**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

Estudo de expressão gênica em citros utilizando modelos lineares

Diógenes Ferreira Filho

Dissertação apresentada para obtenção do título de
Mestre em Ciências. Área de concentração: Estatística
e Experimentação Agronômica

**Piracicaba
2009**

Diógenes Ferreira Filho
Licenciatura em Matemática

Estudo de expressão gênica em citros utilizando modelos lineares

Orientadora:
Profa. Dra. ROSELI APARECIDA LEANDRO

Dissertação apresentada para obtenção do título de
Mestre em Ciências. Área de concentração: Estatística
e Experimentação Agronômica

**Piracicaba
2009**

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP**

Ferreira Filho, Diógenes

Estudo de expressão gênica em citros utilizando modelos lineares / Diógenes Ferreira
Filho. - - Piracicaba, 2009.
125 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2009.
Bibliografia.

1. DNA - Análise 2. Expressão gênica 3. Frutas cítricas - Experimentos 4. Inferência bayesiana 5. Modelos lineares I. Título

CDD 634.3
F383e

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

Aos meus pais, à minha namorada e aos meus irmãos.

DEDICO.

AGRADECIMENTOS

À Deus, por tudo que sou e que conquistei.

Aos meus amados pais Diógenes Ferreira e Aparecida Antonia Vendramel Ferreira por todo o apoio e suporte para eu chegar até aqui, por sempre terem acreditado em mim e por todo seu amor.

À minha namorada, meu amor Thaís Diane Borges dos Santos por sempre estar ao meu lado, mesmo nos momentos mais difíceis, por todo seu apoio, dedicação, paciência, compreensão e por todo seu amor.

Aos meus irmãos Guilherme Ferreira e Daniela Ferreira por todo seu apoio, carinho e por todo seu amor.

Ao meu sogro Aparecido Rodrigues dos Santos, à minha querida sogra Maria Aparecida Borges dos Santos (in memoriam) e ao meu cunhado Talles Eduardo Borges dos Santos por todo seu apoio, carinho e por todo seu amor.

À Professora Roseli Aparecia Leandro, pela experiência e conhecimento que me foram passados, pela paciência e pela amizade.

À Coordenação do Programa de Pós-graduação em Estatística e Experimentação Agronômica, pela oportunidade de realização do curso.

À Solange de Assis Paes Sabadin, secretária do Programa de Pós-graduação em Estatística e Experimentação Agronômica, pelo apoio e disponibilidade.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), pela concessão da bolsa de estudo.

Aos professores do Departamento de Ciências Exatas pelos conhecimentos transmitidos.

À Simone Guidetti-Gonzalez e à professora Helaine Carrer pela gentileza em ceder o banco de dados utilizado para as análises e pelo tempo e atenção cedidos.

À equipe da Divisão de Biblioteca e Documentação, pelo apoio e gentileza no ajuste das normas da dissertação.

Aos meus amigos e colegas de pós-graduação: Mateus Augusto Donegá, Tiago Flor de Santana, Cássio Dessotti, Epaminondas de Vasconcellos Couto, Marcelo Tavares de Lima, Rodrigo Rossetto Pescim, Keliny Martins de Melo Sousa, Marcelino Alves Rosa de Pascoa, Tiago Almeida de Oliveira, Ana Patrícia Bastos Peixoto, Ricardo Alves de Olinda, Renato Nunes Pereira, Elton Rafael Maurício da Silva Pereira, pelos momentos de parceria e alegria.

E a todas as pessoas que contribuíram de forma direta ou indireta para a realização deste trabalho.

SUMÁRIO

LISTA DE FIGURAS	11
LISTA DE TABELAS	13
RESUMO	15
ABSTRACT	17
1 INTRODUÇÃO.....	19
2 DESENVOLVIMENTO.....	21
2.1 Revisão Bibliográfica	21
2.1.1 Revisão da Técnica de <i>Microarray</i>	21
2.1.1.1 Algumas técnicas de <i>Microarray</i>	21
2.1.1.2 Fabricação de <i>Microarrays</i>	22
2.1.2 <i>Microarray</i> de cDNA	24
2.1.2.1 Fixação dos cDNAs nas lâminas	24
2.1.2.2 Extração de RNA e hibridização	24
2.1.2.3 Aquisição de Imagens.....	25
2.1.2.4 Análise das imagens	29
2.1.2.4.1 Endereçamento ou gradeamento.....	30
2.1.2.4.2 Segmentação do sinal (ou dos <i>spots</i>).....	32
2.1.2.4.3 Quantificação da intensidade.....	32
2.1.3 <i>Microarray</i> de uma cor.....	34
2.1.4 Delineamento.....	34
2.1.5 Visualização dos dados.....	36
2.1.6 Correção do background.....	37
2.1.7 Normalização dos dados.....	37
2.1.7.1 Normalização global.....	39
2.1.7.2 Normalização dependente da intensidade de expressão	40
2.1.7.3 Normalização dentro de grupo de impressão	41
2.1.7.4 Normalização dentro do <i>array</i> para escala	42
2.1.7.5 Normalização entre <i>arrays</i>	43
2.1.7.6 Normalização quantílica.....	43

2.1.8 Análise dos dados.....	45
2.1.8.1 Métodos para comparação de dois grupos	46
2.1.8.1.1 Análise usando <i>fold-change</i>	46
2.1.8.1.2 Teste t para dois grupos com variâncias iguais	48
2.1.8.1.3 Teste t para dois grupos com variâncias diferentes.....	49
2.1.8.1.4 Teste de Wilcoxon	49
2.1.8.2 Modelos Lineares	50
2.1.8.2.1 Modelo linear fixo envolvendo apenas efeito de tratamento.	50
2.1.8.2.2 Modelo de análise de variância (ANOVA).....	57
2.1.8.2.3 Modelo de efeitos mistos	59
2.1.8.3 Abordagem Bayesiana Empírica.....	59
2.1.8.4 Tipos de erro e testes Múltiplos	61
2.2 O <i>software</i> R	63
2.2.1 O projeto <i>Bioconductor</i> e o pacote <i>limma</i>	63
3 MATERIAIS E MÉTODOS	65
3.1 Material experimental	65
3.1.1 Delineamento do experimento	68
3.2 Métodos.....	69
3.2.1 Normalização dos dados	69
3.2.1.1 Normalização com a sonda <i>Overgo</i>	69
3.2.1.2 Normalização dependente da intensidade de expressão	69
3.2.1.3 Normalização entre <i>arrays</i> para correção de escala.....	70
3.2.2 Métodos de Análise.....	70
3.2.2.1 Método 1	71
3.2.2.2 Método 2	72
3.2.2.3 Método 3	73
3.2.2.4 Método 4	74
3.2.2.5 Método 5	75
3.2.2.6 Método 6	77
4 RESULTADOS E DISCUSSÃO	79
4.1 Normalização dos dados	79

4.2 Análise dos dados	81
4.2.1 Análise utilizando o método 1	81
4.2.2 Análise utilizando o método 2	82
4.2.3 Análise utilizando o método 3	83
4.2.4 Análise utilizando o método 4	84
4.2.5 Análise utilizando o método 5	86
4.2.6 Análise utilizando o método 6	88
4.3 Comparação dos resultados	89
5 CONCLUSÕES	99
REFERÊNCIAS	101
ANEXOS	105
ANEXO A – ENTRADA DOS DADOS NO R.....	107
ANEXO B – ROTINAS PARA PRÉ-PROCESSAMENTO DOS DADOS.....	108
ANEXO C – ROTINAS PARA A ANÁLISE COM O MÉTODO 1	110
ANEXO D – ROTINAS PARA A ANÁLISE COM O MÉTODO 2	112
ANEXO E – ROTINAS PARA A ANÁLISE COM O MÉTODO 3.....	114
ANEXO F – ROTINAS PARA A ANÁLISE COM O MÉTODO 4.....	117
ANEXO G – ROTINAS PARA A ANÁLISE COM O MÉTODO 5	120
ANEXO H – ROTINAS PARA A ANÁLISE COM O MÉTODO 6	124

LISTA DE FIGURAS

Figura 1 - Imagem de um <i>microarray</i>	21
Figura 2 - Técnicas de fabricação de <i>microarray</i> : (a) <i>Photolithography</i> , (b) <i>Mechanical microspotting</i> , (c) <i>Ink jetting</i>	23
Figura 3 - DNA.....	23
Figura 4 - Robô depositando os cDNAs nas lâminas de vidro.....	25
Figura 5 - Hibridização.....	26
Figura 6 - Leitura de um <i>microarray</i> pelo scanner a laser	26
Figura 7 - Imagens de um experimento de <i>microarray</i>	27
Figura 8 - <i>ScanAlyse</i>	28
Figura 9 - Imagem de <i>microarray</i>	28
Figura 10 - Técnica de cDNA <i>microarray</i>	29
Figura 11 - Termos usados na linguagem de processamento de imagens de <i>microarray</i>	30
Figura 12 - Gradeamento dos blocos.....	31
Figura 13 - Gradeamento dos <i>spots</i>	31
Figura 14 - Exemplos de segmentação do sinal	32
Figura 15 - Cálculo da intensidade de fluorescência do <i>spot</i>	33
Figura 16 - Esquema da técnica de <i>microarray</i> para <i>arrays</i> de duas cores e uma cor	34
Figura 17 - Exemplos de delineamentos para <i>microarrays</i> de duas cores	35
Figura 18 - Gráfico de dispersão de $Cy3 \times Cy5$	36
Figura 19 - Gráfico MA.....	37
Figura 20 - Possível problema em uma imagem de <i>microarray</i>	38
Figura 21 - Gráfico MA para o <i>array swirl.1</i> antes (a) e após (b) a normalização global. A linha azul marca os pontos onde $M = 0$	39
Figura 22 - Gráfico MA com curva de <i>Lowess</i> para o <i>array swirl.1</i> antes (a) e após (b) a normalização dependente da intensidade de expressão	40
Figura 23 - Agulhas de impressão (a) e lâmina com grupos de impressão (b)	41
Figura 24 - Grupos de impressão de um <i>microarray</i> , antes (a) e após (b) a normalização dentro de grupo de impressão, com correspondentes curvas de <i>lowess</i>	42

Figura 25 - <i>Boxplot</i> dos valores M de dois <i>microarrays</i> , antes (a) e após (b) a normalização de escala entre <i>arrays</i>	43
Figura 26 - Densidades das quatro hibridizações antes (a) e depois (b) da normalização quantílica	45
Figura 27 - <i>Boxplot</i> das quatro hibridizações antes (a) e depois (b) da normalização quantílica .	45
Figura 28 - Delineamento com três <i>arrays</i> , duas amostras e uma referência comum	52
Figura 29 - Delineamento em <i>loop</i> com três <i>microarrays</i> e três amostras	56
Figura 30 - Estrutura dos <i>arrays</i>	65
Figura 31 - Distribuição dos genes no <i>array</i>	66
Figura 32 - Imagens das hibridizações dos <i>arrays</i> com as amostras: (A) A ₁ com a amostra CsP, (B) A ₁ com a amostra CsS, (C) A ₂ com a amostra CsP e (D) A ₂ com a amostra CsS	67
Figura 33 - Delineamento do experimento de <i>macroarray</i>	68
Figura 34 - Gráficos de barras dos valores de intensidades de expressão das três repetições do Gene 1 no <i>array</i> A ₁ referentes à amostra CsS (folha de <i>Citrus sinensis</i> sadia) antes (a) e após (b) a normalização com a sonda <i>Overgo</i>	79
Figura 35 - Gráficos MA dos <i>arrays</i> A ₁ e A ₂ antes (A e B, respectivamente) e após (C e D) a normalização dentro do <i>array</i> utilizando regressão não paramétrica (<i>loess</i>).....	80
Figura 36 - <i>Boxplot</i> dos <i>arrays</i> A ₁ e A ₂ após a normalização dentro do <i>array</i> (a) e após a normalização entre <i>arrays</i> (b)	80
Figura 37 - Gráfico <i>Volcano</i> - Método 1.....	81
Figura 38 - Gráfico <i>Volcano</i> da análise utilizando o método 2	83
Figura 39 - Gráfico <i>Volcano</i> da análise utilizando o método 3	84
Figura 40 - Gráfico <i>Volcano</i> da análise utilizando o método 4	85
Figura 41 - Gráfico <i>Volcano</i> da análise utilizando o método 5	87
Figura 42 - Gráfico <i>Volcano</i> - Método 6.....	88
Figura 43 - Distribuição dos denominadores da estatística t ordinária (a) da estatística t moderada (b) para os 273 genes.....	95
Figura 44 - Distribuição dos valores de probabilidade obtidos das estatísticas t ordinárias (a) das estatísticas t moderadas (b)	95

