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

RENATO NUNES PEREIRA

2008

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Pereira, Renato Nunes.

Controle do erro tipo I em um experimento de microarrays com eucalipto / Renato Nunes Pereira. – Lavras: UFLA, 2008.
56 p. : il.

Dissertação (Mestrado)- Universidade Federal de Lavras, 2008.
Orientador: Júlio Sílvio de Sousa Bueno Filho.
Bibliografia.

1. Eucalipto. 2. FDR 3. Microarrays. 4.transformação Box-Cox. I. Universidade Federal de Lavras. II.Título.

CDD - 519.538

RENATO NUNES PEREIRA

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

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

APROVADA em 15 de fevereiro de 2008.

Prof^{a}. Dra. Luzia Aparecida Trinca UNESP

Prof. Dr. Paulo César Lima UFLA

Prof. Dr. Júlio Sílvio de Sousa Bueno Filho
UFLA
(Orientador)

LAVRAS
MINAS GERAIS - BRASIL

AGRADECIMENTOS

Meus sinceros agradecimentos ao professor Júlio Sílvio de Sousa Bueno Filho, pela paciência com que me orientou, disponibilidade em auxiliar-me a qualquer momento, pelas críticas e sugestões.

Aos meus pais, José e Santa, pela confiança, compreensão, carinho e apoio.

Aos meus irmãos, pelo carinho, compreensão e torcida em todos os momentos.

A todos os colegas de mestrado e doutorado em Estatística, especialmente a Natascha, pela amizade e por ter sido minha grande companheira de estudo, e a Dorival, Popó, Ricardo e Tiago, pelo apoio, amizade e por terem sido a minha família aqui em Lavras.

Ao meu amigo Udi Florião que tem demonstrado muito carinho e preocupação comigo. Obrigado pela força e disposição em sempre me ajudar.

À Dona Terezinha, por ter cuidado tão bem de mim aqui em Lavras.

Aos funcionários do Departamento de Ciências Exatas: Edila, Josi, Joyce, Maria, Selminha e Vânia, pela simpatia e boa vontade no atendimento.

Aos professores do Departamento de Ciências Exatas, pelos ensinamentos prestados.

À Universidade Federal de Lavras e ao Departamento de Ciências Exatas, pela oportunidade da realização deste curso.

À CAPES, pela bolsa de estudos, essencial para a realização deste trabalho.

Aos demais que, direta ou indiretamente, contribuíram para a elaboração deste trabalho.

E a Deus, por ter me dado tudo que eu sempre precisei, para que eu pudesse tornar cada sonho em realidade.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Pereira, Renato Nunes. **Controle do erro tipo I em um experimento de microarrays com eucalipto.** 2008. 56 p. Dissertação (Mestrado em Agronomia/ Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, MG. *

No projeto Genolyptus, que busca identificar aspectos da genômica funcional do eucalipto, foi instalado um experimento inicial utilizando 10 microarrays monocromáticos da plataforma Nimblegen. Neste tipo de experimento são realizados milhares de testes simultâneos para diferentes variáveis-resposta na mesma estrutura de unidades experimentais. No presente trabalho, estudaram-se as consequências da utilização da taxa de falsos positivos (FDR,Benjamini & Hochberg,1995) na definição de sondas com maior potencial de manifestar expressão diferencial em uma estrutura de tratamentos que representa altos e baixos teores de lignina. Os cinco tratamentos utilizados nas primeiras unidades experimentais foram: T1: *Eucaliptus grandis*, clone 1, amostra de folha; T2: *Eucaliptus grandis*, clone 1, amostra de tronco; T3: *Eucaliptus grandis*, clone 2, amostra de folha; T4: *Eucaliptus globulos*, clone 3, amostra de folha e T5: *Eucaliptus globulos*, clone 4, amostra de folha. Foram utilizadas repetições biológicas (um clone diferente) para cada tratamento, nas demais unidades experimentais. No caso dos dados não transformados, verificou-se, pelo teste de Shapiro-Wilk, que os resíduos da análise dos dados de 1113 sondas não atendiam à pressuposição de normalidade. A transformação de Box-Cox corrigiu o problema da não normalidade para 801 dessas sondas, tendo sido detectada significância em 12 das 801 análises, para o efeito de tratamentos. Para as demais sondas em que a análise resultou em resíduos aproximadamente normais, 40 revelaram significância para o efeito de tratamentos, perfazendo um total de 52 sondas a serem investigadas para os contrastes de maior interesse. Usando o FDR e concluindo pelo teste "t"de Student, a 1% de significância, foram significativas para o contraste Folha x Xilema em *E.grandis* 33 sondas; para o contraste *E.grandis x E.globulos*, foram significativas 43 sondas; para o contraste entre clones dentro da *E.grandis*, foram significativas 36 sondas e, para o contraste entre clones dentro de *E.globulos*, foram significativas 13 sondas. O uso do procedimento FDR e a transformação Box-Cox apresentaram resultados satisfatórios em reduzir o número de sondas, tornando possível montar estudos posteriores sobre o potencial de expressão diferencial e de validação.

Palavras-chave: eucalipto, FDR, microarrays, transformação Box-Cox.

* **Comitê Orientador:** Júlio Sílvio de Sousa Bueno Filho - UFLA. (Orientador)

ABSTRACT

Pereira, Renato Nunes. **Type I error rate control in a microarrays experiment with eucalyptus.** 2008. 56 p. Dissertation (Master in Agronomy /Statistics and Agricultural Experimentation) Federal University of Lavras, Lavras, MG.*

Genolyptus project aims to identify features of functional genomics of Eucalyptus sp. Within this project a microarray experiment using 10 Nimblegen platform (monochromatic) slides was performed. In this type of experiment thousands of simultaneous tests for different response variables are carried out in the same experimental units. In this work we investigate the consequences of applying False discovery rates (FDR, Benjamini & Hochberg, 1995) to identify probes with greatest potential of differential expression. The treatment structure comprises high and low lignin content. Five treatments were established, namely: T1: *Eucaliptus grandis*, clone 1, leaf sample; T2: *Eucaliptus grandis*, clone 1, stem sample; T3: *Eucaliptus grandis*, clone 2, leaf sample; T4: *Eucaliptus globulos*, clone 3, leaf sample; T5: *Eucaliptus globulos*, clone 4, leaf sample. An extra biological sample (a different clone) was established for each treatment combination. 1113 non-transformed response variables did not pass Shappiro-Wilk normality test and were then transformed according to Box-Cox procedure. From the 801 normally distributed response variables, after transformation, 12 showed significance of treatment effects. For the remaining probes that passed normality test, 40 have shown significance of treatment effects, resulting 52 probes to be investigated on specific contrasts and further testing. Using FDR and "t" test at 1% significance level, for the contrast *E.grandis x E.globulos* there where 43 significant probes. There were 33 significant contrasts between leaves and stems; 36 among clones of *E.grandis* and 13 significant contrasts among clones of *E.globulos*. FDR procedure and Box-Cox transformation had shown satisfactory results in reducing the number of probes to a number that can be handled in further validation studies of differential expression.

Key-words: Box-Cox Transformation, Eucalyptus, FDR, Microarrays.

*Guindance Committee: Júlio Sílvio de Sousa Bueno Filho - UFLA. (Adviser)

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 simultâneos para diferentes variáveis-resposta na mesma estrutura de unidades experimentais.

Na análise de qualquer situação experimental, dois erros podem ser cometidos. O erro tipo I, ou falso positivo, e 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.

Uma estratégia que tem sido encontrada na literatura é a utilização de um procedimento proposto por Benjamini & Hochberg (1995), que controla a taxa de falsos positivos (FDR, do inglês *False discovery rate*). A FDR é definida como a proporção esperada de falsos positivos entre todas as hipóteses nulas rejeitadas.

O procedimento FDR tem sido aplicado em áreas diversas, como por exemplo em análise de microarrays de DNA. 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 um projeto que buscou identificar aspectos da genômica funcional do eucalipto, projeto Genolyptus, foi instalado um experimento inicial utilizando 10 microarrays monocromáticos da plataforma Nimblegen. O objetivo central era determinar espécies, tecidos e órgãos com expressão diferencial para o teor de lignina. O teor de lignina é um caráter de importância tecnológica no aproveita-

mento 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.

Este trabalho foi realizado com o objetivo de analisar um experimento com 10 slides do programa de genética funcional de eucalipto (Genolyptus) e verificar em que a medida transformação de dados e o uso de FDR melhoram os resultados deste experimento.

2 REFERENCIAL TEÓRICO

2.1 A técnica dos microarrays de DNA

Os microarrays de DNA constituem uma tecnologia muito recente e, atualmente, apresentam-se como sendo a de maior potencial na solução de problemas para a biologia, sendo amplamente utilizada para analisar o padrão de expressão gênica, uma vez que permite a verificação da manifestação de milhares de genes simultaneamente. Atualmente, os microarrays podem ser usados para distintos tipos de análises, como por exempo, em análise de expressão gênica, detecção de polimorfismos, re-sequenciação genética, genotipagem e escalagem genômica (Schena et al., 1996).

O primeiro microarray de DNA com 45 sondas de cDNA (ácido desoxirribonucléico complementar, seqüência de um determinado gene) foi introduzido por Schena et al. (1995). O progresso tecnológico dos microarrays de DNA foi extremamente rápido; em 1996, publicações com 1.000 sondas de arrays já haviam sido apresentadas (Schena et al.; Shalon et al.) Hoje, são comuns arrays com dezenas de milhares de sondas. A empresa pioneira no campo foi a Affymetrix (Santa Clara, CA, E.U.A.), que trabalhou com a metodologia de microarrays de uma só cor ou canal.

Na tecnologia de microarrays com lâminas(slides) de vidro, diversas seqüências de DNA conhecidas (produtos de PCR, *polymerase chain reaction, ou reação em cadeia por polimerase*, oligonucleotídeos), as chamadas sondas, é impressa em uma lâmina. A partir do RNA de duas condições distintas é preparado o cDNA, em que são incorporados nucleotídeos marcados com fluorescência distinta Cy3 (verde) e Cy5 (vermelho). No caso da metodologia de microarray de duas cores, há duas populações de cDNA marcados que são submetidas, simultaneamente, à

hibridização com a lâmina contendo as sondas de interesse, permitindo que os alvos presentes em solução sejam pareados e assim possam revelar a expressão dos genes. Haverá competição entre as duas populações de cDNA pelas sondas e a resposta final é um resultado relativo de expressão entre as duas populações.

No caso de microarrays de uma só cor, uma população de cDNA é hibridizada às sondas de um slide. Para a preparação dos microarrays, utilizam-se robôs altamente precisos, que aplicam as diferentes amostras de DNA em diminutos pontos (spots) no centro dos slides, com densidade aproximada de 10.000 pontos/cm^2.

