarray* experiments. **Nat. Rev. Genet.,** v. 3; p. 579-588, 2002.

YANG, Y. H.; BUCKLEY, M. J.; DUDOIT, S.; SPEED, T. P. Comparison of methods for image analysis on cDNA microarray data, **Journal of computational and graphical statistics**, v.11, n°1, p.108-136, 2002.

APLICAÇÃO DE MODELOS LINEARES PARA ANÁLISE DE EXPRESSÃO GÊNICA EM EXPERIMENTOS DE MICROARRAYS

Resumo

O presente trabalho objetivou comparar, utilizando dados de um experimento de *Microarray* com um delineamento simples, os resultados de diferentes testes estatísticos a fim de verificar suas características na detecção de diferenças no nível de expressão dos genes.

Os dados foram provenientes da South Dakota State University-EUA, do Department of Biology and Microbiology, Department of Animal Science, onde toda a parte experimental foi realizada. O material biológico envolveu quatro aves infectadas e quatro não infectadas com o vírus de bronquite infecciosa (IBV). O RNA utilizado foi extraído da camada epitelial da traquéia de animais controle e infectados com o vírus da IBV e, após a transcrição reversa foi marcado com os corantes fluorescentes (Cy3 e Cy5) e hibridizados com o microarray 13k cDNA de aves (FHCRC, Seattle, WA).

A análise de dados dos resultados do experimento de *microarray* englobou dois estágios, sendo o primeiro denominado de Normalização, em que os dados foram pré-processados utilizando o procedimento Loess. A seguir foram realizadas as análises estatísticas propriamente ditas com testes de significância. Utilizou-se um modelo simples de ANOVA e aplicaram-se diferentes metodologias de análise.

A análise das imagens revelou que dos 16192 *spots* em cada slide, apenas 10.926 puderam ser lidos sem defeitos no primeiro slide, 11.633 no segundo slide, 12577 no terceiro e 13.154 no quarto slide. A grande maioria dos *spots* em branco e controles negativos apresentou defeitos que determinaram sua eliminação. Um total de 13.597 *spots* foi lido no conjunto dos quatro slides, mas apenas 9.853 *spots* estavam representados em todos os slides.

Concluiu-se que os experimentos de *microarray*, por tratarem de um conjunto muito grande de observações a serem analisados requerem análises

estatísticas específicas. O método de Cui *et al.* (2005) reduziu a dependência entre a variância residual e o valor de probabilidade do teste F aplicado de acordo com o modelo utilizado, e desta forma, mostra-se adequado para este fim. Todavia, não foi observado neste trabalho aumento do poder de teste.

Palavras-chave: Modelos Mistos; Testes Estatísticos; cDNA; Oligonucleotídeos; Biologia Molecular; Genes

Application of Linear Models for analysis of gene expression in Microarray Experiments

Abstract

The aim of this research was to compare, using real data of an experiment of Microarray with a simple design, the results of different statistical tests in order to verify their characteristics in the detection of differences in the level of expression of the genes.

The data were coming of South Dakota State University-EUA, of the Department of Biology and Microbiology, Department Animal of Science, where the whole experimental part was accomplished. The biological material involved four infected animals and four no infected with the virus of infectious bronchitis (IBV). Used RNA was extracted of the layer epitelial of the windpipe of animals control and infected with the virus of IBV and, after the reverse transcription it was marked with the fluorescent colors (Cy3 and Cy5) and hybridization with the microarray 13k cDNA of birds (FHCRC, Seattle, WA).

The analysis of data of the results of the microarray experiment included two apprenticeships, being the first denominated of Normalization, in that the data were pre-processed using the procedure Loess. To follow the statistical analyses they were accomplished properly said through real data with significant tests. A simple model of ANOVA was used and different analysis methodologies were applied.

The analysis of the images revealed that of the 16192 spots in each slide, only 10.926 could be read without defects in the first slide, 11.633 in the second slide, 12577 in the third slide and 13.154 in the fourth slide. The great majority of the spots in white and negative controls presented defects that determined it elimination. A total of 13.597 spots was read in the group of the four slides, but only 9.853 spots were represented in all of the slides.

It was ended that the microarray experiments, for they treat of a very big group of observations to be analyzed request specific

statistical analyses. The method of Cui et al. (2005) it reduced the dependence between the residual variance and the value of probability of the test applied F in agreement with the used model, and this way, it is shown appropriate for this end. Though, it was not observed in this research increase of the test power.

Key-words: Mixed models; Statistical tests; cDNA; Oligonucleotide; Molecular biology; Genes

INTRODUÇÃO

Milhares de genes em um dado organismo vivo funcionam de modo complexo interagindo entre si, porém a maioria das técnicas em biologia molecular trabalha com um pequeno número de genes em cada situação. Estudos baseados em técnicas que permitem a análise detalhada da expressão dos genes por meio da quantificação dos RNA mensageiros (mRNAs), podem ser muito úteis para o conhecimento e exploração das diferenças entre os organismos. Nos últimos anos, a técnica de análise de expressão gênica utilizando arranjos de DNA em placas (mais conhecida pelo nome em inglês: *Microarray)* tem se expandido devido à facilidade proporcionada a diversos tipos de pesquisa (Mah *et al*., 2004). Essa tecnologia torna possível o monitoramento da expressão de milhares de genes simultaneamente (Schena *et al*., 1995; Lockhardt *et al*., 1996).

Atualmente existem várias plataformas de *Microarray,* que empregam princípios comuns em suas técnicas, entre elas, estão os oligonucleotídeos de cadeia curta e o cDNA que são oligonucleotídeos de cadeia longa.

A técnica com oligonucleotídeos de cadeia curta, conhecida pela marca de seu principal fabricante (*Affymetrix),* emprega seqüências pequenas de DNA que são sintetizadas diretamente na superfície sólida onde ocorre posteriormente a hibridação, e as amostras a serem utilizadas neste processo são marcadas com um único corante fluorescente. A técnica que emprega os oligonucleotídeos de cadeia longa (cDNA), marcados com dupla coloração fluorescente e conhecida como *Microarray* de duas cores é a mais utilizada na maioria dos experimentos com animais domésticos.

Nos experimentos de *Microarray* de duas cores, as amostras contrastadas são identificadas com diferentes colorações (*dyes*). Uma amostra é marcada com verde e a outra com vermelho (Kerr e Churchill, 2001). As amostras são misturadas e lavadas nos slides (*arrays*) onde ocorre o processo de hibridização, em que cada cDNA (DNA complementar) liga-se com o seu molde específico. Assim, a intensidade de fluorescência gerada, permite uma análise detalhada da expressão dos genes.

Juntamente com a aplicação dessa tecnologia, surgiram diversos delineamentos experimentais e desafios para as análises estatísticas (Steibel

et al., 2005). A análise de dados de *Microarray* passou por uma evolução, em que os primeiros trabalhos utilizavam à técnica de Análise de Variância num modelo único, considerando homogeneidade de variância entre todos os genes (Kerr *et al.*, 2000) e cada *spot* como unidade experimental. Posteriormente, Wolfinger *et al.* (2001) sugeriram que a análise fosse realizada utilizando a metodologia dos modelos mistos em dois estágios, em que o primeiro consiste na normalização dos dados, e o segundo, em análises individuais para cada gene, o que levaria em conta a heterogeneidade de variância da medida.

Um dos problemas desta análise refere-se ao teste estatístico utilizado para determinar em quais genes está ocorrendo expressão diferencial. A análise de cada gene não pode ser considerada independentemente das análises de outros genes sob pena de perder-se o controle sobre o erro do tipo I, ou seja, a quantidade de falsos negativos. Adotado um valor do nível de significância individual (α), a probabilidade de que ocorra pelo menos um falso positivo é igual a $(1- \alpha^n)$, em que n é o número de genes analisados.