LISTA DE TABELAS

Tabela 1 - Visão parcial de um banco de dados de <i>microarray</i> com o <i>background</i> corrigido	33
Tabela 2 - Visão parcial do conjunto de dados de expressão gênica de citros com <i>background</i> corrigido.....	68
Tabela 3 - Esquema de análise de variância	74
Tabela 4 - Esquema de análise de variância do modelo gênico	75
Tabela 5 - Esquema de análise de variância para o delineamento em parcelas subdivididas	77
Tabela 6 - <i>Rank</i> dos 10 genes com menores <i>p</i> -valores e <i>rank</i> dos 10 genes com maiores <i>fold-changes</i> para a análise com o Método 1	82
Tabela 7 - <i>Rank</i> dos 10 genes com menores <i>p</i> -valores e <i>rank</i> dos 10 genes com maiores <i>fold-changes</i> para a análise com o Método 2	82
Tabela 8 - <i>Rank</i> dos 10 genes com menores <i>p</i> -valores e <i>rank</i> dos 10 genes com maiores <i>fold-changes</i> para a análise com o Método 3	84
Tabela 9 - <i>Rank</i> dos 10 genes com menores <i>p</i> -valores e <i>rank</i> dos 10 genes com maiores <i>fold-changes</i> para a análise com o Método 4	85
Tabela 10 - Análise de Variância -Método 5.....	86
Tabela 11 - <i>Rank</i> dos 7 genes com menores <i>p</i> -valores e <i>rank</i> dos 7 genes com maiores <i>fold-changes</i> para a análise com o Método 5	87
Tabela 12 - Número de genes diferencialmente expressos simultaneamente	89
Tabela 13 - Significância da diferença de expressão para cada gene em cada método.....	90
Tabela 14 - Genes que se expressaram diferencialmente em pelo menos cinco métodos.....	94
Tabela 15 - Estrutura do banco de dados para leitura dos dados no R.....	107

RESUMO

Estudo de expressão gênica em citros utilizando modelos lineares

Neste trabalho apresenta-se uma revisão da metodologia de experimentos de *microarray* relativas a sua instalação e análise estatística dos dados obtidos. A seguir, aplica-se essa metodologia na análise de dados de expressão gênica em citros, gerados por um experimento de *macroarray*, utilizando modelos lineares de efeitos fixos considerando a inclusão ou não de diferentes efeitos e considerando ajustes de modelos para cada gene separadamente e para todos os genes simultaneamente. Os experimentos de *macroarray* são similares aos experimentos de *microarray*, porém utilizam um menor número de genes. Em geral, são utilizados devido a restrições econômicas. Devido ao fato de terem sido utilizados poucos *arrays* no experimento analisado neste trabalho foi utilizada uma abordagem bayesiana empírica que utiliza estimativas de variância mais estáveis e que leva em consideração a correlação entre as repetições do gene dentro do *array*. Também foi utilizado um método de análise não paramétrico para contornar o problema da falta de normalidade para alguns genes. Os resultados obtidos em cada um dos métodos de análise descritos foram então comparados.

Palavras-chave: *Microarray*; *Macroarray*; Citros; modelo linear fixo; método bayesiano empírico; FDR; *software R*; *Bioconductor*; pacote Limma

ABSTRACT

Gene expression study in citrus using linear models

This paper presents a review of the methodology of microarray experiments for its installation and statistical analysis of data obtained. Then this methodology is applied in data analysis of gene expression in citrus, generated by a macroarray experiment, using linear models with fixed effects considering the inclusion or exclusion of different effects and considering adjustments of models for each gene separately and for all genes simultaneously. The macroarray experiments are similar to the microarray experiments, but use a smaller number of genes. In general, are used due to economic restrictions. Because they have been used a few arrays in the experiment analyzed in this study it was used a empirical Bayes approach that uses estimates of variance more stable and that takes into account the correlation among replicates of the gene within array. A non parametric analysis method was also used to outline the problem of the non normality for some genes. The results obtained in each of the described methods of analysis were then compared.

Keywords: Microarray; Macroarray; Citrus; fixed linear model; empirical Bayes method; FDR; software R, Bioconductor, Limma package

1 INTRODUÇÃO

A comparação dos níveis de expressão de genes de diferentes tecidos ou de um mesmo tecido em diferentes condições pode levar ao entendimento de diversos fenômenos encontrados em um organismo.

Segundo Guidetti-Gonzalez (2009), o desenvolvimento de variedades resistentes a doenças pode ser feito utilizando ferramentas biotecnológicas, entretanto, é necessário identificar os genes envolvidos nas interações planta-patógeno. Isso pode ser feito identificando genes com níveis de expressão diferencial em diferentes condições biológicas (planta sadia versus planta doente).

Segundo Haddad (2007), os métodos de análise de expressão gênica, até pouco tempo atrás, avaliavam apenas poucos genes de cada vez: *Northern Blots*, *dot blots*, *RT-PCR*, entre outros. O surgimento da técnica de *microarray* de DNA representou um grande avanço na pesquisa de expressão gênica, pois, essa técnica permite mensurar os níveis de expressão de milhares de genes simultaneamente.

A análise de dados de *microarray* tem uma série de particularidades que devem ser consideradas, como por exemplo, etapas de pré-processamento como normalização dos dados, utilização de estatísticas de teste apropriadas e correção do nível de significância para testes múltiplos. Os métodos de análise de *microarray* mais comumente encontrados na literatura utilizam modelos lineares. Nesse tipo de análise geralmente é ajustado um modelo linear para cada gene.

Neste trabalho apresenta-se uma revisão da metodologia de experimentos de *microarray* relativas à sua instalação e análise estatística dos dados obtidos. A seguir, aplica-se essa metodologia na análise de dados de expressão gênica em citros, gerados por um experimento de *macroarray*, utilizando modelos lineares de efeitos fixos considerando a inclusão ou não de diferentes efeitos e considerando ajustes de modelos para cada gene separadamente e para todos os genes simultaneamente. Os experimentos de *macroarray* são similares aos experimentos de *microarray*, porém utiliza um menor número de genes, em geral, são utilizados devido a restrições econômicas. Devido ao fato de terem sido utilizados poucos *arrays* no experimento analisado neste trabalho foi utilizada uma abordagem bayesiana empírica que utiliza estimativas de variância mais estáveis e que leva em consideração a correlação entre as repetições do gene dentro do *array*. Também foi utilizado um método de análise não paramétrico para contornar o

problema da falta de normalidade para alguns genes. Os resultados obtidos em cada um dos métodos de análise descritos foram então comparados.

2 DESENVOLVIMENTO

2.1 Revisão Bibliográfica

2.1.1 Revisão da Técnica de *Microarray*

Segundo Haddad (2007), os métodos de análise de expressão gênica, até pouco tempo atrás, avaliavam apenas poucos genes de cada vez: *Northern Blots*, *dot blots*, *RT-PCR*, entre outros. O surgimento da técnica de *microarray* de DNA representou um grande avanço na pesquisa de expressão gênica, pois, essa técnica permite mensurar os níveis de expressão de milhares de genes simultaneamente.

Um *microarray* (Figura 1) é uma lâmina de vidro ou de náilon onde são fixados milhares de “genes”, cujas sequências são conhecidas, para posterior hibridização com cDNA de amostras de tecidos biológicos de interesse.

Figura 1 - Imagem de um *microarray*

A técnica de *microarray* possibilita a comparação dos níveis de expressão gênica entre diferentes tecidos de um mesmo organismo (tronco versus folha, por exemplo), mesmo tecido de um mesmo organismo sob diferentes condições (tecido doente versus tecido saudável, por exemplo), mesmo tecido em organismos diferentes (tecido de rato selvagem versus tecido de rato mutante, por exemplo); experimentos temporais, isto é, experimentos que são realizados em diferentes estágios de desenvolvimentos, dentre outros.

Nesse tipo de experimento são realizados milhares de testes simultâneos na mesma estrutura de unidades experimentais.

2.1.1.1 Algumas técnicas de *Microarray*

As técnicas de *microarray* diferem quanto à forma de fixação do material genético nas lâminas sendo o restante do processo semelhante para todas. Alguns exemplos desta técnica são:

- a) *Illumina bead array* (www.illumina.com);
- b) *Nylon Membrane* (www.schleicher-schuell.com);
- c) *Agilent: Long oligo Ink Jet* (www.home.agilent.com);
- d) *GeneChip Affymetrix* (www.affymetrix.com);
- e) *Microarray* de cDNA;
- f) *Microarrays* de proteínas e *Oligo Microarrays*;
- g) *Macroarrays* (*microarrays* com número menor de genes).

2.1.1.2 Fabricação de *Microarrays*

A construção dos *microarrays* pode ser feita por diferentes técnicas as quais diferem quanto à forma de fixação do material genético nas lâminas. Em Schena et al. (1998) são abordadas três técnicas de fabricação de *microarray* (Figura 2).

Os *microarrays Affymetrix* usam sequências curtas de DNA que são sintetizadas diretamente na superfície na qual deverá ocorrer a hibridação, utilizando um processo de síntese química em áreas fotoativadas (*Photolithography*).

Os *microarrays* de cadeia longa de DNA utilizam sequências sintetizadas em laboratório que são depositadas em uma lâmina em um arranjo de *spots* (local onde cada sequência específica é depositada), nos quais são realizadas as hibridações com as amostras de RNA extraído das populações celulares de interesse.

Segundo Lopes e Pais (2006) apesar da existência de *microarrays* em vários formatos, são usados com mais regularidade duas categorias:

1. *Microarrays* de cDNA compostos por cDNA ou oligonucleotídeos;
2. *Arrays* de grande densidade produzidos comercialmente que contêm oligonucleotídeos sintetizados.

O princípio que rege todos os experimentos de *microarray* é o da capacidade de uma sequência de nucleotídeos hibridizar com sua sequência complementar e formar uma sequência dupla de DNA. Segundo Junqueira e Carneiro (1997), cada nucleotídeo contém resíduos de uma molécula de ácido fosfórico, uma pentose e uma base púrica ou pirimídica, onde as bases púricas do DNA são a adenina e a guanina, e as bases pirimídicas são a timina e a citosina.

Figura 2 - Técnicas de fabricação de *microarray*: (a) *Photolithography*, (b) *Mechanical microspotting*, (c) *Ink jetting*

As bases situam-se dentro da hélice dupla, em planos paralelos entre si e perpendiculares ao eixo da hélice. Cada base de uma cadeia forma par com a base da cadeia complementar e, devido às dimensões das moléculas das bases, a timina faz par somente com a adenina, e a guanina faz par somente com a citosina, das cadeias complementares.

Na Figura 3 observa-se o DNA, composto por duas cadeias, e as ligações entre as bases: a timina faz par com a adenina enquanto a guanina faz par com a citosina.

Figura 3 – DNA

2.1.2 Microarray de cDNA

Segundo Esteves (2007) nos experimentos de *microarray* de cDNA é necessário um banco de clones de sequências expressas (*ESTs*) que são produzidas através da criação de DNA a partir de RNA mensageiro (mRNA) pelo processo de transcrição reversa. Esse procedimento dá origem a moléculas de DNA complementar (cDNA) que são clonados em bactérias. Os clones contendo as sequências específicas dos genes de interesse são selecionados e estas sequências são removidas dos clones, amplificadas por PCR e fixadas em posições específicas da lâmina (*spots*) através de agulhas de impressão controladas por um robô.