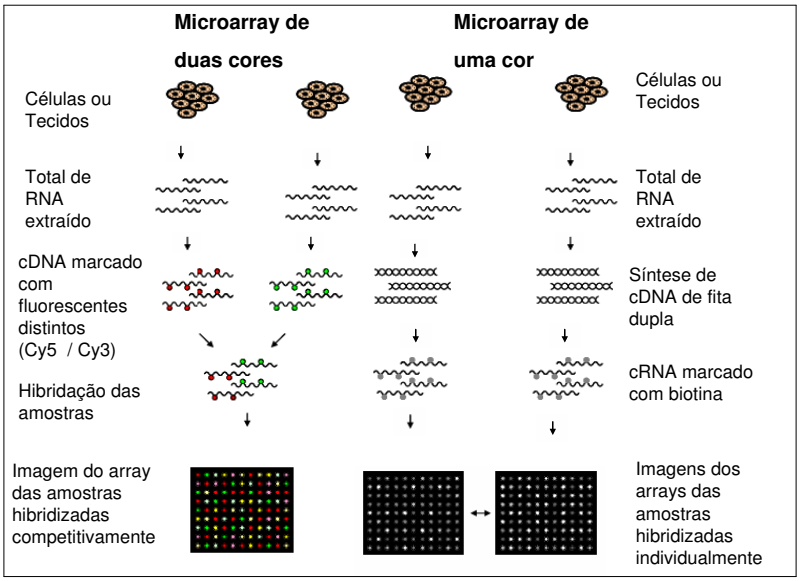


FIGURA 2.1: Esquema da técnica de microarrays para arrays de duas cores e uma cor .

Segundo Kokko (2006), nos arrays de duas cores, duas amostras de RNA são submetidas a reações de transcrição reversa em paralelo, cada uma recebendo um corante diferente Cy3 ou Cy5 . As duas amostras são purificadas, desnaturadas e colocadas, simultaneamente para hibridizar com a lâmina contendo os clones de interesse. Após as devidas lavagens, o slide é lido em um scanner, a imagem é interpretada por um software específico e os dados são analisados por ferramentas estatísticas.

Nos arrays de uma cor, um procedimento de transcrição reversa é usado para produzir cDNA de fita dupla, que é transcrito e amplificado in vitro para cRNA marcado com biotina. O cRNA biotinizado é, então, fragmentado e hibridizado no chip. Após a hibridização, o cRNA não hibridizado é removido do array e o chip é submetido a uma série de lavagens e etapas de coloração, em que o corante fluorescente streptavidin-phycoerythrin (SAPE) liga com a biotina do cRNA marcado. Finalmente, o array é digitalizado usando-se um laser que excita o corante fluorescente. O processo de leitura da imagem é o mesmo que o de arrays de duas cores.

A síntese de microarray usando a tecnologia do sistema Nimblegen, usado na obtenção dos dados analisados neste trabalho, é muito similar à tradicional síntese de "Short oligos"da Affymetrix, com algumas exceções importantes. A técnica de impressão é diferente, pois não usa agulhas, jato de tintas e nem máscaras fotolitográficas, mas microespelhos especialmente desenvolvidos para minimizar artefatos de revelação.

Segundo Cristo (2003), existem outros tipos de substratos que podem ser usados para a fixação dos cDNAs, como é o caso das membranas de náilon (que apresentam certas diferenças em relação às lâminas de vidro). As membranas são fisicamente maiores, assim como seus "spots"(ou pontos de impressão) e a dis-

tância entre eles (plataforma conhecida na literatura por cDNA Array, em vez de microarray); a quantidade de "spots "é, em geral, menor e as superfícies são porosas, ao contrário das lâminas de vidro que são completamente lisas (rígidas). Este último detalhe é importante, pois o material fixado não fica exposto diretamente, como no vidro. Nas membranas de náilon não existe marcação com fluorescência, sendo utilizado um marcador radioativo. Por essa razão, esse tipo de experimento também é conhecido como microarray de um canal, dado que somente uma amostra de mRNA pode ser testada por experimento.

O maior mérito da técnica de microarrays está na possibilidade de observar o padrão de expressão de um grande número de genes em um único ensaio. Lâminas de oligonucleotídeos de alta densidade chegam a conter representantes de todo o genoma de um organismo em uma única lâmina.

A enorme quantidade de dados gerada por esta técnica e a complexidade envolvida dificultam a comparação dos dados encontrados por diferentes grupos. Dados de microarrays , geralmente, não se apresentam em valores absolutos, mas em diferenças relativas. O contexto no qual cada ensaio é realizado, o organismo ou o tipo celular utilizado, as diferentes plataformas de microarray e as formas diversas de extração de imagem e normalização de dados numéricos estão entre os principais fatores que dificultam a comparação e a interpretação dos resultados de experimentos diferentes. Um problema que, em geral, ocorre nestes experimentos é que, devido ao alto custo, em geral, há poucos arrays disponíveis, o que resulta em um experimento com muitas variáveis resposta e poucas unidades experimentais.

Com a intenção de organizar os dados gerados, um grupo de especialistas se reuniu para elaborar o MIAME, ou *minimum information about a microarray experiment* (Brazma et al., 2001). No MIAME, são reunidas sugestões a respeito de informações sobre delineamento do experimento, desenho do chip de DNA, trata-

mentos e obtenção das amostras de RNA, parâmetros de hibridização, medidas de imagens, formas de normalização e até mesmo vocabulário empregado. O objetivo é disponibilizar, de forma ordenada, informações que tornem mais fácil a interpretação correta dos dados gerados em cada experimento. Atualmente, já existem alguns bancos de dados de expressão gênica que permitem a busca e a comparação entre os resultados obtidos (*Gene Expression Omnibus* − NCBI ; *ArrayExpress* − EBI ; DNA *data base of Japan*).

2.2 Normalização

A razão de expressão dos experimentos deveria variar apenas em função da questão biológica, e não da tecnologia empregada, no entanto, como qualquer outra tecnologia, os microarrays têm problemas experimentais, que fazem com que pessoas que os lêem cometam alguns erros, que podem mascarar ou alterar o resultado da análise final. Isto é bastante agravado por se observar um número grande de variáveis respostas. Alguns exemplos de fatores que podem gerar erros são as diferenças na eficiência de incorporação e de brilho dos dois fluoróforos usados, a posição do spot no slide, a intensidade do spot, a qualidade do spotter, a iluminação de fundo, as variações do scanner, as variações entre experimentos e outros. Estes erros experimentais são particularmente problemáticos quando são feitas buscas por variações sutis, quando é possível ocorrer o mascaramento ou, mesmo, a sobreposição do erro sobre a leitura, causando erros de análise. É evidente que a normalização, bem como a análise estatística dos dados, é um passo determinante em ensaios de microarrays.

Para o microarray de duas cores, a razão de expressão costuma ser expressa em escala logarítmica. Ao chamar um dos canais de R (red, ou vermelho, cor associada ao corante Cy5) e o outro de G (green, ou verde, cor associada ao corante

Cy3), é possível dizer que a razão de expressão entre os dois canais (M) é:

$$M = log_2 R - log_2 G = log_2(\frac{R}{G}). \tag{2.1}$$

A intensidade dos spots costuma ser expressa pela equação, representada por A: $A = (log_2 R + log_2 G)/2$, ou seja, a média dos logs das intensidades.

Para o microarray de uma cor, a normalização é simplesmente log da intensidade luminosa(Y), como na expressão 2.2.

$$B = log_2 Y \tag{2.2}$$

Não existe um método ideal para a normalização dos dados de microarrays, mas diferentes métodos que são eficientes em condições específicas.

2.3 Análise de variância e transformação de dados

Segundo Cox & Reid (2000), a análise de variância é uma técnica desenvolvida por Fisher entre 1920 e 1930 para análise e interpretação de dados experimentais. Para a utilização dessa técnica, algumas pressuposições básicas devem ser satisfeitas, entre elas a independência e a normalidade dos erros.

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.

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. Box & Cox (1964) sugeriram uma família de transformações baseada em uma função de verossimilhança, para a escolha de transformações adequadas que tornam o modelo linear, homocedástico e normal, conforme a seguinte expressão:

$$Y^{\lambda} = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \\ log(Y) & \text{se } \lambda = 0, \end{cases}$$

em que: $log(Y)$ é o logaritmo neperiano e esta família é válida para $Y > 0$. O procedimento de Box-Cox utiliza o método de máxima verossimilhança para estimar λ. Uma vez obtido o valor de λ, encontram-se os valores dos dados transformados conforme a equação acima e utilizam-se estes dados para efetuar a análise de variância.

2.4 Tipos de erro e testes múltiplos

Segundo Mood et al. (1974), o erro tipo I é definido como o erro que se comete ao rejeitar a hipótese nula, quando esta é verdadeira (α) e o erro tipo II, como o erro que se comete ao não rejeitar a hipótese nula, quando na verdade ela é falsa (β). O poder de um teste é definido também, pelos mesmos autores, como a probabilidade de rejeitar a hipótese nula H_0, quando ela é falsa $(1 - \beta)$.

Freqüentemente, a preocupação em controlar o erro tipo I ocorre em experimentos em que há muitos tratamentos e contrastes de interesse não ortogonais. O grande número de hipóteses a ser testada simultaneamente gera o problema dos

chamados testes de comparações múltiplas.

Quando trabalhamos neste tipo de pesquisas, surge o problema que se refere ao nível de significância conjunto da análise e, conseqüentemente, o seu poder, pois o nível de significância conjunto aumenta à medida que aumenta o número de testes realizados. Adotando-se um nível de significância (α) para cada teste, tem-se o nível de significância conjunto do teste(α^*). Considerando-se os m testes independentes será:

α^* = probabilidade de rejeitar a hipótese nula verdadeira em pelo menos um teste, ou seja,

α^* = $1-$ probabilidade de não rejeitar hipótese nula verdadeira em nenhum teste, isto é,

$$\alpha^* = 1 - (1 - \alpha)^m$$

Suponha que 10 hipóteses estão sendo testadas com $\alpha = 5\%$ (assumindo testes independentes). A probabilidade de se rejeitar a hipótese nula verdadeira em um teste é de 0,05, mas a probabilidade de se rejeitar hipótese nula verdadeira em pelo menos um teste é de 0,401. Se forem executados 20 testes de hipóteses, a probabilidade aumenta para 0,642. Pelo gráfico da Figura 2.2, observa-se como se comporta o nível de significância, à medida que aumenta o número de testes realizados.

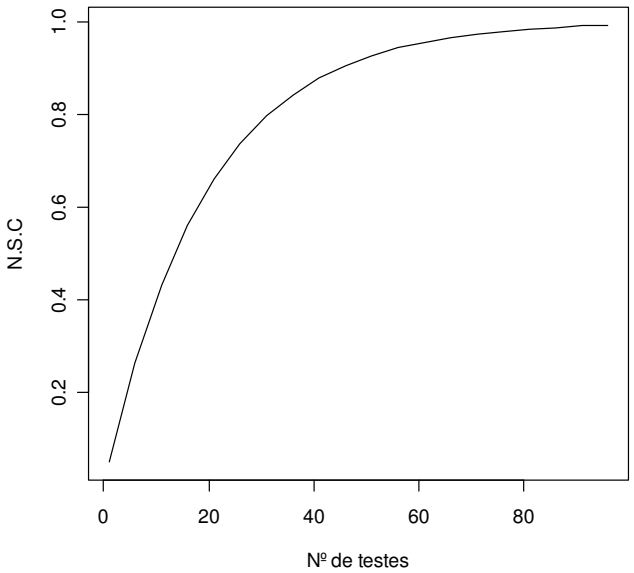


FIGURA 2.2: Nível de significância conjunto em função do número de testes realizados (m), considerando um nível de significância individual de 0,05.