Devido à simplicidade de cálculo e interpretação, a correção de Bonferroni, que consiste em determinar o valor do nível de significância individual (α*) e que proporciona o nível de significância do conjunto (α) desejado (Silva e Vencovsky, 2002), tem sido bastante utilizado. O problema da correção de Bonferroni é que é um teste extremamente conservativo e, portanto, apresenta um poder de detecção muito baixo. Outras abordagens foram propostas por Benjamin & Hoccheberg (1995) que sugeriram controlar a proporção de falsos positivos (FDR), e a metodologia de dois estágios de Miller *et al* (2001). Outro problema que se observa é que existe uma instabilidade da estimação da variância residual, e se, ao acaso, a estimativa da variância residual é pequena, as chances dos testes revelarem-se significativos é bem alta.

Cui *et al.* (2005) propuseram um método que objetivou a regressão das estimativas dos componentes de variância residual de cada gene, na média geométrica destas estimativas Nesse trabalho, todos os diferentes testes F propostos foram detalhados. F_1 testa os efeitos fixos para cada gene (gene-específico) e é, portanto usado quando os componentes de variância são de genes específicos; o teste F_2 utiliza a estimativa de variância da média dos genes e a variância agrupada para cada componente; F_3 aplica o estimador

agrupado de variância para cada componente de variância e esse é bastante utilizado quando todos os componentes de variância são constantes nos genes; e o teste F_S utiliza o estimador comum do componente de variância para cada gene obtido e calcula o teste estatístico F. Esse teste F_S é eficaz quando há informações limitadas para estimar os componentes de variância de cada gene específico. Embora todas as variâncias sejam iguais, a metodologia aplicada usando o estimador comum, aumenta o poder para detectar a diferença existente na expressão dos genes. Ainda nesse trabalho, Cui *et al.* (2005) fizeram um teste de simulação comparando os diferentes testes F, cujos resultados obtidos, levaram a conclusão que o teste F_S proposto é o mais robusto e poderoso do que os outros testes estudados.

Com uma iniciativa pioneira na pesquisa brasileira, o objetivo do presente trabalho foi aplicar algumas técnicas disponíveis na literatura para análise de um experimento de microarray de delineamento simples e com pequeno número de observações, com especial interesse no método de Cui *et al*. (2005), para verificar seu efeito sobre o poder do teste.

MATERIAL E MÉTODOS

O conjunto de dados utilizados no trabalho foram provenientes da South Dakota State University-EUA, do *Department of Biology and Microbiology, Department of Animal Science*, onde toda a parte experimental foi realizada.

O material biológico envolveu quatro aves infectadas e quatro não infectadas com o vírus de bronquite infecciosa (IBV), com um total de 10.000 genes analisados aproximadamente.

IBV é um protótipo da família corona vírus que causa uma doença aguda nas vias respiratórias em aves. A doença, como muitas outras, pode ser prevenida por vacinação com o vírus vivo atenuado, administrado diretamente na mucosa. Estudos recentes têm mostrado que a memória imune de longa duração é detectada nas células do tecido das mucosas quando a vacinação é realizada no local e não quando a imunização é sistêmica. Isso sugere que a interação direta dos tecidos da mucosa com os componentes do patógeno é essencial para a indução da imunidade local. Assim, o estudo foi realizado para avaliar a cinética da transcrição molecular local durante o desenvolvimento da imunidade na mucosa utilizando-se o IBV como um modelo.

O RNA utilizado foi extraído da camada epitelial da traquéia de animais controle e infectados com o vírus da IBV e, após a transcrição reversa foi marcado com os corantes fluorescentes verde (Cy3) e vermelho (Cy5) e hibridizados com o microarray 13k cDNA de aves (FHCRC, Seattle, WA). Cada slide era composto por 32 blocos, cada um consistindo de uma matriz com 22 x 23 *spots*, totalizando 16.192 *spots*. Destes, aproximadamente 700 *spots* eram ocupados por genes controle ou eram vazios.

O delineamento experimental envolveu dois tratamentos (quatro amostras de animais infectados e quatro de animais sadios). As amostras foram distribuídas em dois pares de slides com troca de corantes (*dye-swap*). Cy3 e Cy5 são as típicas colorações usadas no experimento, sendo respectivamente verde e vermelho. Para cada um destes slides, a amostra de um animal contaminado era marcada com o corante verde ou vermelho e a amostra de um animal sadio com a outra cor (TABELA 1). Após a hibridização, a imagem gerada foi interpretada utilizando-se o software Genepix (Axon Instruments..., 2001).

Tabela 1 – Distribuição das amostras de animais contaminados (Cont.) e Sadios (Trat) por slide e corante (Cy3_verde e Cy5_vermelho).

Slide N°	Cy3 _verde	Cy5 _vermelho
1	TRAT	CONT
2	TRAT	CONT
3	CONT	TRAT
4	CONT	TRAT

Removeram-se os dados em que os indicadores de defeitos nos *spots* ("*flags*") do software analisador de imagem informaram que havia problemas, assim como os dados de *spots* com valores de circularidade inferiores a 70. Esses valores representam uma deformidade dos "*spots*" que dificultam a leitura das intensidades de fluorescência dos "*dyes*". Em uma comparação, os valores de 100 para a circularidade representam uma circunferência de leitura precisa dos *spots*.

Foram realizados os processos para a normalização dos dados dentro de slide conforme descrito por Dudoit *et al.* (2002). Inicialmente obteve-se a média (A) do logaritmo na base dois da medida da intensidade dos dois corantes em cada spot e o logaritmo da razão entre eles (M). A seguir, utilizando-se o procedimento *Loess* do pacote estatístico SAS, ajustou-se o modelo $\hat{M} = f(A)$ para cada um dos quatro slides do experimento. O logaritmo da razão (Cy3/Cy5) normalizado ($\overline{M}$) foi então calculado como $\overline{M} = M - \hat{M}$. Optou-se no presente trabalho por trabalhar com cada observação em isolado (e não com o logaritmo da razão), de forma que foi preciso obter o valor de cada observação como desvio da média: $\overline{C}y3 = \left(A + \overline{M}\right)/2$ e $\overline{C}y5 = \left(A - \overline{M}\right)/2$. A normalização realizada balanceou as intensidades de fluorescência para o Cy3 e Cy5 e permitiu a comparação do nível de expressão dos genes entre os slides.

Após a normalização dentro de slides, optou-se pela análise em dois estágios sugerida por Wolfinger *et al.* (2001) O primeiro estágio se refere à normalização global, em que o modelo [1] é aplicado em todo o conjunto de dados.

$$y_{ijkgr} = \mu + \mathrm{A}_i + D_j + (AD)_{ij} + e_{ijkgr}\ , \qquad [1]$$

Sendo que:

y_{ijkg} representa a expressão da intensidade em escala logarítmica, μ é a média, A_i representa o efeito i-éssimo *array* (*i=1,4*), D_j representa o efeito j-éssimo *dye* (corante)*(j=1,2)*, $(AD)_{ij}$ representa o efeito da interação *dye* X *array* que está confundido com o dos tratamentos. A partir da normalização global, os resíduos "$\hat{e}_{ijkgr}$ " foram submetidos a um segundo modelo [2], utilizado para análises de cada gene individualmente:

$$\hat{e}_{ijkgr} = \mu_k + A_{ik} + D_{jk} + T_{gk} + \varepsilon_{ijkgr} \quad , \qquad [2]$$

Sendo:

μ_k a média constante do gene específico, A_{ik} o efeito especifico do *array* sobre o gene *(i=1,4)*, D_{jk} efeito do *dye sobre o* gene *(i=1,2)* e T_{gk} efeito do tratamento *(g=1,2) no gene*. O efeito de *array* foi considerado aleatório e os demais fixos. O efeito de interesse foi o de tratamento. A partir do teste F foram obtidos os valores de probabilidade (valor_p) alcançados para cada gene individualmente.