2.1.2.1 Fixação dos cDNAs nas lâminas

Nesta etapa os clones de cDNA os quais contém, cada um, as sequências de um único gene específico são fixados na lâmina em posições específicas conhecidas como “*spots*”. O interesse é medir o nível de expressão destes genes fixados na lâmina em relação a amostras biológicas. A fixação dos cDNAs é feita seguindo as etapas a seguir:

- a) **Seleção de clones:** Devem ser selecionados clones de cDNA. Esses geralmente são provenientes de algum banco de clones relacionado com algum projeto genoma específico.
- b) **Amplificação dos fragmentos por PCR:** Como em experimentos de *microarray* podem ser utilizadas várias lâminas a quantidade inicial de cDNAs pode não ser suficiente, e, pode-se fazer, por exemplo, a amplificação do material utilizando a técnica de amplificação por PCR (reação em cadeia por polimerase).
- c) **Fixação das cDNAs:** Os fragmentos selecionados são fixados nos *spots* das lâminas de vidro (*microarray*) por um robô. A Figura 4 ilustra a fixação dos cDNAs na lâmina de vidro.

2.1.2.2 Extração de RNA e hibridização

São extraídos o RNA mensageiro (mRNA) ou RNA total das duas populações celulares de interesse (Tratamento versus Controle, por exemplo).

Figura 4 - Robô depositando os cDNAs nas lâminas de vidro

A partir de cada amostra de RNA é produzido cDNA (DNA complementar) por um processo chamado de transcrição reversa. Os cDNAs de cada amostra são “tingidos” com os corantes fluorescentes Cy3 (representado pela cor verde) e Cy5 (representado pela cor vermelha) os quais serão excitados com comprimentos de onda diferentes. Os cDNAs da amostra controle geralmente são “tingidos” com Cy3 e os cDNAs da amostra tratamento são “tingidos” com Cy5.

Os cDNAs alvo de duas amostras distintas (uma com Cy3 e outra com Cy5) são misturados e hibridizados com os cDNAs na lâmina de vidro.

Nesse processo as sequências de cDNA das duas amostras, em contato com a lâmina de vidro, irão se “anelar” com suas sequências complementares dos cDNAs fixos na lâmina. Haverá competição entre as duas amostras.

Na Figura 5 pode-se observar que as duas amostras de cDNA marcadas são misturadas e depositadas sobre a lâmina de vidro e quando uma sequência marcada com corante fluorescente entra em contato com sua sequência complementar que está fixada no *array* elas se “anelam” (hibridizam) formando uma sequência de DNA de fita dupla contendo o corante fluorescente.

2.1.2.3 Aquisição de Imagens

Depois da hibridização a lâmina passa por uma etapa de lavagem para remover o excesso de cDNAs marcados com os corantes fluorescentes que não hibridizaram com os cDNAs fixados. Após a lavagem a lâmina é escaneada para obter a imagem digitalizada onde cada ponto representa a intensidade de fluorescência de cada *spot*, a qual é diretamente proporcional à intensidade de expressão do gene fixado no respectivo *spot*.

Figura 5 - Hibridização

Segundo Esteves (2002) dois tipos de scanners para *microarray* são:

1) Scanner CCD: As lâminas são expostas a uma luz que excita os corantes fluorescentes (Cy3 e Cy5) presentes nas amostras de interesse que foram utilizadas para a hibridização e uma câmera fotografa a imagem decorrente da emissão de intensidade de fluorescência;

2) Scanner a laser: Estes fazem uma varredura na lâmina com um raio laser no comprimento de onda específico de cada corante fluorescente digitalizando a imagem gerada. São feitas duas varreduras, uma no comprimento de onda do corante Cy3 e outra no comprimento de onda do corante Cy5. Os dados brutos resultantes de um experimento de *microarray* de duas cores são imagens monocromáticas, uma para cada corante, geralmente um arquivo com extensão tif.

A Figura 6 ilustra a leitura de um *microarray* pelo scanner a laser. Observe que esta figura é meramente ilustrativa, pois, como foi visto o *laser* não excita os dois corantes simultaneamente.

Figura 6 - Leitura de um *microarray* pelo scanner a laser

Na Figura 7 temos duas imagens de uma mesma lâmina: a Figura 7a corresponde à leitura pelo scanner no comprimento de onda que excita o corante Cy3 e a Figura 7b corresponde à leitura pelo scanner no comprimento de onda que excita o corante Cy5. Elas fazem parte do experimento Apo AI (DUDOIT et al., 2002).

Figura 7 - Imagens de um experimento de *microarray*

Posteriormente um pacote computacional de análise de imagens atribui cores a cada uma das imagens e mescla essas imagens dando origem a uma única imagem. Um exemplo de programa de análise de imagens de *microarray* de DNA é o *ScanAlyse* (EISEN, 1998). A Figura 8 gerada pelo *ScanAlyse* é a composição das imagens das Figuras 7a e 7b com a devida atribuição de cores verde para o corante Cy3 (Figura 7a) e vermelha para o corante Cy5 (Figura 7b).

Na Figura 9 pode-se observar uma imagem de *microarray* com boa qualidade dos sinais dos *spots*, na qual os pontos vermelhos são genes que estão se expressando na amostra marcada com Cy5, os pontos verdes são genes que estão se expressando na amostra marcada com Cy3, os pontos laranja são genes que estão se expressando nas duas amostras e os pontos pretos (sem sinal) são genes que não se expressaram em nenhuma das amostras.

Um resumo da técnica de cDNA *microarray* desde a preparação da lâmina até a obtenção da imagem é ilustrado a Figura 10 (FUJITA, 2007).

Figura 8 - *ScanAlyse* Figura 9 - Imagem de *microarray*

Figura 10 - Técnica de cDNA *microarray*

2.1.2.4 Análise das imagens

As imagens constituem os dados da análise de expressão gênica. Segundo Esteves (2002) alguns termos usados na linguagem de processamento de imagens de *microarray* (Figura 11) são:

- Foreground* ou região de sinal: Região ocupada pelo *spot*;
- Background*: Imagem de fundo da lâmina (região onde não se encontram os *spots*);
- Ruído: Falta de contribuição de sinal devido a moléculas que não se anelam com nenhuma molécula fluorescente;
- Artefato: Sinais inespecíficos decorrentes de sujeira na lâmina ou hibridização inespecífica que contaminam o *background*.

Figura 11 - Termos usados na linguagem de processamento de imagens de *microarray*

O processamento de imagens de *microarray* pode ser dividido em três partes:

1. Endereçamento ou gradeamento;
2. Segmentação do sinal (ou dos *spots*);
3. Quantificação da intensidade.

2.1.2.4.1 Endereçamento ou gradeamento

Nessa etapa são localizados os blocos e os *spots*. Assim devem ser feitos dois tipos de gradeamento:

- a) Blocos;
- b) Região de influência do *spot*.

As Figuras 12 e 13 ilustram o gradeamento dos blocos e gradeamento dos *spots* (Esteves, 2002).

Figura 12 - Gradeamento dos blocos

Figura 13 - Gradeamento dos spots

2.1.2.4.2 Segmentação do sinal (ou dos *spots*)

Verifica se o pixel pertence ao *foreground* ou ao *background*;

Alguns métodos de segmentação são:

- Segmentação de círculo fixo;
- Segmentação de círculo adaptativo;
- Segmentação por histograma;
- Segmentação por variação de intensidade.

Na Figura 14 pode-se observar dois tipos segmentação: segmentação por círculo fixo (Figura 14a) e a segmentação por variação de intensidade (Figura 14b).

Figura 14 - Exemplos de segmentação do sinal

2.1.2.4.3 Quantificação da intensidade

São feitos os cálculos das intensidades do *background* e do *foreground*. Na Figura 15 (CARAZZOLLE, 2008) podemos observar os cálculos da intensidade de fluorescência (vermelha e verde) do *spot*. *Target Median* é a mediana da intensidade dentro do círculo, e *Bkgd Median* é a mediana da intensidade do *background* (região fora do círculo, porém dentro do quadrado). A área é calculada como sendo o número de pixels dentro do círculo cuja intensidade seja maior que todos os pixels fora do círculo, mas dentro do quadrado. Assim a intensidade do *spot* é calculada por:

$$\text{Intensidade} = (\text{Target median} - \text{Bkgd median}) * \text{Area}.$$

Após a análise da imagem os resultado são transcritos em uma tabela contendo os níveis de expressão dos genes.

Figura 15 - Cálculo da intensidade de fluorescência do *spot*

Uma pequena parte do conjunto de dados do experimento Apo AI (DUDOIT et al., 2002) é apresentada na Tabela 1. Cada linha representa um gene e cada coluna representa um *microarray* marcado com um corante fluorescente específico (Cy3 ou Cy5).

Tabela 1 - Visão parcial de um banco de dados de *microarray* com o *background* corrigido

Gene	Corante	Array					
		1	1	2	2	3	3
		Cy3	Cy5	Cy3	Cy5	Cy3	Cy5
1		5592.58	2765.58	4749.89	1768.22	3709.31	1440.54
2		4746.38	2868.43	3088.12	2277.18	2259.42	1599.92
3		2108.48	1236.32	3669.53	1546.84	7774.40	2639.45
4		548.46	383.62	708.16	532.50	462.67	323.55
5		856.48	377.36	715.64	525.44	347.27	280.67
6		629.39	402.09	552.49	493.09	357.78	244.72

2.1.3 Microarray de uma cor

Segundo Pereira (2008), nos arranjos de uma cor, um procedimento de transcrição reversa é usado para produzir cDNA de fita dupla, que é transcrito e amplificado para cRNA marcado com biotina. O cRNA marcado com biotina é, então, fragmentado e hibridizado no *array*. Após a hibridização, o cRNA não hibridizado é removido do *array* e este 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 Figura 16 ilustra as técnicas de *microarray* de uma e duas cores.

Figura 16 - Esquema da técnica de *microarray* para *arrays* de duas cores e uma cor

2.1.4 Delineamento

A representação do delineamento de um experimento de *microarray* de duas cores será feita utilizando flechas para representar os *microarrays* (lâminas) e letras para representar as amostras biológicas (amostras de cDNA). A base da flecha indica que a amostra correspondente foi marcada com Cy3 e a ponta da flecha indica que a amostra correspondente foi marcada com Cy5.

Na Figura 17 são representados quatro exemplos de delineamento:

- a) Delineamento com comparação direta;
- b) Delineamento com Inversão de Corantes (*Dye Swap*);
- c) Delineamento com Referência;
- d) Delineamento em *Loop*.

Figura 17 - Exemplos de delineamentos para *microarrays* de duas cores

Considerando-se os experimentos representados na Figura 17, o experimento (a) tem um único *microarray* no qual a amostra A foi marcada com Cy3 (verde) e a amostra B foi marcada com Cy5 (vermelho). O experimento (b) tem dois *microarrays* sendo que no primeiro a amostra A foi marcada com Cy3 e a amostra B com Cy5 enquanto no segundo *microarray* foi feito o inverso, ou seja, a amostra A foi marcada com Cy5 e a amostra B com Cy3. O experimento (c) tem três *microarrays*, as amostras A e B são comparadas indiretamente através de uma referência comum. O experimento (d) tem três *microarrays* em que as três amostras A, B e C são comparadas.

O delineamento com inversão de corantes (*dye swap*) tem como objetivo eliminar o confundimento devido ao efeito dos corantes.

No delineamento com referência as amostras de interesse são comparadas indiretamente, através de uma referência comum. Esse tipo de delineamento tem a vantagem de outros tipos de amostras poderem ser adicionados posteriormente para serem comparadas com as amostras que já haviam sido comparadas, bastando utilizar a mesma referência comum utilizada nas amostras anteriores.

2.1.5 Visualização dos dados

Os dados de intensidade de fluorescência são tradicionalmente visualizados utilizando-se gráficos de dispersão. Os gráficos a seguir foram feitos com dados do experimento APO AI, (DUDOIT et al. 2002).