Benjamini & Hochberg (1995) sugeriram um novo ponto de vista para o problema dos testes múltiplos. Eles propuseram controlar a FDR, em casos quando todos os m testes são independentes e por Benjamini & Yekutieli (2001) quando os testes são positivamente correlacionados.

2.5 False discovery rate

Segundo Silva (2001), para contornar alguns dos problemas encontrados nos testes de comparações multiplas, Benjamini & Hochberg (1995) propuseram controlar a FDR, 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.

TABELA 2.1: 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

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 2.1 resumem a situação apresentada. R e m são variáveis observaveis 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*, como definido pelos autores, é $P(V \geq 1)$.

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, $Q_e = E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{R}\right]$.

De acordo com Benjamini & Hochberg (1995), duas propriedades dessa razão de erros decorrem imediatamente:

1. se todas as hipóteses (H_0) forem verdadeiras, a FDR é equivalente a FWER. Neste caso, S = 0 e V = R. Portanto, se V = 0 $\Rightarrow$ Q = 0 e V $\geq$ 0 $\Rightarrow$ Q = 1, $P(V \geq 1) = E(Q) = Q_e$. Desse modo, controlar a FDR implica, grosso modo, em controlar a FWER;

2. quando $m_0 \leq m$, a FDR é menor que ou igual a FWER; nesse caso, se V > 0, decorre que $\frac{V}{R} \leq 1$ e tem-se: P (V$\geq$ 1)$\geq$ Q_e. Resulta daí que qualquer procedimento que controle a FWER também controla FDR. No entanto, se um procedimento controlar apenas a FDR, ele poderá ser menos restrito e um ganho em poder deverá ser esperado. Em particular, quanto maior o número de hipóteses

falsas, maior tende a ser S e, conseqüentemente, a diferença entre as razões de erros (FDR e FWER). Decorre, pois, que o potencial de aumento do poder é tanto maior quando maior for o número de hipóteses falsas.

Controlar a variável aleatória Q em cada teste é muito desejável. Porém, isso é impossível, pois, se $m_0 = m$ e se ao menos uma hipótese for rejeitada, $\frac{V}{R} = 1$ e Q não poderá ser controlada. Controlar $\frac{V}{R}|R > 0$ apresenta o mesmo problema, ou seja, é igual a um, na mesma situação; consequentemente, $E(\frac{V}{R}|R > 0)$ não pode ser controlada. A FDR é, então $P(R > 0)E(\frac{V}{R}|R > 0)$, que pode ser controlado, como demonstrado por Benjamini & Hochberg (1995).

O procedimento para determinar o ponto de corte em testes múltiplos, controlando a FDR, pode ser realizado do seguinte modo (Benjamini & 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:

$$q^* \geq \frac{mP_i}{i}.$$

Para melhor compreensão do procedimento, considere-se o exemplo apresentado por Benjamini & Hochberg (1995), em que foram realizados 15 testes de hipóteses, e que os $P_{(i)}s$, ordenados, tenham sido:
$0,0001;\ 0,0004;\ 0,0019;\ 0,0095;\ 0,0201;\ 0,0278\ 0,0298;\ 0,0344;\ 0,0495;$
$0,3240;\ 0,4262;\ 0,5719;\ 0,6528\ 0,7590;\ 1,0000$. Considere-se, ainda, que se deseje obter um nível de significância conjunto, $\alpha^* = 0,05$, que se equivale a um

FDR de 5%. Caso todas as hipóteses sejam verdadeiras, tem-se que

$$0,05 \geq \frac{15P_{(i)}}{i}$$

Assim, para $P_{(4)} = 0,0095 \Rightarrow 0,05 \geq 0,0356$;$P_{(5)} = 0,0201 \Rightarrow 0,05 \leq 0,0603$, portanto, deve-se rejeitar todas as hipóteses com valores p menores ou iguais a 0,0095.

O procedimento FDR tem sido aplicado em áreas diversas, quanto à análise de microarrays, para encontrar genes co-expressados (Tusher et al., 2001; Efron et al., 2001; Efron & Tibshirani, 2002; Dudoit et al., 2003) e nas aplicações psicológicas (Keselman et al., 1999), além de outras aplicações. A técnica da FDR vem sendo empregada com sucesso em microarrays, não porque haja muitos tratamentos, mas porque há muitas variáveis respostas, o que leva à necessidade de reduzir o número de conclusão de falsos positivos para poder manejar, em experimentos posteriores, apenas as variáveis mais promissoras.

3 MATERIAL E MÉTODOS

3.1 Material experimental

Os dados foram gentilmente cedidos pelo professor Alexandre Siqueira Guedes Coelho da Universidade Federal de Goiás-UFG e referem-se a estudos iniciais do projeto genoma funcional do eucalipto (Genolyptus). O objetivo deste estudo usando microarrays foi determinar o controle das vias genéticas da produção de lignina, que é uma substância que confere características estruturais à madeira. As sondas de 21413 cDNAs produzidas pelo programa foram montadas em uma plataforma Nimblegen.

Para o presente estudo, utilizaram -se dados de Eucaliptos em um experimento de microarrays, que visava entender melhor a variabilidade natural dos genes de eucaliptos, como por exemplo as diferenças entre espécies, diferenças 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 T/S (slide dentro de tratamento) é considerado como a estimativa do erro experimental para a presença do efeito de tratamento. Se for tomado o resíduo da análise como a estimativa do erro experimental, o poder do teste estará superestimado, sendo esta a análise padrão da maior parte dos softwares estatísticos. A estrutura experimental para cada uma das variáveis resposta é descrita na Tabela 3.1

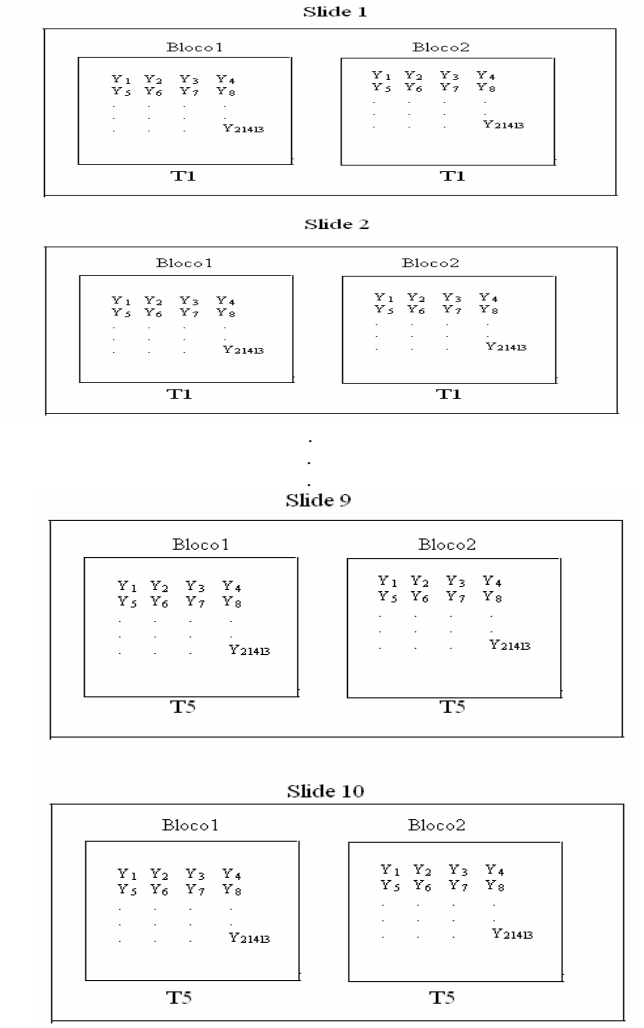
TABELA 3.1: Estrutura de fatores para o experimento com microarray.

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	Xilema
1	2	1	4	*grandis*	1	2	Xilema
1	3	2	5	*grandis*	2	3	Xilema
1	3	2	6	*grandis*	2	4	Xilema
1	4	3	7	*globulos*	3	5	Xilema
1	4	3	8	*globulos*	3	6	Xilema
1	5	4	9	*globulos*	4	7	Xilema
1	5	4	10	*globulos*	4	8	Xilema
2	1	1	1	*grandis*	1	1	Folha
2	1	1	2	*grandis*	1	2	Folha
2	2	1	3	*grandis*	1	1	Xilema
2	2	1	4	*grandis*	1	2	Xilema
2	3	2	5	*grandis*	2	3	Xilema
2	3	2	6	*grandis*	2	4	Xilema
2	4	3	7	*globulos*	3	5	Xilema
2	4	3	8	*globulos*	3	6	Xilema
2	5	4	9	*globulos*	4	7	Xilema
2	5	4	10	*globulos*	4	8	Xilema

grandis : *Eucalyptus grandis*

globulos: *Eucalyptus globulos*

FIGURA 3.1: Croqui da estrutura experimental.



3.2 Modelo de análise

Foram feitas análises de variância para cada sonda. Os cinco tratamentos utilizados foram:

T1: *Eucaliptus grandis*, clone 1, amostra de folha;

T2: *Eucaliptus grandis*, clone 1, amostra de tronco;

T3: *Eucaliptus grandis*, clone 2, amostra de folha;

T4: *Eucaliptus globulos*, clone 3, amostra de folha

T5: *Eucaliptus globulos*, clone 4, amostra de folha.

o modelo estatístico é $y_{ijk} = \mu + t_i + s_{(i)j} + e_{ijk}$

sendo, μ uma constante inerente a todas as observações;

t_i com i = 1,...5 é o efeito fixo de tratamentos;

$s_{(i)j}$ com j = 1,2 é o efeito do slide j dentro do tratamento i;

e_{ijk} com k = 1,2 é o erro experimental.

Os dados da Tabela 3.2 indicam que o teste F correto para a presença de efeito de tratamento é dado pela divisão do quadrado médio de T (tratamento) pelo quadrado médio de T/S (slide dentro de tratamento), que é a estimativa do erro experimental.

Para as sondas que resultaram significativas para efeito de tratamento, foram investigados os contrastes de interesse.

Contraste1: *E.grandis x E.globulos*.