A partir das somas de quadrados do resíduo para cada um dos genes (X_g) obteve-se, utilizando a metodologia descrita por Cui *et al.* (2005), o estimador $\tilde{\sigma}_g^2$ regredido para a média geométrica da soma de quadrados de todos os genes, aplicando-se a seguinte fórmula:

$$\tilde{\sigma}_g^2 = \left(\prod_{g=1}^{G} (X_g / v)^{1/G} \right) Bx \exp\left[\left(1 - \frac{(G-3)V}{\sum (\ln X_g - \overline{\ln X_g})^2} \right)_+ x(\ln X_g - \overline{\ln X_g}) \right]$$

em que, lnΣ= é a correção para viés; G é o número de genes estudados, e *V* é a variância do logarítimo da divisão de uma variável aleatória com distribuição qui-quadrado e graus de liberdade igual a v, por v subtraída da média; e v são os graus de liberdade associados a soma de quadrados do resíduo. Os valores de *B* e *V* foram obtidos em Cui et al. (2005).

Obtidas as estimativas, foram realizados os testes $F_S = \Delta_g / \tilde{\sigma}_g^2$ em que Δ_g é o quadrado médio do tratamento. Cui *et al.* (2005) sugeriram que os valores críticos de F fossem obtidos por permutação, mas devido ao pequeno número de slides utilizados, não foi possível realizar tal procedimento.

Utilizando os valores de probabilidade (valor_p) obtidos nas duas análises foram aplicados os testes conjuntos de Razão de falsas descobertas ou Proporção de falsos positivos e Bonferroni.

Foi usado o pacote estatístico SAS (SAS,1999) para análise dos dados.

RESULTADOS E DISCUSSÂO

A análise das imagens revelou que dos 16192 *spots* em cada slide, apenas 10.926 puderam ser lidos sem defeitos no primeiro slide, 11.633 no segundo slide, 12577 no terceiro e 13.154 no quarto slide. A grande maioria dos *spots* em branco e controles negativos apresentou defeitos que determinaram sua eliminação. Estes defeitos incluíram problemas na forma do *spot* (circularidade) e reflexão da área em torno do *spot* maior que a da área de sinal, entre outros. Um total de 13.597 *spots* foi lido no conjunto dos quatro slides, mas apenas 9.853 *spots* estavam representados em todos os slides.

1-Normalização

A Figura 1 e 2 ilustram o processo de normalização dos dados para o primeiro slide, mostrando a distribuição do logaritmo da razão das intensidades medidas nos dois corantes em função da intensidade média. A Figura 1 representa a distribuição dos logaritmos das razões das duas amostras do primeiro slide (log(cy3/Cy5)), em função da expressão média das duas amostras ((log(Cy3) + Log(cy5))2) antes da normalização. A linha cheia representa o valor predito pela regressão *Loess*. Pode-se notar claramente que existiu uma distorção na razão da expressão dos genes em função da média

de expressão. Depois da normalização pode-se verificar que a distorção foi corrigida (Figura 2).

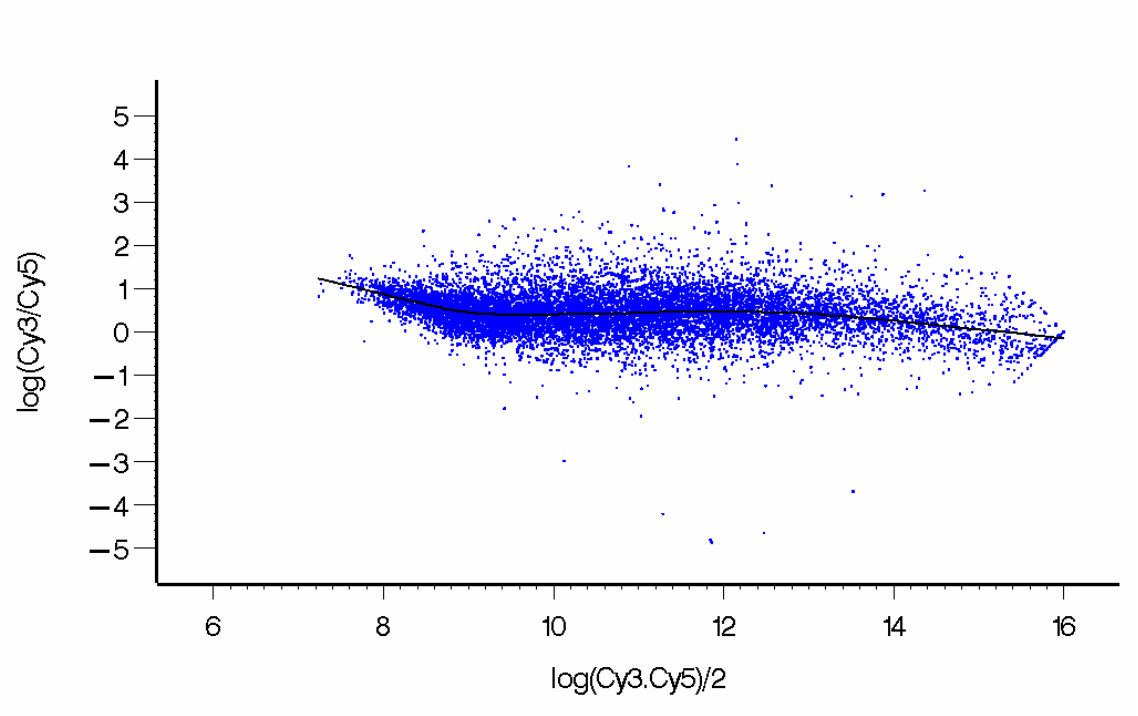

Figura 1: Distribuição do logaritmo da razão de intensidade da expressão medida pelos dois canais (relativos a cada corante) em função da média da intensidade de expressão no slide antes da normalização. A linha cheia representa o valor predito da regressão de Loess.