Os gráficos mais comuns são:

- a) Cy3 $\times$ Cy5 (Figura 18);
- b) Gráfico MA (Figura 19);

em que $M = \log_2(R) - \log_2(G)$ e $A = \frac{1}{2}[\log_2(R) + \log_2(G)]$. Note que R (*Red*) é a intensidade de Cy5 e G (*Green*) é a intensidade de Cy3.

As técnicas de normalização são aplicadas sob os dados transformados: $M = \log_2(R) - \log_2(G)$ e $A = \frac{1}{2}[\log_2(R) + \log_2(G)]$. Essas técnicas buscam reduzir a variação da técnica e tornar as intensidades comparáveis entre as diferentes hibridizações.

Figura 18 - Gráfico de dispersão de Cy3 $\times$ Cy5

Figura 19 - Gráfico MA

2.1.6 Correção do background

Segundo Esteves (2007) a intensidade do *background* é necessária para a correção dos valores de intensidade obtidos, pois, como não é possível obter exatamente os valores de ruído encontrados nos *spots* estes valores são estimados pela intensidade do *background* e assim o efeito deste ruído pode ser minimizado através da correção dos valores de intensidade do *foreground* pelos valores de intensidade do *background*. A correção do *background* é feita subtraindo as intensidades do *background* das intensidades do *foreground*.

2.1.7 Normalização dos dados

Experimentos de *microarray* têm muitas fontes de variação sistemática as quais podem afetar as medições dos níveis de expressão gênica. Para comparar *microarrays* devem ser removidas as fontes de variação de cada um deles. As técnicas de normalização são transformações dos dados que buscam remover essas fontes de variação.

A normalização pode ser feita:

- a) Dentro do *array*;
- b) Entre os *arrays*;

A normalização dentro do *array* é feita em cada um separadamente e pode envolver todos os genes do *array*, ou ser realizada por regiões do *array*. Esse tipo de normalização busca remover, por exemplo, o viés devido aos corantes dentro de cada *array* ou corrigir a locação.

A normalização entre *arrays* é utilizada para permitir a comparação entre *arrays* do experimento.

Algumas causas de variação de um experimento de *microarray* são devidas a fatores tais como:

- a) Diferenças na eficiência da incorporação dos corantes;
- b) Diferenças na quantidade de RNA inicial utilizado para marcação e hibridização;
- c) Diferenças de ajuste de parâmetros do scanner;
- d) Falhas na impressão dos cDNAs;
- e) Imprecisão de equipamentos;
- f) Procedimentos de localização e quantificação adotados pelos pacotes computacionais.

A Figura 20 apresenta uma imagem de *microarray* com uma mancha que aparenta ser a marca de um dedo na lâmina (ESTEVES, 2007).

Figura 20 - Possível problema em uma imagem de *microarray*

Existem vários tipos de normalização de dados de *microarray* entre os quais podemos citar:

- a) Normalização global;
- b) Normalização dependente da intensidade de expressão;

- c) Normalização dentro de grupo de impressão;
- d) Normalização dentro do *array* para escala;
- e) Normalização entre *arrays*;
- f) Normalização quantílica.

2.1.7.1 Normalização global

É feita para correção de locação. É aplicada quando existe uma relação constante entre os dois canais ($Cy5 = k.Cy3$). Segundo Yang et al. (2002) a correção é feita subtraindo-se do logaritmo da razão das medidas de expressão dos dois corantes de cada *spot*, a média destes para todo o slide:

$$\begin{aligned} \text{Log}_2(Cy5/Cy3)_i &= \text{Log}_2(Cy5/Cy3)_i - \mu \\ &= \text{Log}_2(Cy5/k.Cy3)_i \end{aligned}$$

Observa-se na Figura 21 que após a normalização global (Figura 21b) o centro da distribuição de todos os logaritmos da razão fica sendo zero.

Essa normalização tem a desvantagem de não considerar os efeitos da distribuição espacial dos *spots* nos *arrays* na expressão dos genes e por considerar que as diferenças são independentes da intensidade de expressão.

Figura 21 - Gráfico MA para o *array* swirl.1 antes (a) e após (b) a normalização global. A linha azul marca os pontos onde $M = 0$

2.1.7.2 Normalização dependente da intensidade de expressão

Vícios dependentes da intensidade de expressão na distribuição dos logaritmos das razões $M = \log_2(Cy5/Cy3)$.

Note que $A = \frac{1}{2}[\log_2(R) + \log_2(G)]$, ou seja, A é a soma dos logaritmos das intensidades de expressão de Cy3 e Cy5 dividido por dois, assim quando o eixo A é percorrido a intensidade de expressão varia, e pode ser notado no Gráfico MA (Figura 22a) que existe um vício nos valores M dependendo da intensidade de expressão (A). Após a normalização dependente da intensidade de expressão nota-se que esse vício desaparece (Figura 22b).

Utiliza-se um método não paramétrico para regressão de M em A, chamado *Lowess* (*Local-Weighted Regression and Smoothing Scatterplot*). Esse método adapta uma curva linear de regressão localmente ponderada, que é uma das técnicas estatísticas de suavização de gráficos de dispersão.

Figura 22 - Gráfico MA com curva de *Lowess* para o array swirl.1 antes (a) e após (b) a normalização dependente da intensidade de expressão

Tais métodos tentam estimar curvas não lineares capazes de resumir os dados observados. A partir da regressão, a correção é feita considerando-se

$$M^* = \log_2(Cy5^*/Cy3^*)_i = \log_2(Cy5/Cy3)_i - b(A) = \log_2(Cy5/k(A).Cy3)_i$$

em que $b(A)$ é o ajuste de *Lowess* para o gráfico MA e $Cy5^*$ e $Cy3^*$ representam os valores normalizados para os dois canais (YANG et al., 2002).

Segundo Coffey & Cofield (2006) os logaritmos dos valores normalizados $Cy3^*$ e $Cy5^*$ podem ser obtidos por

$$\log(Cy5^*) = A + 0,5M \text{ e } \log(Cy3^*) = A - 0,5M . \quad (1)$$

2.1.7.3 Normalização dentro de grupo de impressão

O processo de depositar os cDNAs nos respectivos pontos (*spots*) nas placas (*arrays*) é chamado de impressão. O conjunto de agulhas utilizado para a impressão pode ser uma fonte de variação sistemática no nível de expressão medido. Na Figura 23a pode-se observar um conjunto de agulhas utilizado na impressão dos cDNAs e na Figura 23b um *microarray* onde os grupos de impressão estão bem evidentes.

Figura 23 - Agulhas de impressão (a) e lâmina com grupos de impressão (b)

Este tipo de normalização é semelhante à normalização dependente da intensidade de expressão sendo, no entanto, realizado por grupo de impressão.

Na Figura 24 pode-se observar os gráficos MA de cada grupo de impressão de um *microarray* e suas respectivas curvas de *Lowess* antes (a) e depois (b) da normalização dentro de grupo de impressão.

Figura 24 - Grupos de impressão de um *microarray*, antes (a) e após (b) a normalização dentro de grupo de impressão, com correspondentes curvas de *lowess*

2.1.7.4 Normalização dentro do *array* para escala

Após o ajuste de locação, ainda é possível a utilização de modelos que assumem variâncias constantes (homocedasticidade) dentro de cada *array* ou entre os diversos *arrays* de um mesmo experimento (Esteves, 2007).

Segundo Stafford (2008) a normalização dentro do *array* para escala ajusta a diferença de escala entre os valores M dentro do *array*. Essa normalização assume que a maioria dos genes não são diferencialmente expressos entre diferentes amostras no banco de dados, portanto, a escala de seus valores M deve ser constante.

Seja M_s' o logaritmo da razão normalizado para escala, então $M_s' = M/s$, onde M é o logaritmo da razão de genes antes da normalização, e s é o fator de escala.

Segundo Yang et al. (2002) uma estimativa robusta de s é dada por

$$\hat{s}_i = \frac{MAD_i}{\sqrt[4]{\prod_{i=1}^I MAD_i}}$$

sendo MAD o desvio mediano absoluto (do inglês, *median absolut deviation*) definido por

$$MAD_i = \text{mediana}_i\{|M_i - \text{mediana}(M_i)|\}$$

em que M_i denota o logaritmo da razão de genes no *array* i antes da normalização e I o número de grupos de impressão.

2.1.7.5 Normalização entre *arrays*

Após a normalização dentro do *array* aplica-se a normalização entre *arrays* para correção dos efeitos de escala entre eles, permitindo assim compará-los.

O mesmo método utilizado para a correção de escala dentro do *array* pode ser diretamente estendido para esta situação.

A Figura 25 ilustra os *boxplots* de dois *microarrays*, em que pode-se observar que as distâncias interquartílicas dos *boxplots* se tornam mais semelhantes após a normalização de escala entre *arrays*.

Figura 25 - *Boxplot* dos valores M de dois *microarrays*, antes (a) e após (b) a normalização de escala entre *arrays*

2.1.7.6 Normalização quantílica

Segundo Bolstad et al (2003) objetivo da normalização quantílica é impor a mesma

distribuição empírica das intensidades para cada *array*. O gráfico *quantile-quantile* irá ter uma linha diagonal reta com inclinação de 45 graus se dois vetores de dados tem a mesma distribuição. Assim, plotando os quantis de dois vetores de dados um contra o outro e então projetando cada ponto sobre uma linha diagonal de 45 graus isso irá produzir uma transformação que proporciona a mesma distribuição para ambos os vetores de dados.

Em n dimensões, um *quantile-quantile plot* onde todos os vetores de dados têm a mesma distribuição teriam os pontos sobre a linha descrita pelo vetor $\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$. Essa extensão para n dimensões motivou o algoritmo da normalização quantílica.

Algoritmo da normalização quantílica:

1. Sejam n vetores de comprimento p , de X , de dimensão $p \times n$, onde cada *array* é uma coluna de X .
2. Organize em ordem crescente cada coluna de X e chame de X_s .
3. Calcule as médias das linhas de X_s e crie X'_s , um arranjo da mesma dimensão de X , mas onde todos os valores de cada linha são iguais as médias da respectiva linha em X_s .
4. Calcule X_n reorganizando cada coluna de X'_s para ter a mesma ordem do correspondente vetor de entrada.

Essa normalização é geralmente aplicada em *microarrays* de um canal, onde cada *array* é hibridizado com apenas uma amostra biológica. Para exemplificar essa normalização considere um experimento de *macroarray* com dois *arrays* onde cada um foi hibridizado duas vezes, uma vez com o tratamento e outra vez com o controle. O resultado desse experimento foram quatro imagens referentes a cada uma das quatro hibridizações (duas em cada um dos *arrays*).

Observe na Figura 26 que após a normalização quantílica as quatro hibridizações passaram a ter a mesma distribuição e os seus gráficos ficaram sobrepostos. Pode-se observar na Figura 27 que as hibridizações tornaram-se comparáveis observando-se seus boxplots antes e depois da normalização quantílica.

Figura 26 - Densidades das quatro hibridizações antes (a) e depois (b) da normalização quantílica

Figura 27 - Boxplot das quatro hibridizações antes (a) e depois (b) da normalização quantílica

2.1.8 Análise dos dados

Nesta etapa, é possível seguir diferentes abordagens matemáticas e estatísticas, dependendo das questões biológicas envolvidas.

Um dos principais objetivos da análise de *microarray* é identificar genes que apresentem expressão diferencial em algum dos grupos biológicos estudados. Esse tipo de análise é

conhecido como busca de genes diferencialmente expressos. Embora muitos estudos de *microarray* façam a comparação dos níveis de expressão dos genes em dois grupos biológicos distintos é possível a comparação entre vários grupos. Comparações entre dois grupos geralmente são feitas utilizando-se testes t, teste de Mann-Withney, ou modelos lineares. Quando mais de dois grupos biológicos são estudados, modelos mais elaborados devem ser utilizados, como modelos de análise de variância (ANOVA).