2T1 + 2T2 + 2T3 - 3T4 - 3T5

Contraste2: Clone dentro de espécie

2:1 T1 + T2 - 2T3

2:1 T4 - T5

Contraste3: Folha versus xilema em *E.grandis*

2T1 - T2 - T3

TABELA 3.2: Esquema da análise de variância com os componentes de variância

FV	G.L.	QM	E(QM)	F
T	I	Q_1	(σ^2 + Kσ_s^2)+ JKσ_t^2	Q_1/Q_2
T/S	I(J-1)	Q_2	(σ^2 + Kσ_s^2)	
Resíduo	IJ(k-1)	Q_3	σ^2	

Segundo Montgomery (2001), a metodologia estatística usada, ANAVA, assume que os dados são normalmentes distribuídos, com variância constante e independente da média dos dados. Quando os dados não satisfazem tais hipóteses, é possível que o uso de transformação não lineares da variável resposta resolva o problema. Inicialmente foi feita uma análise considerando apenas a normalização original, que é o logaritmo da intensidade luminosa, um procedimento comum em experimentos de microarrays. Ao verificar através do teste Shapiro-Wilk que os dados não obedeciam a pressuposição de normalidade, foi feita uma transformação de Box-Cox e em seguida uma nova análise pôde ser realizada. Para a transformação de Box-Cox, considerou-se que λ pode variar no intervalo de $[-3, 3]$. Para aproximar do valor de λ, este intervalo foi dividido em 100 partes iguais. Como um experimento de microarrays mede níveis de expressões para milhares de genes simultaneamente, foi utilizado o critério FDR para garantir um nível global de 0.05 e 0.01.

3.3 Aplicação do critério FDR

A identificação de genes diferencialmente expressos é feita essencialmente por um teste de hipóteses, aplicado a cada sonda. 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. Para cada uma das hipóteses

$H_{0_1}, H_{0_2}, ..., H_{0_{21413}}$, obteve-se o correspondente valor p P_i, $i = 1, 2, ..., 21413$ (probabilidade, sob a hipótese H_0, de obter um valor ou igual ou superior ao obtido para a estatística teste num experimento futuro).

Em seguida, ordenaram-se os valores P_i. Sejam $P_{(1)}, P_{(2)}, ..., P_{(21413)}$ os valores p ordenados e $H_{0_1}, H_{0_2}, ..., H_{0_{21413}}$, as hipóteses correspondentes. Desejam-se um nível de significância conjuto α^*. Caso todas as hipóteses sejam verdadeiras, tem-se que

$$\alpha^* \geq \frac{21413 P_{(i)}}{i}. \tag{3.1}$$

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 de 3.1.

Os cálculos foram feitos utilizando-se o software R, versão, 2.3.1 (R Core Development Team, 2006).

4 RESULTADOS E DISCUSSÃO

Nesta seção encontram-se os resultados que estão organizados da seguinte forma: inicialmente, os resultados referente à análise em que os dados não foram transformados, considerando diferentes níveis de significância, com e sem a aplicação do procedimento FDR. Estes resultados foram apresentados em uma tabela única que, apesar de extensa, permite dar continuidade à seleção das melhores sondas para trabalhos posteriores. Em seguida, apresenta-se um quadro que dispõe os resultados referentes a cada contraste de interesse estudado. Por conveniência prática no estabelecimento de rotinas R que fizessem todas as análises de uma só vez, os mesmos procedimentos foram usados para os dados não tranformados e para os transformados. A apresentação dos resultados segue da mesma forma para os dados transformados e, por último, encontra-se uma figura refente à função de densidade da estatística de Shapiro-Wilk, aplicada respectivamente aos resíduos das análises dos dados não transformados e transformados, que permite discutir o efeito desta transformação.

4.1 Análise considerando os dados sem a transformação de Box-Cox

Para cada uma das 21.413 sondas, foi obtido uma análise de variância como a da tabela 4.1. Esta análise refere-se à sonda de número 4.

TABELA 4.1: Quadro-resumo da análise de variância referente à sonda de número 4.

FV	GL	SQ	QM	F_c	Pr>F
T	4	2,3422x10^{15}	5,8555x10^{14}	56,9103	0,00023
T/S	5	5,1444x10^{13}	1,0289x10^{13}	6,3204	
Resíduo	10	1,6279x10^{13}	1,6279x10^{12}		
Total	19	2,4099x10^{15}			

Nas Figuras 4.1 e 4.2 pode-se observar, inicialmente, o conjunto de todos os valores p e, a seguir, apenas os valores p significativos após a aplicação do procedimento FDR ($\alpha^* = 0,05$). A distribuição dos valores p, antes e depois da aplicação da FDR, indica que a análise não foi excessivamente rigorosa (muito poucos valores p muito baixos). O que se esperaria, neste caso, para a completa aleatoriedade é a distribuição uniforme dos valores p, (ver por exemplo: Storey, 2002). A pequena inclinação negativa nos gráficos é, portanto, indício de adequação da análise.

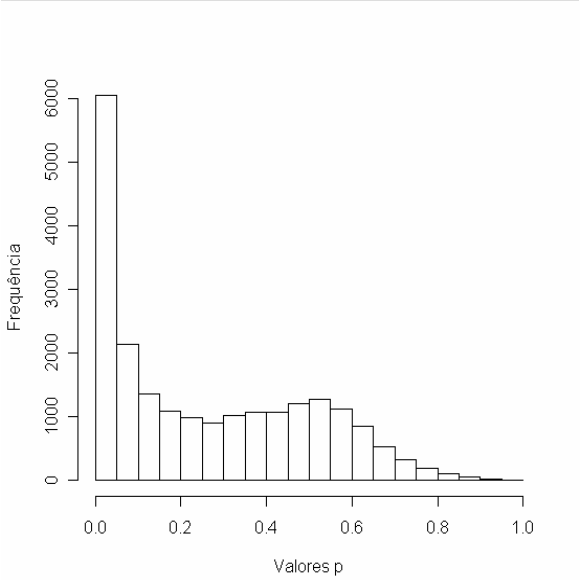


FIGURA 4.1: Histograma de densidade dos 21413 valores p referente às sondas analisadas com $\alpha = 5\%$.

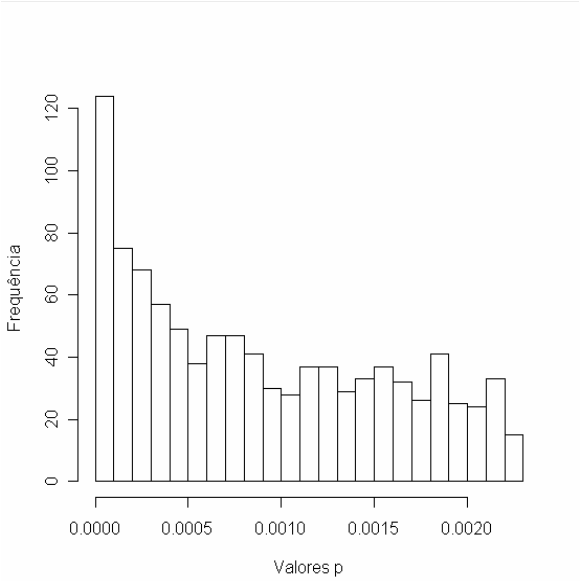


FIGURA 4.2: Histograma de densidade dos valores p que consideram diferenças significaticas no experimento, após a aplicação do FDR com $\alpha^* = 5\%$.

Na terceira coluna da Tabela 4.2 encontram-se os valores p observados no

experimento e, na quarta coluna, os valores p corrigidos segundo o procedimento FDR, para todas as sondas em estudo.

TABELA 4.2: A Lista dos valores p observados, ordenados do menor para o maior. A quarta coluna é o procedimento de controle do Benjamini & Hochberg (1995).

Sonda	Ordem	Valor-p observado	FDR(BH)Thresholds
07239**♣	1	3,87x10^{-9}	8,29x10^{-5}
16533**♣	2	4,56x10^{-9}	4,88x10^{-5}
07466**	3	5,35x10^{-9}	3,82x10^{-5}
07363**♣	4	3,64x10^{-8}	0,000195
03823**	5	2,04x10^{-7}	0,000872
13783**♣	6	4,63x10^{-7}	0,001651
03496**	7	4,76x10^{-7}	0,001456
14908**	8	5,26x10^{-7}	0,001410
20547**	9	7,96x10^{-7}	0,001893
13088**	10	8,19x10^{-7}	0,00175
01304**	11	1,11x10^{-6}	0,00216
00856**	12	1,12x10^{-6}	0,00200
07066**♣	13	2,13x10^{-6}	0,00351
15348**♣	14	2,14x10^{-6}	0,00327
02357**	15	2,22x10^{-6}	0,00317
11455**	16	2,82x10^{-6}	0,00377
19119**	17	3,54x10^{-6}	0,00445
20105**♣	18	3,69x10^{-6}	0,00439
00202**	19	3,85x10^{-6}	0,00434
20778**	20	4,33x10^{-6}	0,00463
01875**	21	4,67x10^{-6}	0,00475
08315**♣	22	4,85x10^{-6}	0,00472
01345**	23	4,89x10^{-6}	0,00455
03997**♣	24	5,47x10^{-6}	0,00488
20457**	25	6,07x10^{-6}	0,00519
08562**	26	6,17x10^{-6}	0,00508
07968**	27	6,53x10^{-6}	0,00517
15403**♣	28	6,64x10^{-6}	0,00508
14059**	29	7,08x10^{-6}	0,00523
19892**♣	30	7,12x10^{-6}	0,00508
04805**	31	7,24x10^{-6}	0,00500

continua...

TABELA 4.2: continuação

Sonda	Ordem	Valor-p observado	FDR(BH)Thresholds
08798**	32	8,29x10^{-6}	0,00555
18062**	33	8,40x10^{-6}	0,00545
18913**♣	34	8,87x10^{-6}	0,00558
02799**	35	9,72x10^{-6}	0,00595
20567**♣	36	1,06x10^{-5}	0,00630
02842**	37	1,07x10^{-5}	0,00620
13770**	38	1,08x10^{-5}	0,00608
13951**	39	1,38x10^{-6}	0,00758
03979**	40	1,40x10^{-5}	0,00747
02261**♣	41	1,42x10^{-5}	0,00980
06529**	42	1,42x10^{-5}	0 0,00725
16118**	43	1,59x10^{-5}	0,00810
11760**	44	1,69x10^{-5}	0,00826
06333**♣	45	1,70x10^{-5}	0,00808
06896**	46	1,79x10^{-5}	0,00855
03318**	47	1,83x10^{-5}	0,00853
04441**	48	1,85x10^{-5}	0,00825
02829**	49	1,94x10^{-5}	0,00847
19958**	50	1,97x10^{-5}	0,00843
00370**♣	51	1,97x10^{-5}	0,00828
15447**	52	2,15x10^{-5}	0,00884
15045**♣	53	2,20x10^{-5}	0,00890
09001**	54	2,27x10^{-5}	0,00900
00321**	55	2,29x10^{-5}	0,00891
20665**♣	56	2,36x10^{-5}	0,00902
08062**♣	57	2,40x10^{-5}	0,00901
07586**	58	2,66x10^{-5}	0,00980
02727**	59	2,67x10^{-5}	**0,00971**
18733*	60	2,97x10^{-5}	0.01059
00557*	61	3,07x10^{-5}	0,01077
02796*	62	3,08x10^{-5}	0,01063
.	.	.	.
.	.	.	.
.	.	.	.
10397*	970	0,002257	0,04981

continua...