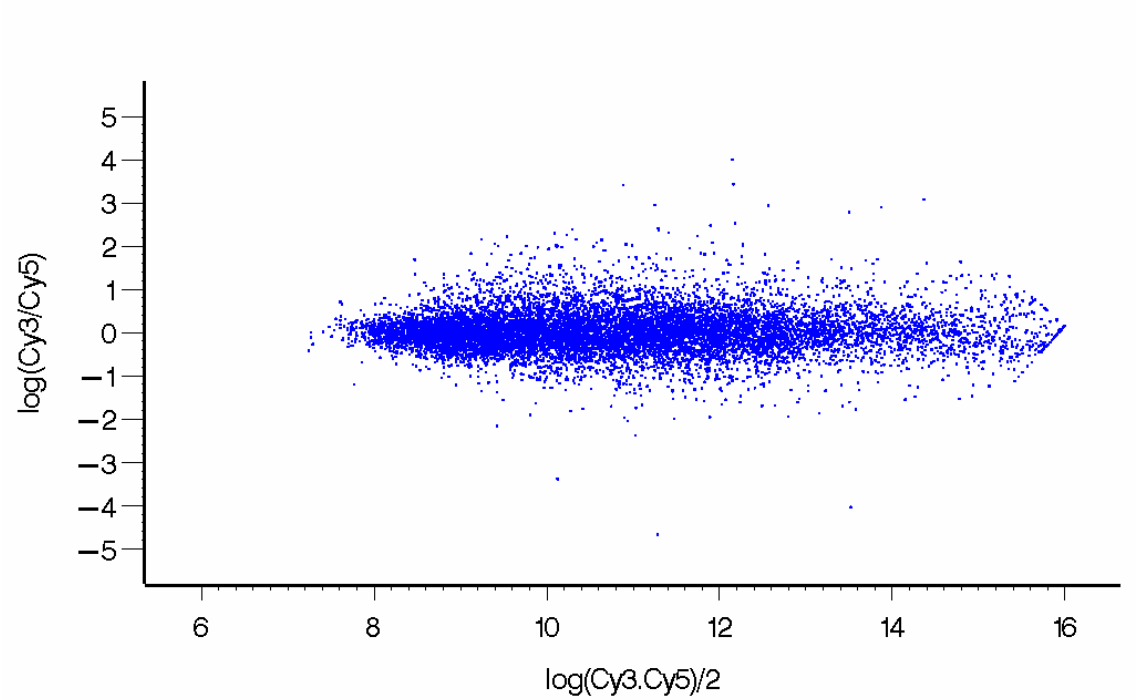

Figura 2: Distribuição do logaritmo da razão de intensidade da expressão medida pelos dois canais (relativos a cada corante) em função da média da intensidade de expressão no slide 1 após a normalização.

Observa-se na Figura 3 a distribuição esquemática das duas amostras nos quatro slides após a normalização. Como os dados sofreram a normalização dentro de slides, não houve diferenças entre as médias das amostras dentro do mesmo slide, uma vez que dentro de slide a amostra está confundida com o corante (*dye*). Contudo, a variação entre slides continua existindo. Notou-se que os slides 2 e 4 apresentam médias mais altas. Esta variação pode ser causada por uma série de diferentes fatores, desde a preparação das amostras no laboratório até a análise das imagens (Draghici, 2003; Rosa et al. 2005)

Após a análise com o modelo [1], também chamada normalização global, foi feito um estudo da distribuição dos resíduos por *slide* (Figura4). Pode-se notar que não houve mais diferenças perceptíveis entre as médias das distribuições nos quatro slides. O objetivo desta normalização foi evitar que as diferenças entre slides pudessem interferir nas análises. É interessante notar que se fossem esperadas diferenças na expressão de todos os genes no

mesmo sentido, então esta análise produziria uma redução nesta diferença e, por conseqüência, no poder de teste da segunda análise.

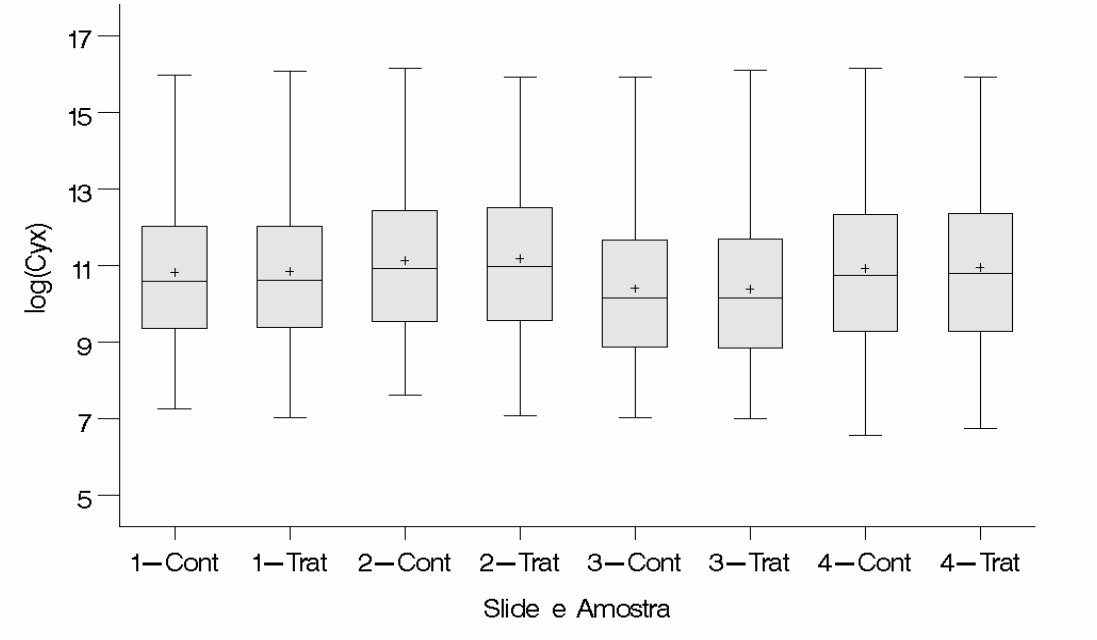

Figura 3 – BoxPlot do logaritmo da razão de intensidade nas duas amostras de cada slide.

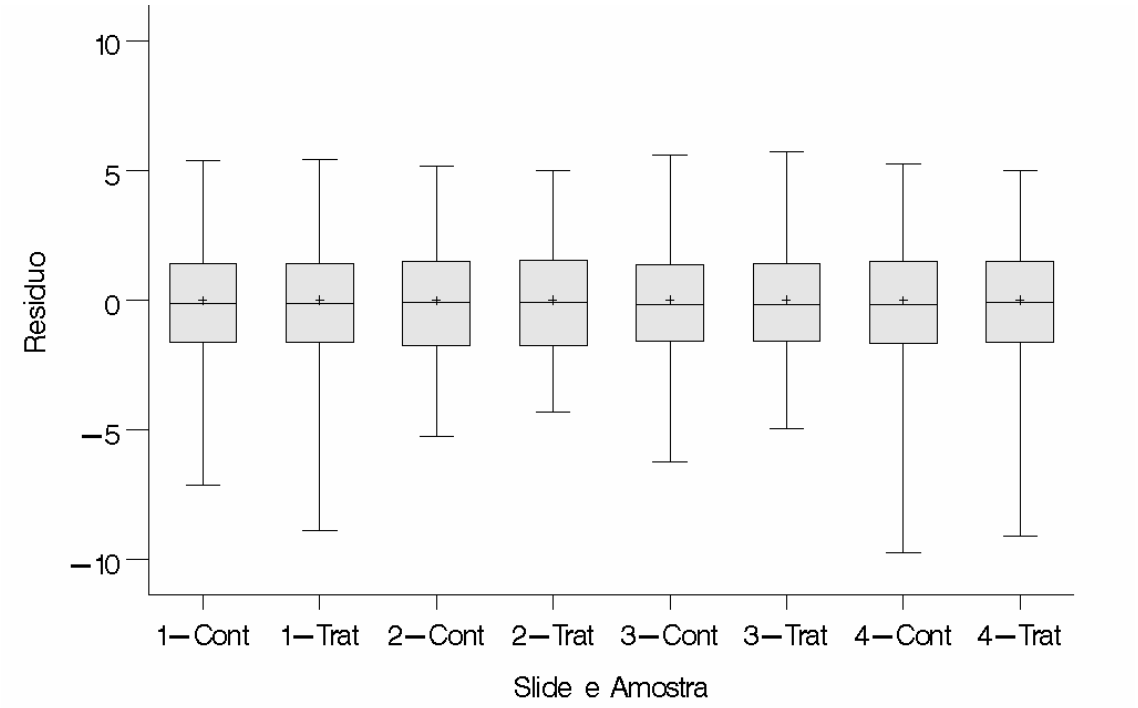

Figura 4 – BoxPlot do resíduo logaritmo da razão da intensidade medida pelos dois canais (corante) após análise pelo modelo [1].

2 – Análises estatísticas.