2.1.8.1 Métodos para comparação de dois grupos

A análise de dois grupos biológicos era realizada com base na razão de expressão (*fold-change*), ou seja, a razão entre os valores de intensidade normalizados dos dois grupos. Era então fixado então um ponto de corte arbitrário para decidir se o gene em questão era diferencialmente expresso ou não. Posteriormente passaram a ser utilizados métodos estatísticos como o teste t ou o teste de Mann-Withney, que é equivalente ao teste da soma de postos de Wilcoxon. O teste de Wilcoxon é um teste não paramétrico que deve ser utilizado caso não seja satisfeita a pressuposição de normalidade.

Apesar de pressuposições de modelos paramétricos raramente serem atendidas em dados de *microarray*, segundo Smyth et al. (2008) os p -valores podem ser utilizados para classificar os genes em um *rank* mesmo na presença de grandes desvios das pressuposições.

2.1.8.1.1 Análise usando *fold-change*

Seja a *fold-change* para o gene g definida como:

$$FC_g = \frac{x_{2g}}{x_{1g}}$$

em que x_{1g} e x_{2g} são os valores de intensidade do gene g nos grupos 1 e 2, respectivamente. Observe que $M_g = \log_2(x_{2g}) - \log_2(x_{1g})$ é igual ao logaritmo na base 2 da *fold-change* para o gene g . O logaritmo da *fold-change* para o gene g será denotado por $\log FC_g$.

Para comparação entre os dois grupos considere um ponto de corte arbitrário, então todos

os genes que tiverem valor absoluto de $\log FC$ maior que o ponto de corte serão considerados diferencialmente expressos. Observe que $\log FC_g$ positivo significa que o gene g está se expressando no grupo 2, se $\log FC_g$ é negativo significa que o gene g está se expressando no grupo 1. Por exemplo, seja $\log FC_g = 1$, segue que

$$\log_2 \left(\frac{x_{2g}}{x_{1g}} \right) = 1 \Rightarrow \frac{x_{2g}}{x_{1g}} = 2^1 \Rightarrow x_{2g} = 2x_{1g}$$

ou seja, o gene g está se expressando duas vezes mais no grupo 2 do que no grupo 1. E se $\log FC_g = -1$, segue que

$$\log_2 \left(\frac{x_{2g}}{x_{1g}} \right) = -1 \Rightarrow \frac{x_{2g}}{x_{1g}} = 2^{-1} \Rightarrow x_{2g} = \frac{x_{1g}}{2}$$

ou seja, o gene g está se expressando duas vezes menos no grupo 2 do que no grupo 1. Se $\log FC_g = 0$, segue que

$$\log_2 \left(\frac{x_{2g}}{x_{1g}} \right) = 0 \Rightarrow \frac{x_{2g}}{x_{1g}} = 2^0 \Rightarrow x_{2g} = x_{1g}$$

ou seja, o gene g está se expressando igualmente nos dois grupos.

No caso do gene g ser replicado dentro do *array*, define-se a média dos logaritmos dos valores de intensidade nas duas condições para o gene g como

$$\bar{y}_{1g} = \frac{\sum_{i=1}^{n_1} y_{1ig}}{n_1} \text{ e } \bar{y}_{2g} = \frac{\sum_{i=1}^{n_2} y_{2ig}}{n_2}$$

em que y_1 e y_2 são os valores do logaritmo na base 2 dos valores das intensidade das condições 1 e 2, e n_1 e n_2 são os tamanhos amostrais das duas condições. Neste caso é definido o $\log FC_g$ médio como:

$$\log FC_g = \bar{y}_{2g} - \bar{y}_{1g}, \quad g = 1, \dots, G \quad (2)$$

Deve-se observar, entretanto, que a *fold-change* não é um método estatístico e, conseqüentemente, não são definidas propriedades estatísticas para testes de hipóteses e não

podem ser calculados intervalos de confiança ou p -valores para as comparações. Devem ser utilizados então métodos estatísticos para identificar genes diferencialmente expressos como, por exemplo, testes t .

2.1.8.1.2 Teste t para dois grupos com variâncias iguais

Sejam μ_{1g} e μ_{2g} as médias verdadeiras e desconhecidas dos níveis de expressão para o gene g nas duas condições. O teste t para dois grupos com variâncias iguais considera a hipótese bilateral de médias iguais dos níveis de expressão para cada gene:

$$H_0 : \mu_{1g} = \mu_{2g} \quad \text{versus} \quad H_A : \mu_{1g} \neq \mu_{2g}$$

Sejam $\bar{y}_{1g}$ e $\bar{y}_{2g}$ as médias amostrais do logaritmo na base 2 dos valores observados nas duas condições para o gene g ($g = 1, \dots, G$), e seja a estimativa de variância comum dos valores do logaritmo na base 2 dos níveis de expressão do gene g definida como

$$s_g^2 = \frac{\sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ijg} - \bar{y}_{ig})^2}{n_1 + n_2 - 2}$$

Segue que a estatística t para comparar os níveis de expressão gênica de dois grupos com variâncias iguais é dada por:

$$t_g = \frac{\bar{y}_{2g} - \bar{y}_{1g}}{\sqrt{s_g^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (3)$$

Observe que o numerador da estatística t (eq. (3)) é similar ao $\log FC_g$ (eq. (2)). A hipótese testada é igual à comparação usando a *fold-change*, porém leva em conta a variabilidade do gene no cálculo da estatística t . Quando dividimos cada diferença pela sua estimativa de variância estamos “padronizando” a *fold-change* (COFFEY; COFIELD, 2006). O teste t verifica mudanças relativas no nível de expressão, assim ele irá identificar genes com pequena expressão diferencial se essas diferenças ocorrerem com pouca variação. Tais diferenças não podem ser encontradas utilizando o método da *fold-change*.

2.1.8.1.3 Teste t para dois grupos com variâncias diferentes

O teste t descrito acima é valido sob a suposição de que os níveis de expressão gênica têm variâncias iguais nas duas condições. Quando as variâncias nas duas condições não são iguais, uma versão alternativa do teste t é produzida levando em conta essa diferença.

Para o gene g ($g = 1, \dots, G$), seja a média amostral dos logaritmos dos valores de intensidade nas duas condições, $\bar{y}_{1g}$ e $\bar{y}_{2g}$, como anteriormente, e seja a estimativa de variância dos níveis de expressão do gene g (em logaritmo na base 2) para os dois grupos definida como

$$s_{1g}^2 = \frac{\sum_{i=1}^{n_1} (y_{1ig} - \bar{y}_{1g})^2}{n_1 - 1} \quad \text{e} \quad s_{2g}^2 = \frac{\sum_{i=1}^{n_2} (y_{2ig} - \bar{y}_{2g})^2}{n_2 - 1}$$

Segue que a estatística t para comparar níveis de expressão gênica de dois grupos com variâncias iguais é

$$t_g = \frac{\bar{y}_{2g} - \bar{y}_{1g}}{\sqrt{\frac{s_{1g}^2}{n_1} + \frac{s_{2g}^2}{n_2}}} \quad (4)$$

Ao contrário da estatística da eq. (3), a distribuição da estatística da eq. (4) tem uma distribuição t apenas aproximada (COFFEY; COFIELD, 2006). Segundo Borges e Ferreira (1999) pode ser utilizada a aproximação de Satterthwaite para estimar número apropriado de graus de liberdade, onde os números de graus de liberdade são calculados por

$$v = \frac{\left(\frac{s_{1g}^2}{n_1} + \frac{s_{2g}^2}{n_2} \right)^2}{\frac{\left(\frac{s_{1g}^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_{2g}^2}{n_2} \right)^2}{n_2 - 1}}$$

2.1.8.1.4 Teste de Wilcoxon

O teste da soma de postos de Wilcoxon é um teste não paramétrico que pode ser utilizado

no caso de os dados não seguirem uma distribuição normal. Assumindo que a distribuição das duas amostras tem o mesmo formato, mas talvez diferentes médias, então o teste de Wilcoxon pode ser utilizado para testar se as médias são iguais. Este teste utiliza os postos das observações ao invés dos valores para comparar duas amostras.

Se os dados forem normalmente distribuídos o teste t deve ser utilizado, pois ele garante mais poder do que o teste de Wilcoxon quando os dados são normalmente distribuídos.

2.1.8.2 Modelos Lineares

O procedimento de fazer comparações dos logaritmos das razões das amostras de RNA utilizando o teste t considera a variação dentro do gene e isso o torna substancialmente melhor do que a análise apenas pela *fold-change*. Segundo Kerr (2003) esse procedimento pode ser reformulado como um modelo linear, e essa reformulação têm a vantagem de considerar outros delineamentos experimentais mais gerais. Outra abordagem de modelos lineares, como modelos de análise de variância, leva em consideração as causas de variabilidade tais como possíveis variações na resposta devido ao corante ou ao *array*, e, extensões desses modelos consideram ainda efeitos aleatórios (COFFEY; COFIELD, 2006).

O modelo linear apropriado para cada situação depende do tipo de efeitos incluídos no modelo. O termo efeito fixo é aplicado quando estes efeitos são os únicos de interesse para o pesquisador e, caso o experimento fosse repetido, estes efeitos seriam repetidos exatamente iguais no experimento atual. Nesse tipo de experimento geralmente o interesse é comparar as médias da variável resposta entre vários níveis de um efeito fixo. O termo efeito aleatório é aplicado para efeitos que representam uma amostra aleatória de uma população. Em experimentos envolvendo efeitos aleatórios, o interesse é geralmente em fazer inferência para a população como um todo ao invés de valores específicos observados no experimento.

Modelos fixos são modelos envolvendo apenas efeitos fixos, modelos aleatórios são modelos envolvendo apenas efeitos aleatórios e modelos que envolvem efeitos aleatórios e efeitos fixos são chamados de modelos mistos.

2.1.8.2.1 Modelo linear fixo envolvendo apenas efeito de tratamento.

Nesse modelo os parâmetros são comparações entre as amostras de RNA, ou seja, o único efeito envolvido nesse modelo é o efeito de tratamento (amostras de RNA). Assim o modelo será

definido como

$$y = X\alpha + \varepsilon \quad (5)$$

em que Y é um vetor de logaritmos na base 2 de valores observados para o gene g (ou logaritmo das razões entre as amostras de RNA), X é uma matriz delineamento, α é um vetor de parâmetros desconhecidos e ε é o erro.

Note que o teste t para dois grupos pode ser pensado como um caso especial de modelo linear definindo X como sendo uma matriz $n \times 2$ com a primeira coluna com todos os elementos iguais a 1 para representar o intercepto e na segunda coluna os elementos são iguais a 1 se a observação é do grupo 2 e iguais a 0 se a observação é do grupo 1. Correspondente à matriz delineamento, o vetor α contém dois elementos: α_0 representando o intercepto e α_1 representando a diferença entre a média do valor da resposta para o grupo 2 e a média do valor da resposta para o grupo 1. Na forma de escalares, essa comparação entre os dois grupos pode ser escrita como

$$y_{ij} = \alpha_0 + \alpha_1 x_{ij} + \varepsilon_{ij}, \quad j = 1, 2$$

em que

$$x_{ij} = \begin{cases} 0, & \text{se } j = 1 \\ 1, & \text{se } j = 2 \end{cases}$$

e $\varepsilon_{ij} \sim N(0, \sigma^2)$. Se $\mu_1 = \alpha_0$ representa a média dos níveis de expressão no grupo 1 e $\mu_2 = \alpha_0 + \alpha_1$ representa a média dos níveis de expressão nos grupos 2, segue que o teste de expressão gênica diferencial, que é equivalente ao teste t para dois grupos assumindo variâncias iguais entre os dois grupos é provido testando $H_0: \alpha_1 = 0$ nesse modelo linear. Esta parametrização do teste t para dois grupos corresponde ao esquema de referência comum, onde o primeiro grupo é usado como referência.

Segundo Smyth (2004) o delineamento de qualquer experimento de *microarray* pode ser representado em termos de um modelo linear para cada gene, como definido pela eq. (5). Essa

abordagem requer que duas matrizes sejam definidas: a matriz delineamento (*design matrix*) e a matriz contraste (*contrast matrix*).