TABELA 4.2: continuação

Sonda	Ordem	Valor-p observado	FDR(BH)Thresholds
20606*	971	0,002257	0,04978
06459*	972	0,002259	0,04978
06143*	973	0,002269	**0,04994**
.....	.....	.....	.....
12460	974	0,002280	0,05012
16612	975	0,002282	0,05012
00932	976	0,002288	0,05021
.	.	.	.
.	.	.	.
.	.	.	.
01568	2545	0,00997	0,0839
15973	2546	0,00999	0,0840
17695	2547	**0,01000**	0,0841
00212	2548	0,01001	0,0841
10123	2549	0,01001	0,0841
13185	2550	0,01002	0,0841
14933	2551	0,01002	0,0841
.	.	.	.
.	.	.	.
.	.	.	.
04515	6057	0,04995	0,1766
21125	6058	0,04996	0,1766
07635	6059	0,04999	0,1767
08964	6060	**0,04999**	0,1767
03325	6061	0,05006	0,1768
18638	6062	0,05006	0,1768
03900	6063	0,05009	0,1769
.	.	.	.
.	.	.	.
.	.	.	.
20130	21411	0,9803	0,9804
00116	21412	0,9864	0,9864
03501	21413	0,9904	0,9904

**significativo para $\alpha = 1\%$.

*significativo para $\alpha = 5\%$.

♣ Não normalidade

Observa-se, pelos dados da Tabela 4.2, 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 falsos positivos, segundo a metodologia empregada para corrigir o problema ocasionado pelo fato de diversas hipóteses estarem sendo testadas simultaneamente. Com a correção da FDR, foram rejeitadas apenas as hipóteses com os p-valores menores ou iguais a 0,002269.

Mesmo com a aplicação do 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^*, apenas os valores p menores ou iguais a 2,67x10^{-5} foram rejeitados, num total de 59 sondas com efeito de tratamento. Destas 59 sondas, apenas 40 foram considerados com efeito de tratamento, pois os dados referentes aos outros 19 não seguiam a pressuposição de normalidade, segundo o teste de Shapiro-Wilk. Para as 40 sondas com efeito de tratamento, investigam-se os contrastes de interesse descritos conforme na Tabela 4.3.

Para o contraste *E.grandis x E.globulos* constatou-se que 31 das 40 hipóteses testadas foram consideradas significativas.

TABELA 4.3: Número de sondas com contrastes significativos a 1% dentre as 40.

Contrastes	Sem o FDR	Com o FDR
E.grandis x E.glóbulos	31	31
Clone dentro da *E.grandis*	24	24
Clone dentro de *E.glóbulos*	13	12
Folha x Xilema em *E.grandis*	21	21

Para os contrastes clone dentro da *E. grandis* e clone dentro de *E. globulos* foram, respectivamente 24 e 13 hipóteses consideradas significativas, num total de

40 hipóteses testadas. O contraste folha x xilema em *E. grandis* foi significativo para 21 sondas em relação ao total testado. Aplicando-se o critério FDR, apenas o contraste clone dentro da *E. grandis* teve uma redução de 13 para 12 e os demais contrastes permaneceram com o mesmo número de valores p.

4.2 Análise considerando os dados com a transformação de Box-Cox

Para cada uma das 21.413 sondas obteve-se uma análise de variância como a que segue abaixo. Os dados da Tabela 4.4 referem-se à sonda de número 4.

TABELA 4.4: Quadro-resumo da análise de variância referente à sonda de número 4.

FV	GL	SQ	QM	F_c	Pr>F
T	4	2,9636	0,7409	86,1512	8,4740x10^{-5}
T/S	5	0,0431	0,0086	0,0086	
Resíduo	10	0,0011	0,0001		
Total	19	3,0078			

Assim como para os dados não transformados, houve um número muito grande de sondas (1529) em que o efeito dos tratamentos foi significativo ($\alpha^* = 0,05$). Sendo assim, considerou-se também necessário fazer uso do critério FDR e reduzir o nível de significância para $\alpha^* = 0,01$.

Nas Figuras 4.3 e 4.4 pode-se observar, inicialmente, o conjunto de todos os valores p e, a seguir apenas os valores p significativos após a aplicação do procedimento FDR ($\alpha^* = 0,05$). Estes gráficos também indicam adequação da análise.

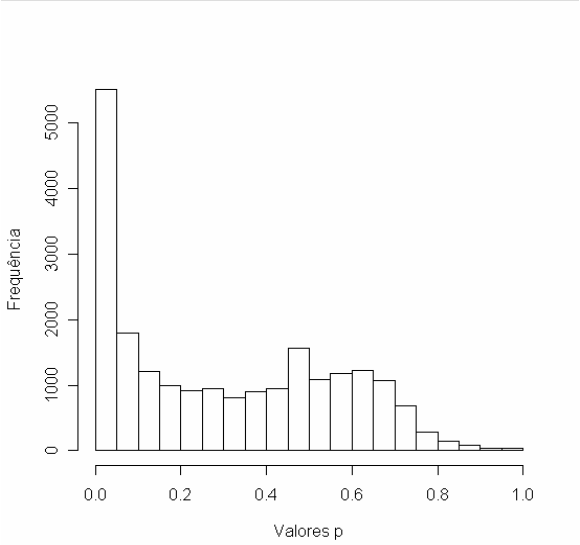


FIGURA 4.3: Histograma de densidade dos 21.413 valores p referentes às sondas analisadas, com $\alpha = 5\%$.

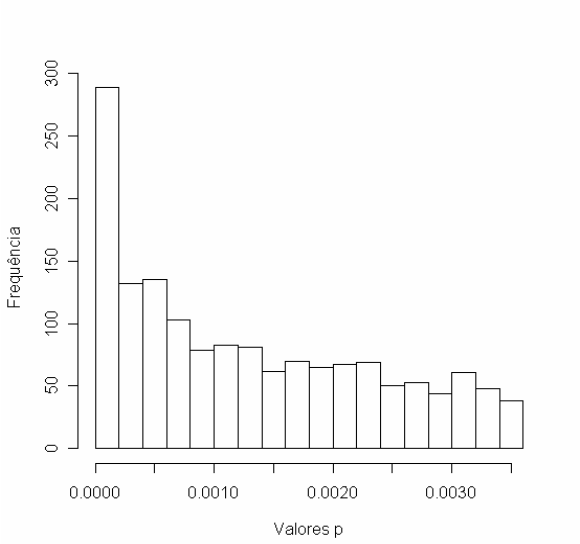


FIGURA 4.4: Histograma de densidade dos valores p que consideram diferenças significaticas no experimento após a aplicação do FDR, com $\alpha^* = 5\%$.

Na terceira coluna da Tabela 4.5 encontram-se os valores p observados no experimento e, na quarta coluna, os valores p segundo o procedimento FDR, refe-

rentes às sondas em estudo.

TABELA 4.5: A lista dos valores p observados, ordenados do menor para o maior. A quarta coluna é o procedimento de controle do Benjamini & Hochberg (1995).

Sonda	Ordem	Valor-p observado	FDR(BH)Thresholds
10700**♣	1	4,86x10^{-12}	1,04x10^{-7}
05757**♣	2	2,45x10^{-11}	2,62x10^{-7}
07819**♣	3	3,29x10^{-9}	2,35x10^{-5}
01926**♣	4	4,78x10^{-9}	2,56x10^{-5}
04748**♣	5	4,78x10^{-9}	2,05x10^{-5}
00125**♣	6	1,57x10^{-8}	5,59x10^{-5}
01325**♣	7	1,57x10^{-8}	4,79x10^{-5}
02127**♣	8	1,57x10^{-8}	4,19x10^{-5}
03581**♣	9	1,57x10^{-8}	3,73x10^{-5}
06338**♣	10	1,57x10^{-8}	3,35x10^{-5}
10624**♣	11	1,57x10^{-8}	3,05x10^{-5}
11557**♣	12	1,57x10^{-8}	2,79x10^{-5}
15389**♣	13	1,57x10^{-8}	2,58x10^{-5}
15543**♣	14	1,57x10^{-8}	2,39x10^{-5}
16382**♣	15	1,57x10^{-8}	2,23x10^{-5}
17698**♣	16	1,57x10^{-8}	2,09x10^{-5}
19119**♣	17	4,52x10^{-8}	5,69x10^{-5}
00566**♣	18	1,24x10^{-7}	0,00015
07499**♣	19	2,33x10^{-7}	0,00026
08324**♣	20	2,33x10^{-7}	0,00025
12043**♣	21	2,33x10^{-7}	0,00024
07171**♣	22	2,45x10^{-7}	0,00024
12853**♣	23	2,45x10^{-7}	0,00023
14908**♣	24	7,02x10^{-7}	0.00063
06394**♣	25	8,51x10^{-7}	0,00073
04968**♣	26	1,04x10^{-6}	0,00085
00856**	27	1,12x10^{-6}	0.00089
06044**♣	28	1,69x10^{-6}	0,00129
06084**♣	29	1,69x10^{-6}	0,00125
19992**♣	30	2,37x10^{-6}	0,00169
21066**♣	31	2,37x10^{-6}	0,00164

continua...

TABELA 4.5: continuação

Sonda	Ordem	Valor-p observado	FDR(BH)Thresholds
02538**♣	32	2,45x10^{-6}	0,00164
14999**♣	33	2,47x10^{-6}	0,00160
07466**	34	3,04x10^{-6}	0,00189
16533**	35	3,09x10^{-6}	0,00189
12005**♣	36	3,12x10^{-6}	0,00186
07254**	37	3,68x10^{-6}	0,00213
20778**	38	4,54x10^{-6}	0,00256
03496**	39	4,83x10^{-6}	0,00265
01875**	40	4,99x10^{-6}	0,00267
01595**♣	41	5,17x10^{-6}	0,00270
03648**	42	6,02x10^{-6}	0,00307
14059**	43	6,05x10^{-6}	0,00301
02702**♣	44	6,21x10^{-6}	0,00302
00202**	45	6,68x10^{-6}	0,00318
08562**	46	7,09x10^{-6}	0,00330
08262**♣	47	7,09x10^{-6}	0,00323
11455**	48	7,88x10^{-6}	0,00352
07901**	49	8,36x10^{-6}	0,00365
00201**	50	8,61x10^{-6}	0,00368
03318**	51	8,84x10^{-6}	0,00371
00094**	52	9,19x10^{-6}	0,00378
13951**	53	1,01x10^{-5}	0,00410
06376**♣	54	1,05x10^{-5}	0,00418
03823**	55	1,07x10^{-5}	0,00418
15120**♣	56	1,14x10^{-5}	0,00438
11760**	57	1,16x10^{-5}	0,00435
14626**	58	1,36x10^{-5}	0,00504
02842**♣	59	1,37x10^{-5}	0,00498
12496*	60	1,48x10^{-5}	0,00531
16725**♣	61	1,53x10^{-5}	0,00537
14656**	62	1,62x10^{-5}	0,00561
18062**	63	1,70x10^{-5}	0,00577
04833**	64	1,72x10^{-5}	0,00575
15427**	65	1,72x10^{-5}	0,00567
15348**	66	1,72x10^{-5}	0,00560

continua...