A análise estatística propriamente dita realizada para cada gene individualmente com o modelo [2], após a correção pelo modelo [1] gerou estimativas de variância residual para cada um dos genes estudados. Observando-se a Figura 5, pode-se notar que a distribuição dos quadrados médios dos resíduos foi altamente assimétrica, apresentando uma concentração em valores muito próximos de zero com média aproximadamente igual a duas vezes a variância. A forma da distribuição dos quadrados médios dos resíduos aproximou-se da forma da distribuição qui-quadrado com poucos graus de liberdade. Ainda que se pudesse supor homocedasticidade para a expressão dos genes estudados neste trabalho, esta análise geraria um grande número de estimativas abaixo da estimativa real. Como a estimativa da variância residual obtida por este método não é viciada e a distribuição destas estimativas é assimétrica, com desvio a direita, a maior parte destes valores

seriam subestimativas da variância verdadeira. A variação em torno da média depende do número de replicações, e assim, quanto menor o número replicações, maior será o de valores subestimados.

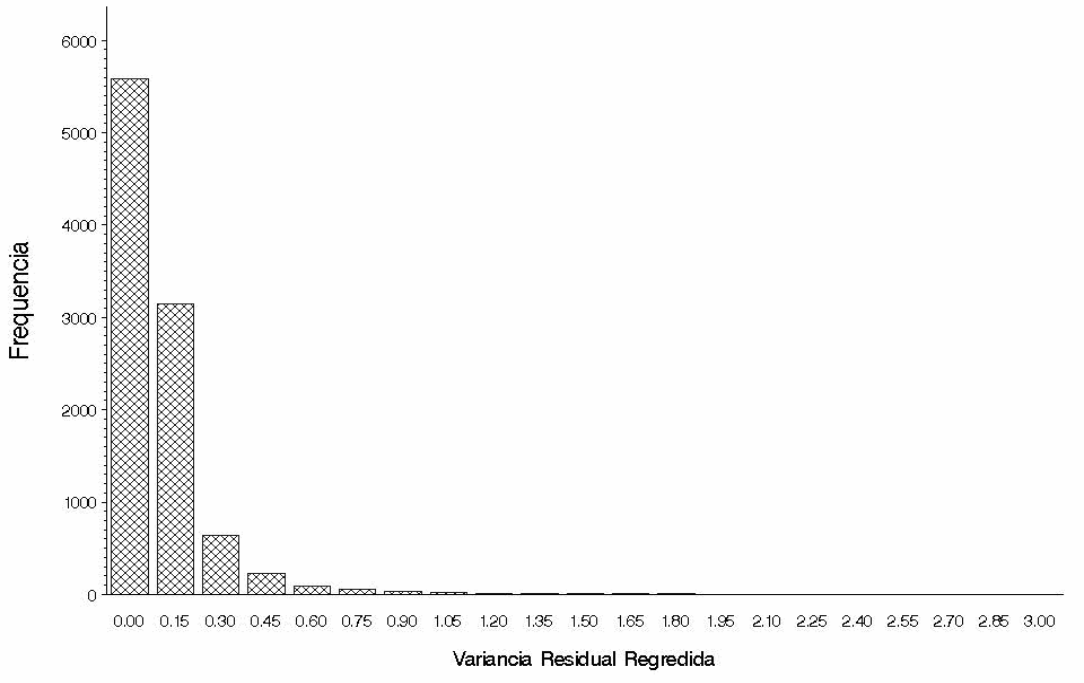

Figura 5 – Distribuição dos quadrados médios dos resíduos para cada gene analisado**.**

A distribuição de freqüência dos valores de probabilidade na análise com o modelo [2] está apresentados na Figura 6. Considerando-se que a hipótese nula seja verdadeira para todos os genes, seria esperado que a distribuição dos valores de probabilidade tivessem uma distribuição uniforme entre 0 e 1 (Nettleton, 2006). Pode-se ver que existe uma tendência de que os valores mais baixos de probabilidade apresentem uma freqüência maior, o que indica que pode existir uma expressão diferencial em uma parte dos genes. Para 1196 dos genes analisados, o valor do teste F ficou acima do valor crítico normalmente considerado de significância no teste individual (p=0,05). Pode-se supor que a distribuição dos valores de probabilidade mostrada na Figura 5 seja uma mistura de uma distribuição uniforme para os genes que não apresentaram expressão diferencial e de valores que tendem a se concentrar em valores menores para aqueles genes que apresentam expressão diferencial. Estes resultados são conseqüências do problema da estimação da

variância residual que o denominador do teste F utilizado e seriam esperados de acordo com Cui *et al.* (2003). De qualquer forma, espera-se que entre estes 1196 genes exista uma importante fração de falsos positivos, mas também se espera que uma parte destes genes esteja realmente se expressando de maneira diferente nas amostras referentes aos diferentes tratamentos.

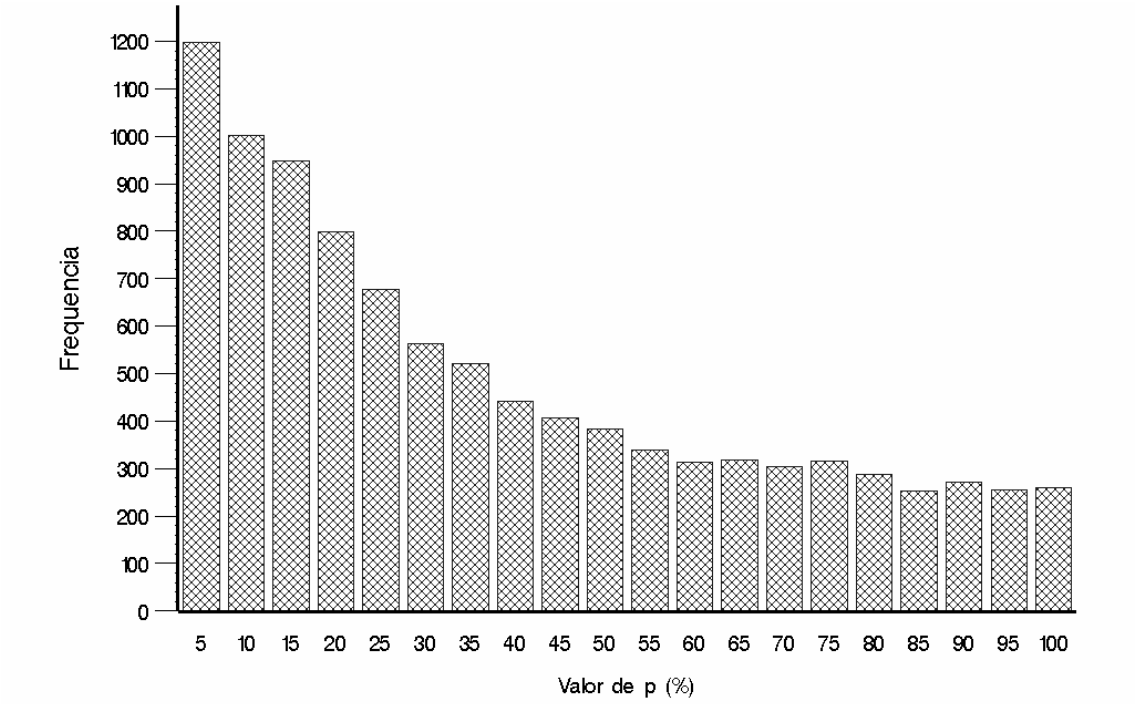

Figura 6 – Distribuição dos valores de probabilidade obtidos no teste F de acordo utilizando-se o modelo [2].