2.1.8.2.1.1 Matriz Delineamento

Representa as diferentes amostras de RNA que serão hibridizadas no *array*. É uma matriz de coeficientes onde cada linha representa um *array* e cada coluna corresponde a um parâmetro. Os parâmetros devem ser independentes, ou seja, nenhum deles poderá ser obtido como combinação dos demais.

Exemplo:

No seguinte exemplo (THORNE, 2006) tem-se um delineamento com três *arrays* (Figura 28) onde duas amostras (*A* e *B*) são comparadas indiretamente através de uma referência comum (Ref). Cada *array* é representado por um parâmetro (comparação entre duas amostras), o qual é obtido pela diferença entre a amostra na ponta da flecha e a amostra na base da flecha. O *array* 1, por exemplo, é representado por (*A* – Ref).

Figura 28 - Delineamento com três *arrays*, duas amostras e uma referência comum

Quando se utiliza um delineamento com inversão de corantes (*dye swap*) deve-se indicar na matriz delineamento o *array* no qual os corantes foram invertidos invertendo o sinal do respectivo parâmetro. Assim o *array* 1 será representado por (*A* – Ref), o *array* 2 será representado por (Ref – *A*) ou – (*A* – Ref) e o *array* 3 será representado por (*B* – Ref)

$$\begin{pmatrix} \text{array 1} \\ \text{array 2} \\ \text{array 3} \end{pmatrix} = \begin{pmatrix} A - \text{Ref} \\ -(A - \text{Ref}) \\ B - \text{Ref} \end{pmatrix}$$

Se forem considerados parâmetros $(A - \text{Ref})$ e $(B - \text{Ref})$, por exemplo, tem-se:

$$\alpha_j = \begin{pmatrix} A - \text{Ref} \\ B - \text{Ref} \end{pmatrix}$$

Assim a matriz delineamento deverá ser uma matriz X tal que

$$\begin{pmatrix} \text{array 1} \\ \text{array 2} \\ \text{array 3} \end{pmatrix} = X \begin{pmatrix} A - \text{Ref} \\ B - \text{Ref} \end{pmatrix}$$

ou

$$\begin{pmatrix} A - \text{Ref} \\ -(A - \text{Ref}) \\ B - \text{Ref} \end{pmatrix} = X \begin{pmatrix} A - \text{Ref} \\ B - \text{Ref} \end{pmatrix}$$

Logo a matriz delineamento será

$$\begin{array}{cc} \begin{array}{c} A - \text{Ref} \\ B - \text{Ref} \end{array} & \begin{array}{c} A - \text{Ref} \\ B - \text{Ref} \end{array} \\ \text{array 1} & \begin{pmatrix} 1 & 0 \end{pmatrix} \\ \text{array 2} & \begin{pmatrix} -1 & 0 \end{pmatrix} \\ \text{array 3} & \begin{pmatrix} 0 & 1 \end{pmatrix} \end{array}$$

O sinal (-1) na segunda linha indica que houve inversão de corante (*dye swap*) no *array 2*. Então

$$\begin{pmatrix} A - \text{Ref} \\ -(A - \text{Ref}) \\ B - \text{Ref} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} A - \text{Ref} \\ B - \text{Ref} \end{pmatrix}$$

Assim o modelo linear para o gene g em forma matricial usando a matriz delineamento pode ser escrito como:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} A - \text{Ref} \\ B - \text{Ref} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

Observe que o modelo acima faz comparações apenas entre a amostra A e a amostra referência e entre a amostra B e a amostra referência, porém a comparação de interesse que seria entre as amostras A e B não é feita. Para comparar as amostras A e B pode-se utilizar a matriz contraste, ou, como nesse caso existe apenas uma comparação de interesse, pode-se simplesmente especificar esse parâmetro no vetor α_j . Dessa forma o vetor α_j contendo o parâmetro $B - A$ poderá ser escrito como:

$$\alpha_j = \begin{pmatrix} A - \text{Ref} \\ B - A \end{pmatrix}$$

Resulta que

$$\begin{pmatrix} A - \text{Ref} \\ -(A - \text{Ref}) \\ B - \text{Ref} \end{pmatrix} = X \begin{pmatrix} A - \text{Ref} \\ B - A \end{pmatrix}$$

Assim a matriz delineamento será

$$X = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \end{pmatrix}$$

Logo

$$\begin{pmatrix} A - \text{Ref} \\ -(A - \text{Ref}) \\ B - \text{Ref} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} A - \text{Ref} \\ B - A \end{pmatrix}$$

Assim o modelo linear para o gene g em forma matricial usando a matriz delineamento pode ser escrito como:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} A - \text{Ref} \\ B - A \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

Observe que esse modelo contém o parâmetro de interesse ($B - A$).

Depois de ajustar um modelo linear para cada gene os parâmetros são estimados e submetidos a testes de significância.

2.1.8.2.1.2 Matriz Contraste

Como na matriz delineamento todos os parâmetros devem ser independentes pode ocorrer de algum contraste de interesse não estar incluído. Deve-se então fazer uma nova matriz para incluir esses contrastes na análise, a matriz contraste.

Essa matriz permite que os parâmetros definidos pela matriz delineamento sejam combinados em contrastes, onde cada contraste corresponde uma comparação de interesse entre as amostras de RNA. Para experimentos simples, como no exemplo anterior, a matriz contraste pode ser omitida.

Os contrastes de interesse são dados por:

$$\beta_g = C^T \alpha_g$$

onde C é a matriz contraste.

Exemplo:

Considere o delineamento em *loop* (Figura 29) com três *arrays* e três amostras

Figura 29 - Delineamento em *loop* com três *microarrays* e três amostras

Pode haver interesse nos contrastes em todas as comparações entre as amostras, ou seja, pode-se ter interesse nos contrastes $(B - A)$, $(C - B)$ e $(C - A)$, porém a matriz delineamento define apenas dois destes, pois eles devem ser independentes. Entretanto pode-se utilizar a matriz contraste para fazer combinações dos dois parâmetros definidos pela matriz delineamento obtendo-se todos os contrastes de interesse.

Considere o seguinte modelo em que a matriz delineamento define os contrastes $B - A$ e $C - B$:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} B - A \\ C - B \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

Considerando a matriz

$$C = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

os contrastes de interesse são obtidos por

$$\beta_g = C^T \alpha_g = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} B - A \\ C - B \end{pmatrix} = \begin{pmatrix} C - B \\ B - A \\ C - A \end{pmatrix}$$

2.1.8.2.2 Modelo de análise de variância (ANOVA)

Ao contrário do teste t para dois grupos, o modelo linear de efeitos fixos permite comparar mais de dois grupos ou o controle de outros efeitos fixos incluindo termos adicionais no modelo (eq. (6)) como efeitos de grupos (variedades), genes, corantes, e *array*. Quando esses termos adicionais estão associados com a resposta, a variância do erro (residual) pode ser substancialmente diminuída. Desde que o denominador da estatística de teste usada para comparar os dois grupos envolve essa variância do erro, isso permite aumentar o poder na comparação dos dois grupos. Conseqüentemente, em experimentos de *microarray*, levando em conta causas de variação conhecidas aumenta-se o poder do experimento para observar diferenças significativas nos níveis de expressão para um dado gene (COFFEY; COFIELD, 2006).

Dependendo quais causas de variação são consideradas de maior importância, diferentes modelos podem ser de maior utilidade em um estudo particular. Por exemplo, a escolha de um modelo poderia incluir parâmetros para considerar os quatro fatores de efeitos principais e interações que considerem (VG) (efeito de interesse), variação de *spot* para *spot*, e interações entre corantes e genes:

$$y_{ijk} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \varepsilon_{ijk} \quad (6)$$

em que y_{ijk} é o logaritmo do valor de intensidade de expressão observado no *array* i (A_i), no corante j (D_j), na variedade k (V_k), no gene g (G_g), μ é a média geral, $(AG)_{ig}$ é a interação (*array* × gene), $(DG)_{jg}$ é a interação (corante × gene) e ε_{ijk} são os resíduos, os quais são assumidos ser variáveis aleatórias independentes, identicamente distribuídas com média 0 e variância σ_g^2 . É assumida homogeneidade de variâncias para observações de um mesmo gene, mas é permitida heterogeneidade de variâncias para níveis de expressão em diferentes genes. Para o teste de hipóteses é necessária a suposição que os erros seguem uma distribuição normal.

Kerr et al. (2002) também propõem um modelo alternado (eq. (7)), que inclui o termo da interação *array* × corante $(AD)_{ij}$:

$$y_{ijk} = \mu + A_i + D_j + (AD)_{ij} + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \varepsilon_{ijk} \quad (7)$$

Segundo Kerr e Churchill (2001) quando os genes são replicados no *array* o efeito $(AG)_{ig}$ deve ser substituído pelo efeito dos *spots* $(AG)_{igs}$.

Como os erros aleatórios diferem de gene para gene, Wolfinger et al. (2001) sugeriram separar o modelo acima em dois modelos, onde o primeiro modelo (eq. (8)) contém os componentes “globais” e pode ser visto como um modelo de normalização

$$y_{ijk} = \mu + A_i + D_j + (AD)_{ij} + r_{ijk} \quad (8)$$

Os resíduos normalizados do primeiro modelo são então utilizados como entrada no modelo gênico (eq. (9)) (modelo de efeitos específicos para cada gene de interesse)

$$r_{ijk} = G_g + (VG)_{kg} + (DG)_{jg} + (AG)_{ij} + \varepsilon_{ijk} \quad (9)$$

Novamente o termo (VG) representa o efeito de interesse.

Uma grande vantagem do modelo de ANOVA é que pode ser estendido para delineamentos de *microarray* mais complexos.

Casella (2008) propõe uma abordagem que considera um experimento de *microarray* como um experimento em parcelas subdivididas (eq. (10)). As amostras de RNA são consideradas os tratamentos aplicados às parcelas e os genes são os tratamentos aplicados às subparcelas. Assumindo os *arrays* como blocos o modelo será

$$y_{ijg} = \mu + A_i + T_j + \varepsilon_{ij} + G_g + (AG)_{ig} + (TG)_{jg} + \delta_{ijk} \quad (10)$$

em que y_{ijg} é a variável resposta, μ é o efeito da média geral, A_i é o efeito do *array* i, T_j é o efeito do tratamento j aplicado à parcela, ε_{ij} é o efeito do erro devido à parcela em que $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, G_g é o efeito do gene g, $(AG)_{ig}$ é a interação entre o *array* i e o gene g, $(TG)_{jg}$ é a interação entre o tratamento j e o gene g e δ_{ijk} é o efeito do erro devido à subparcela em que $\delta_{ijk} \sim N(0, \sigma_\delta^2)$.

2.1.8.2.3 Modelo de efeitos mistos

Modelos de efeitos fixos permitem a inclusão de efeitos de fatores, entretanto esses modelos são baseados nas suposições de independência entre as observações e variância homogênea entre os níveis de expressão de um mesmo gene. A presença de efeitos aleatórios em experimentos de *microarray* pode introduzir correlação entre os níveis de expressão. Por exemplo, em um típico experimento de *microarray*, o efeito do *array* pode ser considerado um efeito aleatório, como os *arrays* usados no experimento representam um subgrupo de todos os possíveis *arrays*. Modelos mistos incluem ambos os efeitos fixos e aleatórios como pode-se observar na eq. (11). Wolfinger et al. (2001) utiliza um modelo em dois estágios como descrito acima o qual considera aleatórios os efeitos de *array* e *spots* seguindo uma distribuição normal.

Em notação matricial um modelo de efeito misto é definido como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad (11)$$

em que:

- a) $\mathbf{y}$ é um vetor de observados logaritmos na base 2 das variáveis resposta;
- b) $\mathbf{X}$ é uma matriz delineamento de constantes conhecidas para os efeitos fixos;
- c) $\boldsymbol{\beta}$ é um vetor de parâmetros de efeitos fixos desconhecidos;
- d) $\mathbf{Z}$ é uma matriz delineamento de constantes conhecidas para os efeitos aleatórios;
- e) $\mathbf{u}$ é um vetor de parâmetros de efeitos aleatórios desconhecidos;
- f) $\boldsymbol{\varepsilon}$ é um vetor de erros aleatórios.