TABELA 4.5: continuação

Sonda	Ordem	Valor-p observado	FDR(BH)Thresholds
02468**	67	1,85x10^{-5}	0,00592
15419**	68	1,96x10^{-5}	0,00619
19958**	69	2,01x10^{-5}	0,00624
07256**♣	70	2,07x10^{-5}	0,00635
09515**	71	2,13x10^{-5}	0,00643
20302**	72	2,14x10^{-5}	0,00636
05384**	73	2,22x10^{-5}	0,00650
07372**	74	2,89x10^{-5}	0,00837
20547**	75	2,94x10^{-5}	0,00838
06181**	76	3,44x10^{-5}	0,00970
15086**	77	3,57x10^{-5}	0,00994
05594**	78	3,62x10^{-5}	0,00994
00037**	79	3,67x10^{-5}	**0,00995**
02727**	80	4,06x10^{-5}	0,01087
18733*	81	4,29x10^{-5}	0,01134673
.	.	.	.
.	.	.	.
.	.	.	.
03818*	1527	0,00356	0,04999
03956*	1528	0,00356	0,04997
03289*	1529	0,00357	**0,04998**
.....	.....	.....	.....
07212	1530	0,00358	0,0501
02861	1531	0,00359	0,0502
05006	1532	0,00359	0,0502
.	.	.	.
.	.	.	.
.	.	.	.
04413	2625	0,00997	0,0814

continuação...

TABELA 4.5: continuação

Sonda	Ordem	Valor-p observado	FDR(BH)Thresholds
20021	2626	0,00998	0,0814
17298	2627	**0,00999**	0,0814
21352	2628	0,01001	0,0815
00109	2629	0,01002	0,0816
08792	2630	0,01002	0,0816
.	.	.	.
.	.	.	.
.	.	.	.
18643	5514	0,04994	0,1766
20198	5515	0,04995	0,1766
17814	5516	0,04997	0,1767
06362	5517	0,05000	0,1767
11068	5518	**0,05000**	0,1768
11579	5519	0,05003	0,1768
03882	5520	0,05004	0,1769
.	.	.	.
.	.	.	.
.	.	.	.
02904	21410	0,9996	0,9997
03009	21411	0,9996	0,9997
03724	21412	0,9996	0,9996
06384	21413	0,9996	0,9996

**significativo para $\alpha = 1\%$.
*significativo para $\alpha = 5\%$.
♣ Não normalidade

Verifica-se, pelos dados da Tabela 4.5, 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, foram rejeitadas apenas as hipóteses com os valores p menores ou iguais a 0,00357, a 5% de significância. Do total resultante de 1.529 rejeições ($\alpha = 5\%$), apenas 79 foram também significativas com $\alpha = 1\%$.

4.3 Número de sondas combinando as duas análises

Apesar de ter sido feita análises com e sem transformação, não convém a comparação dos dois resultados, pois o objetivo é combinar as duas análises. 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, serão combinadas as 40 sondas da análise sem transformação com aquelas dentre as 79 que na análise inicial não atendiam a pressuposição de normalidade.

Verificou-se, pelo teste de Shapiro-Wilk, que os resíduos da análise dos dados de 1.113 sondas não atendiam à pressuposição de normalidade. A transformação de Box-Cox corrigiu o problema da não normalidade para 801 destas sondas, tendo em apenas 12 das 801 análises sido detectada significância para o efeito de tratamentos. Na primeira coluna da Tabela 4.5 encontram-se em negrito estas 12 sondas.

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 4.6.

TABELA 4.6: Número de sondas com contrastes significativos, a 1% de significância, dentre as 52 com efeito de tratamento significativo.

Contrastes	P-Valor Observado	FDR(BH)
E. grandis x E. globulos	52	43
Clone dentro de *E. grandis*	52	36
Clone dentro de *E. globulos*	52	13
Folha x xilema em *E. grandis*	52	33

Das 52 sondas que interferem na expressão diferencial entre tratamentos, 33 estão relacionadas a diferenças entre folhas e xilema, 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 é importante, pois a identificação de genes associados às sondas não podem ser realizados com um número muito grande de sondas candidatas. Supondo que um laboratório esteja interessado em investigar algumas sondas dentre as 52 com expressão diferencial para tratamentos, encontram-se em anexo, a análise completa para 15 dessas sondas.

A distribuição da estatística de Shapiro-Wilk revela que houve grande aumento no número de variáveis com estatística baixa (mais próxima da distribuição normal), apesar que, aumentou o número de sondas que deixaram de atender a pressuposição de normalidade.

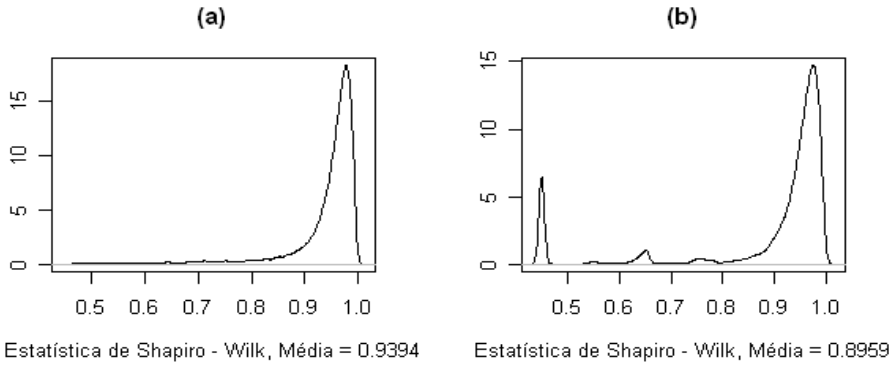


FIGURA 4.5: Estatística de Shapiro-Wilk para os dados não transformados (a) e os dados transformados (b)

As 52 sondas com expressão diferencial para tratamentos (resume as sondas importantes para os contrastes) e as sondas com contrastes significativos a 1% dentre as 52 com efeito de tratamento significativo estão relacionadas em anexo.

5 CONCLUSÕES

Os resultados obtidos pelo presente estudo permitem concluir que:

O critério FDR reduziu o número de variáveis em estudo, com segurança de se concentrar atenção nas mais relevantes.

A transformação de Box-Cox auxiliou a conferir validade para as análises de variáveis originalmente não transformadas.

6 REFERÊNCIAS BIBLIOGRÁFICAS

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistics Society**, London, v. 57, n. 1, p. 289-300, 1995.

BENJAMINI, Y.; YEKUTIELI, D. On the control of discovery rate in multiple testing under dependency. **The Annals of Statistics.** London, v. 29, n. 4, p. 1165-1188, 2001.

BOX, G. E. P.; COX, D. R. An analyis of transformations. **Journal of the Royal Statistical Society**, London, v. 26, n. 2, p. 42-56, 1964.

BRAZMA, A. et al. "Minimum information about a microarray experiment (MIAME): toward standards for microarray data". **Nature Genetics**, New York, v. 29, n. 4, p. 365-371, Dec. 2001.

CRISTO, E. B. **Métodos estatísticos na análise de experimentos de microarray.** 2003. 107 p. Dissertação (Mestrado em Estatística)-Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo.

COX, D. R.; REID, N. **The theory of the design of experiments**. London: Chapman & Hall, 2000. 340 p.

DUDOIT, S.; SHAFFER, J. C. Multiple hypothesis testing in microarrays experiments. **Statistical Science**, Berkeley, v. 18, p. 71-103, Aug. 2003.

EFRON, B.; TIBSHIRANI, R. Microarrays, empirical bayes methods, and false discovery rates. **Technical Report**, Chicago, v. 24, p. 201-217, July 2002.

EFRON, B.; TIBSHIRANI, R.; STOREY, J. D.; TUSHER, V. Empirical bayes analysis of a microarray experiment. **Journal of American Statistical Association**. New York, v. 96, p. 1151-1160, Dec. 2001.

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**. London, v. 18, p. 58-59, Mar. 1999.

KOKKO, A. **expression microarray technology as a tool in cancer research**. 2006. 73 p. Thesis (Doctoral science in technology) - Institute of Science, Helsinki University of Technology, Espoo, Finland.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. New York: J. Wiley, 1974. 564 p.

MONTGOMERY, D. C. **Design and analysis of experiments**. New York : J. Wiley, 1991. 649 p.

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. 2007.

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

SCHENA, M.; SHALON, D.; HELLER, R.; CHAI, A.; BROWN, P. O.; DAVIS, R. W. Parallel human genome analysis: microarray - based expression monitoring of 1000 genes. **Proceedings of the national academy of sciences**, New York, v. 93, p. 10614-10619, Mar. 1996.

SHAPIRO, S. S.; WILK, M. B. An Analysis of variance for normality: complete sample. **Biometrika**, London, v.52, p. 549-611, Jan. 1965.

STOREY, J. D. **False discovery rate theory and applications to DNA microarrays**. 2002. 114 p. Doctoral Dissertation of philosophy. Stanford university.

SILVA, H. D. **Aspectos biométricos da detecção de QTL'S ("Quantitative Trait Loci") em espécies cultivadas**. 2001. 166 p. Tese (Doutorado em Agronomia) - Escola Superior de Agronomia Luiz de Queiroz, Piracicaba, SP.

TUSHER, V. G.; TIBSHIRANI, R.; CHU,G. Significance analysis of microarray applied to the ionizing radiation response.**Proceedings of the national academy of sciences**, New York, v. 98, n. 5, p. 116-5121, 2001.

ANEXOS

ANEXO A **Páginas**

TABELA 1A: Relação das 52 sondas com expressão diferencial para tratamentos.

07466	03823	03496	14908	20547
13088	01304	00856	02357	11455
19119	00202	20778	01875	01345
20457	08562	07968	14059	04805
08798	18062	02799	02842	13951
03979	06529	16118	11760	06896
03318	04441	02829	19958	15447
09001	00321	07586	02727	13770
07254	07901	00201	00094	12496
04833	15427	02468	15419	07372
06181	16533			

TABELA 2A: Relação das 43 sondas com o contraste *E. grandis x E. globulos* significativos a 1% dentre as 52 com efeito de tratamento significativo.

07466	13088	18062	01304	13951
20547	03823	14908	02357	03496
07968	08798	00202	08562	02727
19119	04805	20457	11760	20778
01345	01875	16118	13770	06529
07586	15447	02799	00321	02829
06896	16533	07254	07901	00201
00094	12496	04833	15427	02468
15419	07372	06181		

TABELA 3A: Relação das 36 sondas com o contraste clone dentro de *E. grandis* significativos a 1% dentre as 52 com efeito de tratamento significativo.

16533	07466	03823	03496	14908
20547	13088	02357	11455	20778
01345	00202	16118	20457	04805
02799	07586	06529	11760	00321
02829	13770	15447	18062	16533
07254	07901	00201	00094	12496
04833	15427	02468	15419	07372
06181				

TABELA 4A: Relação das 13 sondas com o contraste clone dentro de *E. globulos* significativos a 1% dentre as 52 com efeito de tratamento significativo.