O Figura 7 apresenta a distribuição dos valores de p em função do logaritmo da razão da expressão das duas amostras para cada gene. Os quadrados médios dos resíduos foram distribuídos em classes e estão representados no gráfico, pelas cores dos pontos. A linha contínua representa o valor de -log10(0,05), ou seja, o valor limite de p. Verifica-se neste gráfico que a maior proporção de valores de p menores que 0,05 (valores de –log10(p) maiores que 1.30) estão na faixa central do gráfico, que coincide com a faixa de menor variância residual, apesar de esta faixa apresentar as menores diferenças entre os tratamentos na expressão dos genes. Os pontos nos cantos superiores deste gráfico representariam os genes com maiores significância e diferença. Se a variância dos genes fosse considerada homogênea, as diferenças na expressão dos genes (eixo horizontal) seriam

determinantes para os valores de p. e os pontos formariam uma linha sólida em forma de V. Um dos problemas identificados por vários autores (Baldi *et al.*, 2001; Cui *et al.*; 2003; Smyth, 2004; Cui *et al.* 2005) é que, para muitos dos genes o valor da variância é subestimado e para alguns destes, o valor pode ser muito baixo e uma pequena diferença na expressão dos genes é suficiente para que os valores de p permaneçam abaixo do patamar de 0,05. Isto ocorre especialmente quando o número de repetições é muito pequeno.

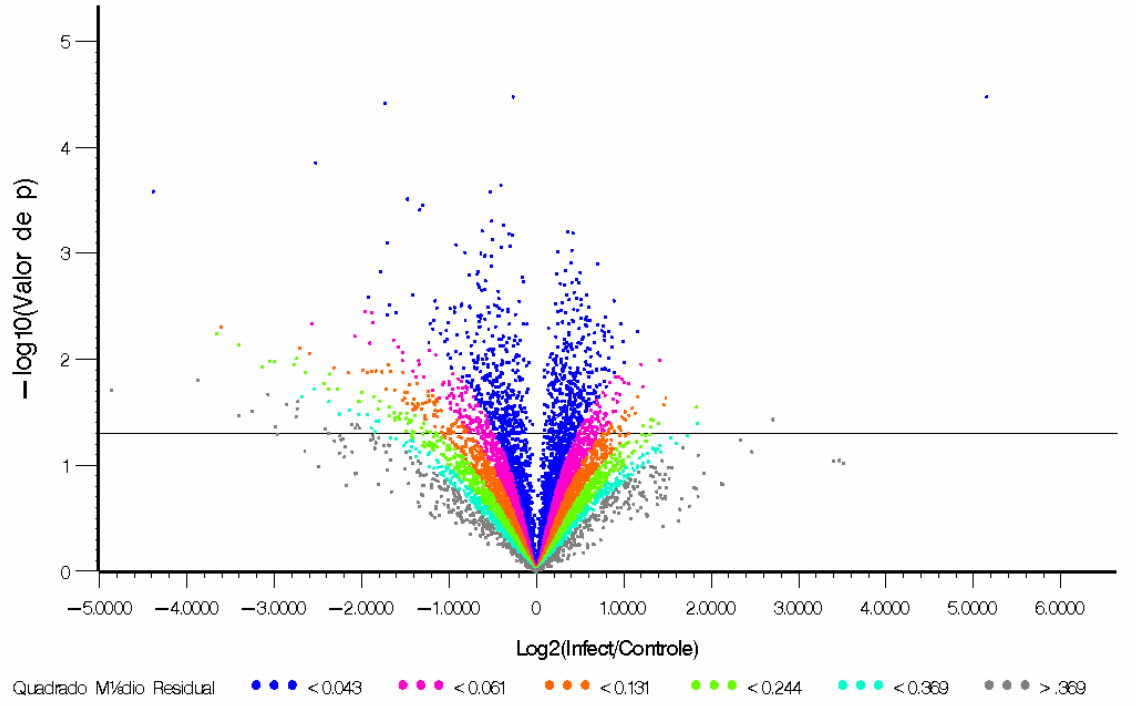

Figura 7 – Distribuição do negativo do logaritmo dos valores de p em função do logaritmo da razão das expressões dos tratamentos (infectados/controle). A Linha contínua representa o negativo do logaritmo de 0,05.

A correção para os testes múltiplos é necessária nos experimentos de *microarray* (Nettleton, 2006; Allison *et al.*, 2006) sendo o teste de Bonferroni eficiente neste sentido, porém extremamente conservador (Speedy, 2003), reduzindo o poder do teste. O método de controle da taxa de falsos positivos (FDR), desenvolvido por Benjamini e Hochberg (1995) pode ser mais eficiente e foi adotado neste trabalho. A aplicação do FDR mostrou que nenhuma das diferenças encontradas pode ser considerada estatisticamente significativa (p>0,05). Apesar da grande proporção de valores de p abaixo do patamar de

0,05, estes resultados poderiam ser esperados uma vez que o número de repetições utilizados neste experimento foi muito pequeno. Segundo Allison *et al.* (2002) e Tsai *et al.* (2003), um número mínimo a ser aplicado em um experimento de microarray seria de cinco replicações biológicas. Este número seria abaixo do ideal, mas pode funcionar se o objetivo do experimento for apenas para encontrar as diferenças entre duas amostras.

3. Regressão das variâncias residuais

O método de Cui *et al.* (2005) regride a variância residual de cada gene para a média geométrica das variâncias, criando uma nova estimativa da variância residual que é um valor ponderado entre este valor global e o efeito de cada gene em particular. A variação observada no conjunto das estimativas dos quadrados médios residuais é utilizada para determinar o fator de ponderação, ou seja, quão próximo da média geométrica ou da estimativa individual as novas estimativas deverão ficar. No caso do presente trabalho, a média geométrica das estimativas individuais ficou em 0,0522 (a unidade destas medidas refere-se ao quadrado do logaritmo da quantidade de pixel refletindo cada cor e será omitida daqui em diante). A variância destas mesmas estimativas ficou em 6,689. Como este valor era relativamente pequeno, quando comparado com a variância esperada caso houvesse homocedasticidade para a expressão de todos os genes (5,155), a regressão teve um efeito acentuado. Desta forma, as distribuições das novas estimativas dos quadrados médios aproximaram-se bastante da média. O Gráfico 8 apresenta a distribuição dos valores das estimativas de variância regredidas. Pode-se observar, em comparação com a distribuição dos quadrados médios originais, que a amplitude de variação dos valores regredidos é bem menor.

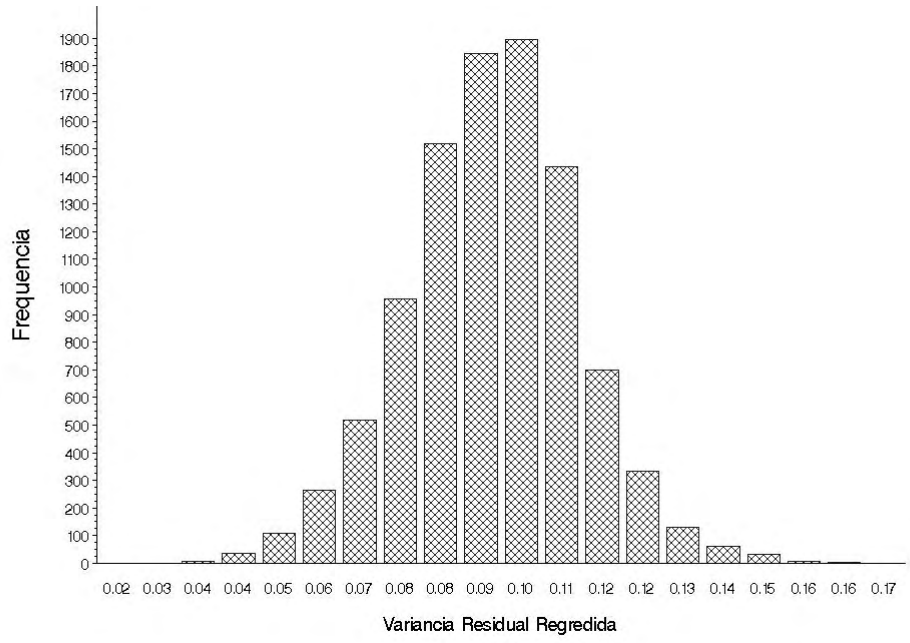

Figura 8 – Distribuição da variância residual de cada gene regredida segundo método de Cui et al. (2005).