As suposições sobre $\mathbf{u}$ e $\boldsymbol{\varepsilon}$ são: $E[\mathbf{u}] = 0$, $E[\boldsymbol{\varepsilon}] = 0$, $\text{var}[\mathbf{u}] = \mathbf{G}$ e $\text{var}[\boldsymbol{\varepsilon}] = \mathbf{R}$, onde $\mathbf{u}$ e $\boldsymbol{\varepsilon}$ são não correlacionados.

2.1.8.3 Abordagem Bayesiana Empírica

Em geral em estudos com *microarrays* o número de lâminas n_h é pequeno, e, portanto as estimativas da variância por gene vão ter muito poucos graus de liberdade e serão muito instáveis (ROSA; SOLER, 2004; apud HUBER et al., 2003).

Para contornar o problema do número de lâminas pequeno podem ser utilizados métodos Bayesianos empíricos, onde uma distribuição a priori é assumida nos dados e os parâmetros

desconhecidos desta distribuição são estimados a partir do próprio conjunto de dados.

Segundo Esteves (2007) métodos bayesianos empíricos aplicados à análise de dados de *microarray* foram inicialmente propostos por Lönnstedt & Speed (2002), onde a estatística proposta era interessante apenas para experimentos que comparam apenas dois grupos biológicos sem a utilização da amostra de referência. Assim, Smyth (2004) propôs uma extensão da estatística proposta por Lönnstedt & Speed (2002) para a aplicação de testes de hipóteses baseados em modelos de ANOVA, tanto para testes F como testes t envolvendo contrastes específicos.

A estatística proposta por Smyth (2004) é a estatística t moderada. A interpretação desta é a mesma da estatística t ordinária exceto que os erros padrões foram moderados ao longo dos genes utilizando um modelo Bayesiano. A estatística t moderada conduz a p-valores do mesmo modo que a estatística t ordinária, com exceção dos graus de liberdade que são aumentados.

O modelo linear para o gene g tem variância residual σ_g^2 com valor amostral s_g^2 e graus de liberdade f_g . A matriz de covariância da estimativa de $\hat{\beta}_g$ é

$$\sigma_g^2 C^T (X^T V_g X)^{-1} C$$

onde V_g é uma matriz de pesos determinada a priori.

O método bayesiano empírico assume uma distribuição qui-quadrado inversa à priori para σ_g^2 com média s_0^2 e graus de liberdade f_0 . Os valores a posteriori para a variância residual são dados por

$$\tilde{s}_g^2 = \frac{f_0 s_0^2 + f_g s_g^2}{f_0 + f_g}$$

onde f_g são os graus de liberdade residuais para o g -ésimo gene. A estatística t moderada é

$$\tilde{t}_{gk} = \frac{\hat{\beta}_{gk}}{\tilde{s}_g \sqrt{u_{gk}}}$$

onde u_{gk} é o k-ésimo elemento da diagonal de

$$C^T V_g C .$$

Pode ser mostrado que esta estatística segue uma distribuição t com graus de liberdade $f_0 + f_g$ se $\beta_{gk} = 0$. Os graus de liberdade extra f_0 representam a informação extra que é extraída do conjunto de genes para fazer inferência sobre cada gene (SMYTH, 2005).

2.1.8.4 Tipos de erro e testes Múltiplos

O erro do Tipo I é o erro que se comete ao rejeitar a hipótese nula H_0 quando na verdade ela é verdadeira, o erro do Tipo II é o erro que se comete ao aceitar a hipótese nula H_0 quando ela é falsa. Os erros do Tipo I e II são denotados por α e β , respectivamente. O poder de um teste é definido como a probabilidade de se rejeitar a hipótese nula H_0 quando ela é falsa $(1 - \beta)$.

Segundo Banzatto e Kronka (1995), geralmente tenta-se controlar apenas o erro do Tipo I, através do nível de significância do teste (α), que consiste na probabilidade máxima de se cometer um erro do Tipo I. Na prática, é comum o nível de significância ser fixado em 5% ou em 1%, ou seja, $\alpha = 0,05$ ou $\alpha = 0,01$.

Em experimentos de *microarray* realizamos testes de hipótese para cada gene, e, segundo Pereira (2008), o grande número de hipóteses a ser testada simultaneamente gera o problema dos chamados testes de comparações múltiplas.

Para introduzir a idéia de testes múltiplos considere o seguinte exemplo (ROSA; SOLER, 2004): para cada gene g foi conduzido um teste de hipótese para verificar expressão diferencial em que verificou-se que os genes g_1 e g_2 são diferencialmente expressos, com um nível de significância individual $\alpha = 0,05$. Isso significa que a probabilidade de um falso positivo é de 5%, isto é, de g_1 não ser diferentemente expresso e ter sido identificado como tal. O mesmo vale para g_2 . Sejam $P_1 = 1 - 0,05$ e $P_2 = 1 - 0,05$ as probabilidades dos genes g_1 e g_2 , respectivamente, serem identificados corretamente, assumindo independência entre as

inferências, tem-se que essa probabilidade é dada por $P_1 \times P_2 = (1 - 0,05)^2 = 0,9025$. É imediato que a probabilidade de cometer algum erro será de $1 - 0,9025 = 0,0925$, ou seja, apesar de ser tolerado um erro de 5% individualmente, quando se considera a probabilidade de todas as afirmações estarem corretas no conjunto de testes realizados, o erro cometido é quase duas vezes maior.

Quanto maior o número de hipóteses rejeitadas, ou de genes identificados segundo esse critério, maior será a probabilidade de que hajam falsos positivos. Deve-se então utilizar procedimentos que controlem a taxa de erro global (FWER, do inglês familywise error rate). Nesse contexto a taxa de erro global é a probabilidade de que haja pelo menos um falso positivo dentre os genes identificados como diferentemente expressos em uma família de testes realizados (ROSA; SOLER, 2004). Para controlar a taxa de erro global pode-se utilizar, por exemplo, a correção de Bonferroni ou o ajuste por FDR.

Segundo Wit e McClure (2004), a correção de Bonferroni classifica todos os genes que tem um p-valor associado menor que α/n como diferencialmente expressos, onde α é o nível de significância global e n é o número de hipóteses testadas. Este procedimento garante que a FWER seja menor ou igual a α . Como em experimentos de *microarray* são realizados muitos testes simultâneos o critério de Bonferroni pode ser muito restritivo, pois o nível de significância individual de cada teste ficará excessivamente baixo e os testes perderão poder.

Segundo Rosa e Soler (2004), uma alternativa a esse controle estrito da taxa de erro global foi proposta por Benjamini & Hochberg (1995) através do conceito da taxa de falsas descobertas ou FDR. A taxa de falsas descobertas é definida como

$$FDR = E \left[\frac{V}{R} \mid R > 0 \right] P(R > 0)$$

onde R é o número de genes identificados como diferencialmente expressos e V é o número desses genes que foram erroneamente identificados (falsos positivos).

Pereira (2008) descreve o procedimento para determinar o ponto de corte em testes múltiplos controlando a FDR proposto em Benjamini e Hochberg (1995): para cada uma das hipóteses testadas $H_{01}, H_{02}, \dots, H_{0m}$, obter o valor da estatística de teste e o correspondente p-valor

$P_1, P_2, \dots, P_m$. Colocar os $P_i, i = 1, \dots, m$ em ordem crescente. Sejam $P_{(1)}, P_{(2)}, \dots, P_{(m)}$ os p-valores ordenados, definindo $q = \frac{mP_{(i)}}{i}$, a FDR pode ser controlada em um nível q^* , determinando-se o maior i para o qual

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

Assim para controlar a FDR a um nível $q^* = 5\%$ o ponto de corte será o $P_{(i)}$ com maior i que satisfaça a condição: $5\% \geq \frac{mP_{(i)}}{i}$, ou seja, rejeita-se as hipóteses com p-valor menor ou igual a $P_{(i)}$.

2.2 O software R

R (R CORE DEVELOPMENT TEAM, 2009) é uma linguagem computacional matemática e um ambiente para análise de dados e gráficos. É similar à linguagem S, porém é gratuito. Além disso, existem pacotes (*library*) que são usados para certas análises estatísticas escritas para R (ou S). Devido à sua diversidade matemática e por ser gratuito, o R tem se tornado uma das ferramentas mais populares para análise de *microarrays* (WIT; MCCLURE, 2004). Uma cópia do *software* R pode ser obtida no *site*: < <http://cran.r-project.org/> >.

2.2.1 O projeto *Bioconductor* e o pacote *limma*

Bioconductor (GENTLEMAN, et al., 2004) é um projeto de *software* de código e desenvolvimento abertos pra prover ferramentas para análise e compreensão de dados genômicos. Nele são desenvolvidos pacotes que funcionam no ambiente computacional estatístico R.

Para instalação do *Bioconductor* basta digitar diretamente no R os comandos abaixo (é necessário conexão com a internet):

```
> source("http://www.bioconductor.org/getBioC.R")
> getBioC()
```


Serão instalados um conjunto de pacotes para análise de dados genômicos, entre os quais está o pacote *limma* (SMYTH, 2004).

O *limma* é um pacote para análise de expressão gênica em dados de *microarray*, o qual faz parte do projeto *Bioconductor*. Ele utiliza modelos lineares para fazer as análises, permitindo fazer comparações de muitas amostras de RNA simultaneamente e tem características que tornam a análise estável até mesmo para experimentos com pequeno número de *arrays*. O pacote *limma* também tem funções para normalização dos dados e análise exploratória (SMYTH et al., 2008).

3 MATERIAIS E MÉTODOS

Segundo Guidetti-Gonzalez (2009), o desenvolvimento de variedades resistentes a doenças através do melhoramento genético convencional dos citros apresenta uma série de dificuldades como, por exemplo, esterilidade de pólen e óvulo, incompatibilidade sexual, etc. Estas dificuldades podem ser superadas através da biotecnologia, para tanto é necessário identificar os genes envolvidos nas interações planta-patógeno, ou seja, identificar genes que apresentem níveis de expressão diferencial em diferentes condições biológicas (planta sadia versus planta doente). Isto pode ser feito utilizando-se a técnica de *microarrays* de DNA, que permite analisar os níveis de expressão de milhares de genes simultaneamente. Nessa seção serão analisados dados de expressão gênica de amostras de citros em duas condições obtidos por meio de um experimento de *macroarray* utilizando modelos lineares de efeitos fixos considerando a inclusão ou exclusão de diferentes efeitos; e considerando ajustes de modelos para cada gene separadamente e para todos os genes simultaneamente.

3.1 Material experimental

Os dados utilizados nesse trabalho foram gentilmente cedidos pela Dra. Simone Guidetti-Gonzalez e pela Prof^a. Dra. Helaine Carrer da ESALQ – USP. Os mesmos referem-se a um estudo de expressão gênica considerando-se genes de citros. Para obtenção dos dados foi utilizada a técnica de *macroarrays*, técnica similar a técnica de *microarrays*. O experimento foi instalado em dois *macroarrays* (*arrays* de náilon), cada um contendo 864 *spots* distribuídos em 24 linhas e 36 colunas, denotados por A₁ e A₂, os quais foram confeccionados manualmente (Figura 30).

Figura 30 - Estrutura dos arrays

Em cada *array* foram fixados cDNAs contendo as sequências de 288 genes (Figura 31), sendo que 273 são genes de citros, os quais foram replicados três vezes cada, em *spots* consecutivos, e mais 2 genes controle (gliceraldeído-3-fosfato desidrogenase-GAPDH e β -actina), os quais foram replicados 24 e 21 vezes, respectivamente, em sequências de três *spots* consecutivos distribuídas pelo *array*. Não há interesse em verificar expressão diferencial nos genes controle, tendo em vista que eles servem apenas para propósitos de normalização, assim eles serão excluídos da análise sendo então considerados apenas os 273 genes de citros.