00856	14059	11455	19958	03979
09001	00441	19119	01875	03318
06896	2829	7254		

TABELA 5A: Relação das 33 sondas com o contraste Folha x xilema em *E. grandis* a 1% dentre as 52 com efeito de tratamento significativo.

07466	14908	20547	11455	13088
20778	20457	04805	13770	06529
11760	15447	00202	03496	03823
02357	19119	01345	08798	02799
04441	16533	07254	07901	00201
00094	12496	04833	15427	02468
15419	07372	06181		

TABELA 6A: Análise de variância referente à sonda 202, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$3{,}85\text{x}10^{-6}$	0,0043	$6{,}68\text{x}10^{-6}$	0,0032
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	$1{,}57\text{x}10^{-7}$	$4{,}88\text{x}10^{-7}$	$2{,}44\text{x}10^{-7}$	$6{,}65\text{x}10^{-7}$
Clone dentro da *E.grandis*	1	$6{,}55\text{x}10^{-7}$	$2{,}46\text{x}10^{-6}$	$1{,}24\text{x}10^{-6}$	$1{,}40\text{x}10^{-5}$
Clone dentro da *E.globulos*	1	0.4133	–	0.4028	–
Folha x xilema em *E.grandis*	1	$1{,}30\text{x}10^{-8}$	$3{,}84\text{x}10^{-8}$	$2{,}36\text{x}10^{-8}$	$2{,}33\text{x}10^{-7}$

TABELA 7A: Análise de variância referente à sonda 856, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$1{,}13\text{x}10^{-6}$	0,002	$1{,}13\text{x}10^{-6}$	0,001
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	0,999	–	0,999	–
Clone dentro da *E.grandis*	1	0,999	–	0,999	–
Clone dentro da *E.globulos*	1	$5{,}56\text{x}10^{-8}$	$3{,}28\text{x}10^{-6}$	$5{,}56\text{x}10^{-8}$	$1{,}46\text{x}10^{-6}$
Folha x xilema em *E.grandis*	1	0,910	–	0,999	–

TABELA 8A: Análise de variância referente à sonda 1875, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$4{,}67\text{x}10^{-6}$	0,0047	$4{,}99\text{x}10^{-6}$	0,0027
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	$5{,}55\text{x}10^{-7}$	$1{,}05\text{x}10^{-6}$	$9{,}07\text{x}10^{-7}$	$2{,}50\text{x}10^{-6}$
Clone dentro da *E.grandis*	1	0,9999	–	0,9999	–
Clone dentro da *E.globulos*	1	0,0002	0,0012	0,0001	0,0006
Folha x xilema em *E.grandis*	1	0,8382	–	0,9999	–

TABELA 9A: Análise de variância referente à sonda 2842, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$1{,}07\text{x}10^{-5}$	0,0062	$1{,}37\text{x}10^{-5}$	0,0050
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	0,9999	–	0,9999	–
Clone dentro da *E.grandis*	1	0,9999	–	0,9999	–
Clone dentro da *E.globulos*	1	$3{,}78\text{x}10^{-6}$	$3{,}72\text{x}10^{-5}$	$3{,}20\text{x}10^{-6}$	$2{,}53\text{x}10^{-5}$
Folha x xilema em *E.grandis*	1	0,9558	–	0,9999	–

TABELA 10A: Análise de variância referente à sonda 3496, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$4{,}76\text{x}10^{-7}$	0,0015	$4{,}83\text{x}10^{-6}$	0,0027
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	$8{,}75\text{x}10^{-8}$	$3{,}44\text{x}10^{-7}$	$1{,}71\text{x}10^{-6}$	$3{,}67\text{x}10^{-6}$
Clone dentro da *E.grandis*	1	$3{,}51\text{x}10^{-8}$	$3{,}45\text{x}10^{-7}$	$3{,}41\text{x}10^{-7}$	$8{,}98\text{x}10^{-6}$
Clone dentro da *E.globulos*	1	0,0632	–	0,01159	–
Folha x Xilema em *E.grandis*	1	$1{,}07\text{x}10^{-9}$	$1{,}05\text{x}10^{-8}$	$1{,}01\text{x}10^{-8}$	$1{,}98\text{x}10^{-7}$

TABELA 11A: Análise de variância referente à sonda 7466, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$5{,}35\text{x}10^{-9}$	$3{,}82\text{x}10^{-5}$	$3{,}04\text{x}10^{-6}$	0,00189
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	$4{,}84\text{x}10^{-10}$	$9{,}52\text{x}10^{-9}$	$1{,}09\text{x}10^{-7}$	$6{,}15\text{x}10^{-7}$
Clone dentro da *E.grandis*	1	$7{,}07\text{x}10^{-10}$	$1{,}39\text{x}10^{-8}$	$3{,}04\text{x}10^{-6}$	$1{,}41\text{x}10^{-5}$
Clone dentro da *E.globulos*	1	0,0092	0,042	0,0019	0,0083
Folha x xilema em *E.grandis*	1	$1{,}41\text{x}10^{-11}$	$2{,}77\text{x}10^{-10}$	$1{,}18\text{x}10^{-8}$	$2{,}84\text{x}10^{-6}$

TABELA 12A: Análise de variância referente à sonda 8562, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$6{,}17\text{x}10^{-6}$	0,0051	$7{,}09\text{x}10^{-6}$	0,0033
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	$1{,}70\text{x}10^{-7}$	$2{,}18\text{x}10^{-7}$	$7{,}09\text{x}10^{-6}$	$3{,}11\text{x}10^{-5}$
Clone dentro da *E.grandis*	1	0,9999	–	0,9999	–
Clone dentro da *E.globulos*	1	0,7184	–	0,7367	–
Folha x xilema em *E.grandis*	1	0,9999	–	0,9999	–

TABELA 13A: Análise de variância referente à sonda 11455, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$2{,}82\text{x}10^{-6}$	0,0038	$7{,}88\text{x}10^{-6}$	0,0035
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	0,9999	–	0,9999	–
Clone dentro da *E.grandis*	1	$3{,}25\text{x}10^{-7}$	$1{,}98\text{x}10^{-6}$	$6{,}48\text{x}10^{-6}$	$2{,}33\text{x}10^{-5}$
Clone dentro da *E.globulos*	1	$1{,}36\text{x}10^{-6}$	$2{,}67\text{x}10^{-5}$	$4{,}08\text{x}10^{-5}$	$2{,}48\text{x}10^{-4}$
Folha x xilema em *E.grandis*	1	$6{,}75\text{x}10^{-9}$	$3{,}62\text{x}10^{-8}$	$2{,}85\text{x}10^{-8}$	$2{,}05\text{x}10^{-7}$

TABELA 14A: Análise de variância referente à sonda 11760, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$1{,}69\text{x}10^{-5}$	0,0083	$1{,}16\text{x}10^{-5}$	0,0044
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	$4{,}97\text{x}10^{-7}$	$1{,}45\text{x}10^{-6}$	$5{,}72\text{x}10^{-7}$	$2{,}05\text{x}10^{-6}$
Clone dentro da *E.grandis*	1	$2{,}63\text{x}10^{-6}$	$5{,}93\text{x}10^{-6}$	$1{,}33\text{x}10^{-6}$	$1{,}05\text{x}10^{-5}$
Clone dentro da *E.globulos*	1	0,0171	–	0,0789	–
Folha x xilema em *E.grandis*	1	$6{,}65\text{x}10^{-8}$	$1{,}42\text{x}10^{-7}$	$3{,}66\text{x}10^{-8}$	$2{,}40\text{x}10^{-7}$

TABELA 15A: Análise de variância referente à sonda 13951, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$1{,}38\text{x}10^{-5}$	0,0076	$1{,}02\text{x}10^{-5}$	0,0041
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	$2{,}28\text{x}10^{-8}$	$1{,}68\text{x}10^{-7}$	$1{,}60\text{x}10^{-8}$	$1{,}58\text{x}10^{-7}$
Clone dentro da *E.grandis*	1	0,0115	–	0,0645	–
Clone dentro da *E.globulos*	1	0,0115	–	0,0277	–
Folha x xilema em *E.grandis*	1	0,9986	–	0,9866	

TABELA 16A: Análise de variância referente à sonda 14059, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$7{,}08\text{x}10^{-6}$	0,0052	$6{,}05\text{x}10^{-6}$	0,0030
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	0,9999	–	0,9999	–
Clone dentro da *E.grandis*	1	0,9999	–	0,9999	–
Clone dentro da *E.globulos*	1	$1{,}69\text{x}10^{-7}$	$4{,}98\text{x}10^{-6}$	$2{,}45\text{x}10^{-7}$	$3{,}22\text{x}10^{-6}$
Folha x xilema em *E.grandis*	1	0,9999	–	0,9999	–

TABELA 17A: Análise de variância referente à sonda 18062, com e sem transformação.

		Não Transf.		Transf.	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$8{,}40\text{x}10^{-6}$	0,0055	$1{,}70\text{x}10^{-5}$	0,0058
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	$1{,}49\text{x}10^{-8}$	$1{,}46\text{x}10^{-7}$	$2{,}76\text{x}10^{-8}$	$2{,}18\text{x}10^{-7}$
Clone dentro da *E.grandis*	1	0,0003	0,0004	0,0090	0,0284
Clone dentro da *E.globulos*	1	0,9439	–	0,9919	–
Folha x xilema em *E.grandis*	1	0,9991	–	0,9784	–

TABELA 18A: Análise de variância referente à sonda 19958, com e sem transformação.

		Não Transf.		Transf.	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$1{,}97\text{x}10^{-5}$	0,0084	$2{,}01\text{x}10^{-5}$	0,00624
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	0,9999	–	0,9999	–
Clone dentro da *E.grandis*	1	0,9999	–	0,9999	–
Clone dentro da *E.globulos*	1	$2{,}62\text{x}10^{-6}$	$3{,}86\text{x}10^{-5}$	$2{,}86\text{x}10^{-6}$	$2{,}51\text{x}10^{-5}$
Folha x xilema em *E.grandis*	1	0,9999	–	0,9999	–

TABELA 19A: Análise de variância referente à sonda 20547, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$7{,}96\text{x}10^{-7}$	0,0019	$2{,}94\text{x}10^{-5}$	0,0084
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	$5{,}25\text{x}10^{-8}$	$2{,}82\text{x}10^{-7}$	$5{,}89\text{x}10^{-7}$	$1{,}94\text{x}10^{-6}$
Clone dentro da *E.grandis*	1	$6{,}23\text{x}10^{-8}$	$4{,}08\text{x}10^{-7}$	$2{,}97\text{x}10^{-6}$	$1{,}47\text{x}10^{-5}$
Clone dentro da *E.globulos*	1	0,7657	–	0,9065	–
Folha x xilema em *E.grandis*	1	$2{,}27\text{x}10^{-9}$	$1{,}49\text{x}10^{-8}$	$1{,}39\text{x}10^{-7}$	$4{,}22\text{x}10^{-7}$

TABELA 20A: Análise de variância referente à sonda 20778, com e sem transformação.