Observa-se também que, em contraste com a distribuição dos valores originais, a distribuição destes valores é aproximadamente simétrica, sendo a média e a mediana relativamente próximas. Não houve, portanto, uma maior concentração de valores próximos de zero. O teste estatístico (FS), calculado utilizando a estimativa regredida da variância residual, reduziu sensivelmente a proporção de valores de p abaixo do patamar de 0,05. Apenas 487 genes, o que representaria um pouco menos de 5% dos genes estariam enquadrados nesta faixa. Apesar disto, a distribuição dos valores de p (Gráfico 9) não se apresenta uniforme, com uma concentração maior de valores entre 0,05 e 0,30.

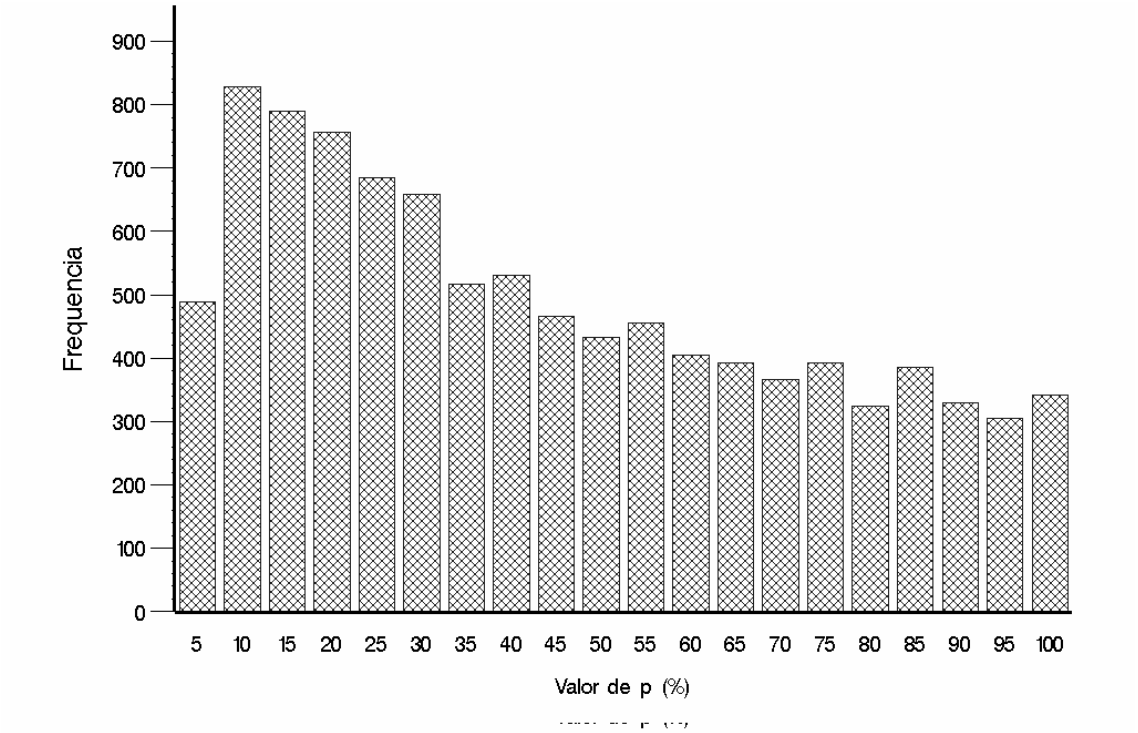

Figura 9 – Distribuição dos valores de p para o teste FS (Cui et al., 2005).

Na Figura 10 apresenta-se a distribuição do negativo do logaritmo dos valores de p em função do logaritmo da razão da expressão dos genes nas amostras dos animais infectados para aqueles dos animais não infectados, utilizando-se o teste FS (Cui *et al.*, 2005). As cores no gráfico também representam classes de variância residuais originais. Como se poderia esperar, em função da maior concentração dos valores de variância residual utilizados no teste, a dispersão dos valores de p em função das diferenças de expressão gênica entre os tratamentos, foi bem menor neste caso. Como o denominador do teste é menos disperso e o numerador continua o mesmo do teste anterior, os valores de p tendem a depender bem mais da diferença entre os genes. Também é possível perceber que a proporção de valores de p abaixo de 0,05 não é mais dependente da estimativa da variância residual ou, pelo menos, que a maior proporção destes não está na classe com menor variância residual.

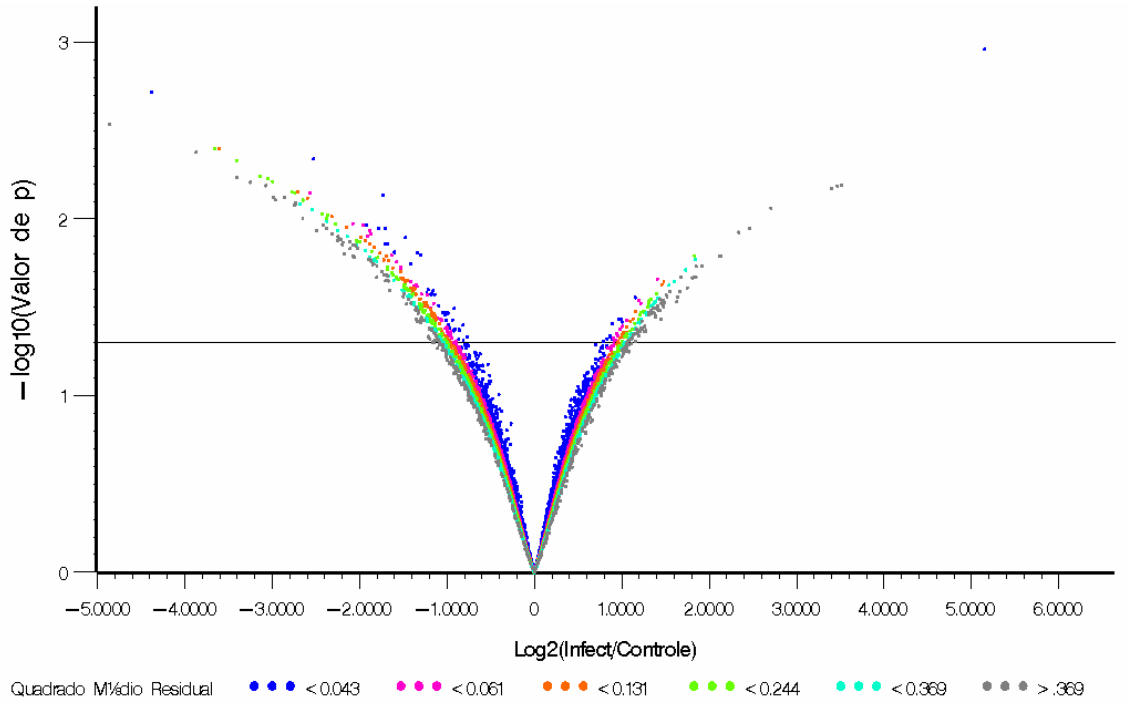

Figura 10 - Distribuição do negativo do logaritmo dos valores de p obtidos no teste FS (Cui et al., 2005) em função do logaritmo da razão das expressões os tratamentos (infectados/controle). A Linha continua representa o negativo do logaritmo de 0,05.