		Não Transformado		Transformado	
FV	G.L.	P>F	FDR	P>F	FDR
T	4	$4{,}33\text{x}10^{-6}$	0,0046	$4{,}55\text{x}10^{-6}$	0,0026
T/S	5	–	–	–	–
Resíduo	10	–	–	–	–
E.grandis x E.globulos	1	$5{,}15\text{x}10^{-7}$	$1{,}05\text{x}10^{-6}$	$4{,}18\text{x}10^{-7}$	$1{,}57\text{x}10^{-6}$
Clone dentro da *E.grandis*	1	$3{,}60\text{x}10^{-7}$	$1{,}52\text{x}10^{-6}$	$4{,}52\text{x}10^{-7}$	$8{,}93\text{x}10^{-6}$
Clone dentro da *E.globulos*	1	0,7621	–	0,9794	–
Folha x xilema em *E.grandis*	1	$1{,}07\text{x}10^{-8}$	$3{,}95\text{x}10^{-8}$	$1{,}20\text{x}10^{-8}$	$1{,}58\text{x}10^{-7}$

ANEXO B **Páginas**

PROGRAMA 1B : Rotina para análise dos dados.

```
   #Programa para os dados não transformados

d     <- read.table(''Nimb.design.txt'', header=TRUE )
dados <- read.table(''NimbData.txt'', header=TRUE )
y     <- 0*rnorm(20)
attach(d)
T     <- factor(trat)
Ind   <- factor(indiv)
Ram   <- factor(ramet)
B     <- factor(bloco)
S     <- factor(slide)
n     <- length(dados[,1])/2
for ( i in 1:n )
{
for ( j in 1:10 )
{
y[j]    <- dados[ ( 2*(i-1)) +1, (j+2) ]
y[10+j] <- dados[ ( 2*(i-1)) +2, (j+2)]
}
modelo  <- lm(y ~ T/S)
summary <- summary(modelo)
anova   <- anova(modelo)
write(cbind(dados[((2*(i-1))+1),1],
anova[1,3]),''saida.txt'', append=TRUE)
write(cbind(dados[((2*(i-1))+1),1],
anova[2,3]),''saidaint.txt'', append=TRUE)
#saida é o QM de T
#saidaint é o QM de T/S
QM    <-cbind(saida$V2,saidaint$V2)
F     <- QM[,1]/QM[,2]
PV    <- 1-pf(F[,1],4,5)
PVord <-sort(PV[,1])
t     <- 0
t1    <- 0
j     <- 1
PVord1<- read.table(''PVord.txt'')
for (i in 1:n) {
if (((21413*PVord1[i,1])/i)< = 0.05) {
t[j]  <- PVord1[i,1]
```

```
t1[j] <- i
j <- j+1
}
}
Rotina para contrastes
d     <- read.table(''Nimb.design.txt'',h=T)
dados <- read.table(''NimbData.txt'',h=T)
y     <- 0*rnorm(20)
QM    <-read.table(''QM.txt'')
attach(d)
n  <- length(dados[,1])/2
a  <-c(2,2,2,2,2,2,-3,-3,-3,-3,2,2,2,2,2,2,-3,-3,-3,-3)
b  <-c(1,1,1,1,-2,-2,0,0,0,0,1,1,1,1,-2,-2,0,0,0,0)
c  <-c(0,0,0,0,0,0,1,1,-1,-1,0,0,0,0,0,0,1,1,-1,-1)
d  <-c(2,2,-1,-1,-1,-1,0,0,0,0,2,2,-1,-1,-1,-1,0,0,0,0)
matrix<-cbind(a,b,c,d)
for( i in 1:n)
{
for(j in 1:10)
{
y[j]    <- dados[(2*(i-1))+1,(j+2)]
y[10+j] <- dados[(2*(i-1))+2,(j+2)]
}
if (PV[i,1]<= 0.00226910840747907)
yestim  <- t(matrix)%*%y
ta      <- yestim[1,]/sqrt((sum(a^2)*QM[i,2])/4)
tb      <- yestim[2,]/sqrt((sum(b^2)*QM[i,2])/4)
tc      <- yestim[3,]/sqrt((sum(c^2)*QM[i,2])/4)
td      <- yestim[4,]/sqrt((sum(d^2)*QM[i,2])/4)
t       <- cbind(ta,tb,tc,td)
write(cbind(dados[((2*(i-1))+1),1],t),''t.txt'',
 append=TRUE)
}
}
Pa      <-1-pt(t[,2],5)
Pb      <-1-pt(t[,3],5)
Pc      <-1-pt(t[,4],5)
Pd      <-1-pt(t[,5],5)
Paord   <-sort(Pa[,1])
```

```
Pbord   <-sort(Pb[,1])
Pcord   <-sort(Pc[,1])
Pdord   <-sort(Pd[,1])
#Rotina para determinar o ponto de corte do FDR para (a)
Paord   <-read.table('Paord.txt')
ta      <- 0
ta1     <- 0
j       <- 1
Paord1  <- read.table(''Paord.txt'')
for (i in 1:n) {
if (((59*Paord1[i,1])/i)<=0.01) {
ta[j]  <- Paord1[i,1]
ta1[j] <- i
j      <- j+1
}
}
```

```
#Programa para os dados  transformados
d     <- read.table(''Nimb.design.txt'',h=T)"
dados <- read.table(''NimbData.txt'',h=T)"
y     <- 0*rnorm(20)"
attach(d)"
T    <- factor(trat)
Ind  <- factor(indiv)
Ram  <- factor(ramet)
B    <- factor(bloco)
S    <- factor(slide)
n    <- length(dados[,1])/2
for( i in 1:n)
{
for(j in 1:10)
{
y[j]    <- dados[(2*(i-1))+1,(j+2)]
y[10+j] <- dados[(2*(i-1))+2,(j+2)]
}
tr     <- boxcox(y ~ T/S, lambda = seq(-3, 3,length=100),
 plotit = FALSE)
er     <- cbind(tr$y,tr$x)
ind    <- which(max(tr$y)==er)
lambda <- er[ind,2]# estimatima do meu lambda
write(lambda,''lambda.txt'', append=TRUE)
}
for( i in 1:n)
{
for(j in 1:10)
{
y[j]   <- dados[(2*(i-1))+1,(j+2)]
y[10+j]<- dados[(2*(i-1))+2,(j+2)]
 }
if (lambda[i]!=0){
yt <- ((y^lambda[i])-1)/(lambda[i])
} else {
yt <- log(y)
}
modelo  <- lm(yt ~ T/S)
summary <- summary(modelo)
```

```
anova   <- anova(modelo)
write(cbind(dados[((2*(i-1))+1),1],anova[1,3]),
''saida.txt'', append=TRUE)
write(cbind(dados[((2*(i-1))+1),1],anova[2,3]),
''saidaint.txt'', append=TRUE)
}
#saida é o QM de T
#saidaint é o QM de T/S
QM    <-cbind(saida$V2,saidaint$V2)
F     <- QM[,1]/QM[,2]
PV    <- 1-pf(F[,1],4,5)
PVord <-sort(PV[,1])
t     <- 0
t1    <- 0
j     <- 1
PVord1<- read.table(''PVord.txt'')
for (i in 1:n) {
if (((21413*PVord1[i,1])/i)< = 0.05) {
t[j]  <- PVord1[i,1]
t1[j] <- i
j <- j+1
}
}
Rotina para contrastes
d     <- read.table(''Nimb.design.txt'',h=T)
dados <- read.table(''NimbData.txt'',h=T)
y     <- 0*rnorm(20)
fin   <-cbind(lambda,PV)
attach(d)
n     <- length(dados[,1])/2
a  <-c(2,2,2,2,2,2,-3,-3,-3,-3,2,2,2,2,2,2,-3,-3,-3,-3)
b  <-c(1,1,1,1,-2,-2,0,0,0,0,1,1,1,1,-2,-2,0,0,0,0)
c  <-c(0,0,0,0,0,0,1,1,-1,-1,0,0,0,0,0,0,1,1,-1,-1)
d  <-c(2,2,-1,-1,-1,-1,0,0,0,0,2,2,-1,-1,-1,-1,0,0,0,0)
matrix<-cbind(a,b,c,d)
for( i in 1:n)
{
for(j in 1:10)
{
```

```
y[j]    <- dados[(2*(i-1))+1,(j+2)]
y[10+j] <- dados[(2*(i-1))+2,(j+2)]
}
if (fin[i,2]<=0.00356911344348132) {
if (fin[i,1]!=0) {
yt   <- ((y\^fin[i,1])-1)/(fin[i,1])
} else  {
yt   <- log(y)
}
yestim <- t(matrix)%*%yt
ta     <- yestim[1,]/sqrt((sum(a^2)*QM[i,2])/4)
tb     <- yestim[2,]/sqrt((sum(b^2)*QM[i,2])/4)
tc     <- yestim[3,]/sqrt((sum(c^2)*QM[i,2])/4)
td     <- yestim[4,]/sqrt((sum(d^2)*QM[i,2])/4)
t      <- cbind(ta,tb,tc,td)
write(cbind(dados[((2*(i-1))+1),1],t),
''t.txt'', append=TRUE)
}
}
Pa    <- 1-pt(t[,2],5)
Pb    <- 1-pt(t[,3],5)
Pc    <- 1-pt(t[,4],5)
Pd    <- 1-pt(t[,5],5)
Paord <- sort(Pa[,1])
Pbord <- sort(Pb[,1])
Pcord <- sort(Pc[,1])
Pdord <- sort(Pd[,1])
# Rotina para o ponto de corte do FDR para o contraste (a)
Paord   <-read.table('Paord.txt')
ta      <- 0
ta1     <- 0
j       <- 1
Paord1  <- read.table(''Paord.txt'')
for (i in 1:n) {
if ((79*Paord1[i,1])/i)<=0.01) {
ta[j]  <- Paord1[i,1]
ta1[j] <- i
j      <- j+1
}
```

```
}
#Rotina para aplicação do teste de Shapiro - Wilk
d     <- read.table(''Nimb.design.txt'',h=T)
dados <- read.table(''NimbData.txt'',h=T)
y     <- 0*rnorm(20)
attach(d)
T    <- factor(trat)
Ind  <- factor(indiv)
Ram  <- factor(ramet)
B    <- factor(bloco)
S    <- factor(slide)
n    <- length(dados[,1])/2
for( i in 1:n)
{
for(j in 1:10)
{
y[j]    <- dados[(2*(i-1))+1,(j+2)]
y[10+j] <- dados[(2*(i-1))+2,(j+2)]
}
modelo <- lm( y ~ T/S)
norm   <- shapiro.test(modelo$residuals)
norm1[i,1] <- norm$p.value
write.table(norm1,''norm1.txt'')
}
# o teste de Shapiro - wilk para os dados
transformados segue o mesmo raciocínio.
```