A aplicação da metodologia para teste múltiplos (FDR) também não revelou, neste caso, significância (p>0,05) para nenhum dos genes estudados. Segundo Cui *et al.,* (2005) seria mais interessante gerar um valor critico para F e FS a partir de uma técnica de permutação dos slides, já que a distribuição F não seria totalmente adequada. Talvez esta técnica permitisse a identificação de expressão diferencial dos genes, mas não foi possível utilizá-la uma vez que as permutações devem sempre envolver os slides como um todo (e não cada gene em particular) e como o número de slides era relativamente pequeno não seria possível aplicar neste estudo. Segundo Draghici (2003) são necessários pelo menos seis slides para gerar um mínimo de permutações que garanta a confiabilidade do teste. De qualquer forma, uma vez que menos de 5% dos genes apresentaram valor de p menor que 0,05, este resultado era esperado.

CONCLUSÕES

Os experimentos de *microarray*, por tratarem de um conjunto muito grande de observações a serem analisados requerem análises estatísticas específicas. O método de Cui *et al.* (2005) reduziu a dependência entre a variância residual e o valor de probabilidade do teste F aplicado de acordo com o modelo utilizado, e desta forma, mostra-se adequado para este fim. Todavia, não foi observado neste trabalho aumento do poder de teste.

REFERÊNCIAS BIBLIOGRÁFICAS

Allison, D.B., Cui, X., Page, G.P., and Sabripour, M. Microarray data analysis: From disarray to consolidation and consensus. **Nat. Rev. Genet.** 7, 55–65, 2006.

Allison, D. B. et al. A mixture model approach for theanalysis of microarray gene expression data. **Comput. Stat. Data Analysis** 39, 1–20, 2002.

Axon Instruments Inc., GenePix Pro 3.0, Technical manual Axon Instruments, Union City, CA, 2001.

Baldi, P. & Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. ***Bioinformatics*** **17**, 509–519, 2001.

BENJAMINI, Y;Hocheberg, Y. Controling the False Discovery Rate: a Pratical and Powerful Approach to multipleTesting. **Journal of the Royal Statistics Society B,** v. 57, p.289-300, 1995

CUI, X.; HWANG,J.T.G.; QIU, J.; BLADES, N.J.; CHURCHILL, G.A. Improved statistical tests for differential gene expression by shrinking variance components estimates. **Biostatistics**, v.6, 1, pp.59-75, 2005.

CUI, X. e CHURCHILL, G. A. Statistical tests for differential expression in cDNA microarray experiments. **Genome Biology.** V.4. (4):210, 2003.

Draghici, S. **Data Analysis Tools for DNA Microarray.** 1.ed. **I**n Chapman & Hall/CRC editors, New York, 517 p., 2003.

DUDOIT, S.; YANG, Y.H.; CALLOW, M. J.; SPEED, T.P.. Statistical methods for identifying differentially expressed gene in replicated cDNA microarray experiments. **Statistical Science**, v. 12, n° 1, p.111-139, 2002.

Edwards, J. W. *et al.* Empirical Bayes estimation of genespecific effects in micro-array research. ***Funct. Integr.Genomics*** **5**, 32–39, 2005.

KERR, M. K. e CHURCHILL, G. A. Experimental design for gene expression *microarrays.* **Biostatistics,** v. 2; p. 183-201, 2001.

KERR, M. K.; MARTIN, M.; CHURCHILL, G. A. Analysis of variance for gene expression *microarray* data. **J. Comp. Biol.,** v. 7; p. 819-837, 2000.

LOCKHARDT, D.J., DONG, H.L., BYRNE, M.C., FOTTETTIE, M.T., GALLO, M.V., CHEE, M.S., MITTMANN, M., WANG, C., KOBAYASHI, M. and HORTON, H. . Expression monitoring by hybridization to high-density oligonucleotide arrays. **Nat. Biotechnol.**., 14, 1675-1680. , 1996.

MAH, N.; THELIN, A.; LU, T.; NIKOLAUS, S; KÜHBACHER, T.; GURBUZ, Y.; EICKHOFF, H.; KLÖPPEL, G.; LEHRACH, H.; MELLGARD, B.; COSTELLO, C.M.; SCHREIBER, S. A comparison of oligonucleotide and cDNA – based microarray systems. **Physiol Genomics**, v.16, p.361-370, 2004.

MILLER RA, Galecki A, Shmookler-Reis RJ. Interpretation, design, and analysis of gene array expression experiments. ***J Gerontol A-Biol***; 56: B52-B57, 2001.

Nettleton,D. A Discussion of Statistical Methods for Design and Analysis

of Microarray Experiments for Plant Scientists. **The Plant Cell**; 18: 2112-2121, 2006.

ROSA, G.J.M.; STEIBEL, J.P. ; TEMPELMAN, R. J. Reassessing design and analysis of two-color microarray experiments using mixed effects models. **Comparative and Functional Genomics**, v. 6; p.123-131, 2005 .

SAS - Statistical Analysis Sistem. SAS® User's guide: Statistical, ver. 9. Cary: SAS Institute Inc, 1999.

SCHENA, M.; SHALON, D.; DAVIS, R.; W; BROWN, P. O. Quantitative monitoring of gene-expression patterns with a complementary DNA *microarray.* **Science,** v. 270; p 467-470, 1995.

SILVA H.D. ; VENCOVSKY R. Poder de detecção de "Quantitative Trait Loci", da análise de marcas simples e da regressão linear múltipla. **Scielo agric.** (Piracicaba, Braz) v.59 n.4, 2002.

SMYTH, G.K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. **Statistical Applications in Genetics and Molecular Biology**. V.3. Issue 1.Article 3. 2004.

Speedy, T . **Statistical Analysis of Gene Expression .Microarray.** 1.ed .CRC PRESS, California-EUA., 216 p. , 2003.

STEIBEL, J. P. ; TEMPELMAN R. J.; ROSA, G.J.M. Power and sample size determinations for two color microarray experiments based on different levels of replication. *Submetido à publicação.*

Tsai, C. A., Hsueh, H. M. & Chen, J. J. Estimation offalse discovery rates in multiple testing: application to gene microarray data. **Biometrics** 59, 1071–1081, 2003.

WOLFINGER, R. D.; GIBSON, G.; WOLFINGER, E. D.; BENNETT, L.; HAMADEH, H.; BUSHEL, P.; AFSHARI, C.; PAULES, R. S. Assessing gene significance from cDNA *microarray* expression data via mixed models. **J. Comp. Biol.,**v. 8; p. 625-637, 2001.

IMPLICAÇÕES

Observa-se pelo contexto do trabalho que as metodologias de análises estatísticas propostas para o estudo da expressão gênica foram de grande importância. A maior parte dos estudos envolvendo a técnica de *microarray* avalia principalmente o processo laboratorial, ficando a análise estatística como um complemento do trabalho. Já no nosso estudo, enfatizamos a análise estatística propriamente dita, o que contribuiu consideravelmente para muitos estudos da área, trazendo resultados mais precisos e coerentes que serão aplicados na prática.

O estudo da expressão gênica utilizando experimento de *microarray* é muito utilizado em vários países do mundo, sendo que no Brasil ainda é pouco explorado comparado aos Estados Unidos e outros países europeus. Portanto, além da importância citada anteriormente, esse trabalho contribuirá com a pesquisa brasileira na área.

Deve-se ressaltar que esse trabalho foi pioneiro em pesquisas brasileiras para esse estudo, ficando registrado a sua relevância.