Figura 31 - Distribuição dos genes no *array*

Após a fixação dos genes, os dois *arrays* foram hibridizados com a sonda *Overgo* (“amostra” marcada feita para hibridizar com todos os cDNAs fixados nos *arrays*) e, posteriormente, foram feitas leituras para obtenção das imagens das hibridizações as quais foram digitalizadas para posterior quantificação dos sinais dos *spots*. O objetivo da hibridização com a sonda *Overgo* é quantificar o material fixado em cada *spot*, permitindo assim eliminar a variação devido à diferença na quantidade de cDNA fixado em cada um deles (GUIDETTI-GONZALEZ, 2009). A seguir foi realizado um processo de lavagem para remover a sonda *Overgo* sem remover os cDNAs fixados nos *arrays*.

Segundo Guidetti-Gonzalez (2009), foi obtido o RNA de folha de *Citrus sinensis* sadia e com sintomas de pinta preta, e a partir desses RNAs foram obtidas duas amostras de cDNA marcadas com “corante”: CsS (folha de *Citrus sinensis* sadia) e CsP (folha de *Citrus sinensis* com sintomas de pinta preta). Os dois *arrays* foram hibridizados com as duas amostras marcadas. Na primeira etapa o *array* A₁ foi hibridizado com a amostra CsS e o *array* A₂ com a amostra CsP,

sendo posteriormente realizadas leituras dos *arrays* para obtenção das imagens as quais foram digitalizadas para posterior quantificação dos sinais dos *spots* (Figura 28 A e B). Foi realizado um processo de lavagem para remoção do material fixado. Em uma segunda etapa, a ordem da hibridização foi invertida: o *array* A₁ foi hibridizado com a sonda CsP e o *array* A₂ com a sonda CsS. Foram feitas leituras para obtenção das imagens das hibridizações (Figura 29 C e D) as quais foram digitalizadas para posterior quantificação dos sinais dos *spots*.

Observe que esse procedimento de inversão das amostras hibridizadas antes e depois da lavagem minimizou uma possível variação devido ao processo de lavagem. Por exemplo, se na primeira etapa de hibridização tivesse sido utilizada a amostra CsS nos dois *arrays* e na segunda etapa de hibridização tivesse sido utilizada a amostra CsP nos dois *arrays* então o efeito da lavagem ficaria completamente confundido com o efeito de tratamento.

Figura 32 - Imagens das hibridizações dos *arrays* com as amostras: (A) A₁ com a amostra CsP, (B) A₁ com a amostra CsS, (C) A₂ com a amostra CsP e (D) A₂ com a amostra CsS

Depois de quantificados os sinais do *foreground* e do *background* das duas amostras, foi feita a correção de *background* e o resultado foi transcrito em uma tabela de dados contendo as intensidades de sinal dos *spots* para as amostras CsS (folha de *Citrus sinensis* sadia), CsP (folha

de *Citrus sinensis* com sintomas de pinta preta) e também os valores de intensidade da sonda *Overg*, em cada *array* (Tabela 2).

Tabela 2 - Visão parcial do conjunto de dados de expressão gênica de citros com *background* corrigido

Gene	Array A ₁			Array A ₂		
	<i>Overg</i>	CsS	CsP	<i>Overg</i>	CsS	CsP
1	1,428	1,944	6,089	2,111	0,601	5,100
1	2,450	4,068	13,702	2,717	0,844	9,640
1	1,603	2,056	8,527	1,885	0,567	5,093
2	1,492	2,514	7,297	2,294	1,484	10,262
2	2,241	3,770	13,389	2,104	1,419	9,769
2	1,926	3,128	10,495	2,431	1,805	12,872

3.1.1 Delineamento do experimento

Na Figura 33 pode ser visualizada a representação do delineamento do experimento utilizando flechas.

Figura 33 - Delineamento do experimento de *macroarray*

Observe que cada flecha representa um *array* e CsS e CsP são as amostras controle e tratamento, respectivamente. Note que nesse caso as amostras não foram marcadas com “corantes” diferentes, então a base e a ponta da flecha não representam diferentes corantes. Entretanto, as duas amostras não foram hibridizadas no *array* ao mesmo tempo e houve um processo de lavagem entre uma hibridização e a outra, podendo-se assim questionar se esse processo de lavagem influenciou a hibridização seguinte, dessa forma ao invés de efeito de corante pode-se considerar o efeito da “lavagem”, então a base da flecha representa a hibridização antes da lavagem e a ponta da flecha representa a hibridização após a lavagem.

Em experimentos de *microarray* de duas cores as colunas do banco de dados são representadas pelo nome do corante com o qual a amostra foi marcada (Cy3 ou Cy5) e não pelo nome da amostra. Ao contrário de experimentos com dois corantes, pode-se observar na Tabela 2 que foi utilizado o nome da própria amostra para representar a coluna contendo as intensidades de

expressão da mesma, sendo assim, não haverá necessidade de indicar a inversão das hibridizações na matriz delineamento invertendo o sinal do coeficiente.

3.2 Métodos

3.2.1 Normalização dos dados

Antes de serem analisados, os dados passaram por algumas etapas de normalização para reduzir os efeitos da variação causada pela técnica.

3.2.1.1 Normalização com a sonda *Overg*

Primeiramente os valores de intensidade das duas amostras, CsS (folha de *Citrus sinensis* sadia) e CsP (folha de *Citrus sinensis* com sintomas de pinta preta), em cada *array*, foram divididos pelos valores de intensidade da sonda *Overg* no respectivo *array*. Assim, o valor de intensidade das duas amostras em cada *spot* foi dividido pelo valor de intensidade da sonda *Overg* do *spot* correspondente, diminuindo a variação devida à diferença de material depositado em cada *spot*.

3.2.1.2 Normalização dependente da intensidade de expressão

Após a normalização com a sonda *Overg*, foram removidos os genes controle (β -actina e GAPDH), pois estes não serão analisados. Foi então realizada outra normalização para eliminar o viés devido à intensidade de expressão dentro do *array* utilizando regressão não paramétrica (*Loess*). Essa normalização foi realizada sobre a transformação *MA*, sendo $M = \log_2(CsP) - \log(CsS)$ e $A = 0,5 \times [\log_2(CsP) - \log(CsS)]$ (em que CsP é a amostra de folha de *Citrus sinensis* com sintomas de pinta preta e CsS é a amostra de folha de *Citrus sinensis* sadia).

Note que não aparecem “Cy3” ou “Cy5” em *M* ou em *A*, pois nenhum destes corantes foi utilizado na marcação da amostra. Na realidade foi utilizado o mesmo “corante” para marcar as duas amostras, as quais foram hibridizadas nos *arrays*, uma de cada vez, após processos de lavagem para remover a hibridização anterior.

3.2.1.3 Normalização entre *arrays* para correção de escala

Após a normalização para eliminar o viés devido à intensidade de expressão dentro do *array* foi realizada uma normalização entre os *arrays* para correção de escala.

Como as normalizações foram realizadas sobre a transformação *MA* e, em algumas análises serão necessários os logaritmos dos valores de intensidade, estes valores podem ser obtidos como na eq. (1), fazendo

$$\log_2(CsP^*) = A^* + 0,5M^* \text{ e } \log_2(CsS^*) = A^* - 0,5M^*$$

em que A^* e M^* são os valores normalizados de A e M .

3.2.2 Métodos de Análise

Foram utilizados modelos lineares de efeitos fixos para análise dos dados sendo que cada modelo levou em consideração diferentes efeitos e também diferentes abordagens em relação ao número de genes analisados (um gene por vez ou todos os genes conjuntamente). Foram realizados testes de significância sobre os efeitos de interesse em cada um dos modelos utilizando a estatística t e os resultados obtidos a partir de cada modelo foram comparados. Também foi utilizada em um dos modelos uma abordagem bayesiana empírica para comparar os resultados obtidos utilizando-se a estatística t ordinária com os resultados obtidos a partir da estatística t moderada. Devido ao fato de que nem todos os genes seguem uma distribuição normal foi também feito um teste de Wilcoxon.

Os métodos de análise utilizados foram:

1. Análise com um modelo linear fixo considerando apenas o efeito de tratamento;
2. Análise com um modelo linear fixo considerando apenas o efeito de tratamento e utilizando um método bayesiano empírico associado com uma estimativa de correlação entre as repetições dos genes;
3. Análise com um modelo de Análise de Variância para cada gene considerando efeitos de tratamento e *array* (*spots*);
4. Análise em dois estágios em que o primeiro modelo (modelo de normalização) considera os efeitos globais [*Array*, Tratamento, (*Array* $\times$ Tratamento)] e o segundo modelo

(modelo gênico) considera os efeitos ao nível de cada gene [Gene, (Gene $\times$ Tratamento) e (Gene $\times$ Array)];

5. Análise considerando um delineamento em parcelas subdivididas em que o *array* foi considerado como bloco, as amostras de cDNA (CsS e CsP) foram consideradas como os tratamentos aplicados às parcelas e os genes foram considerados como os tratamentos aplicados às subparcelas;
6. Análise dos dois grupos utilizando o teste da soma de postos de Wilcoxon (equivalente ao teste de Mann-Whitney).

3.2.2.1 Método 1

Foi ajustado um modelo linear fixo para cada gene para comparar os dois grupos biológicos representados pelas amostras *CsS* e *CsP* em que os valores da variável resposta foram $y_{gij} = \log_2 \left(\text{CsP}_{gij} / \text{CsS}_{gij} \right)$. O modelo ajustado foi:

$$y = X\alpha + \varepsilon$$

em que y é um vetor 6×1 de valores M para o gene g , $X = \mathbf{1}_{6 \times 1}$ é uma matriz delineamento 6×1 , $\alpha_g = (CsP - CsS)$ é um vetor 1×1 de parâmetros desconhecidos e ε é um vetor 6×1 de erros aleatórios. Note que cada gene tem três repetições dentro de cada um dos dois *arrays*, e, como a variável resposta é $y_{gij} = \log_2 \left(\text{CsP}_{gij} / \text{CsS}_{gij} \right)$, ou seja, o logaritmo da razão entre os valores de intensidade das duas amostras resulta em seis observações para cada gene ($2 \text{ arrays} \times 3 \text{ repetições}$).

Note que como os genes foram replicados três vezes dentro de cada *array* e foram utilizados dois *arrays* resultando em seis valores M para cada gene, cada linha da matriz delineamento representa uma repetição do gene g no *array* i .

Se μ_g é a verdadeira média de $y_{gij} = \log_2 \left(\text{CsP}_{gij} / \text{CsS}_{gij} \right)$ para o gene g , então a hipótese a ser testada será $H_0 : \mu_g = 0$. Note que

$$\begin{aligned}
\mu_g = 0 &\Rightarrow \log_2(CsP_g / CsS_g) = 0 \\
&\Rightarrow \log_2(CsP_g) - \log_2(CsS_g) = 0 \\
&\Rightarrow \log_2(CsP_g) = \log_2(CsS_g).
\end{aligned}$$

O teste utilizado para testar a hipótese $H_0 : \mu_g = 0$ foi o teste t com 5 graus de liberdade. Também foi utilizado um ajuste por FDR para controlar a taxa de falsos positivos.

3.2.2.2 Método 2

Para realizar essa análise utilizou-se o pacote *limma* (SMYTH, 2004; SMYTH, et al. 2005b) do projeto *Bioconductor*.

Smyth et al. (2005) propuseram um método para extrair informação das repetições dos genes dentro do *array* estimando a correlação entre elas. A correlação entre as repetições dos genes é estimada para cada gene separadamente utilizando máxima verossimilhança restrita (REML, do inglês *restricted maximum likelihood*), baseado em um modelo linear de efeitos mistos. Uma estimativa global da correlação entre as réplicas é calculada como uma média das correlações individuais em escala arco tangente hiperbólico (*atanh*), e esta estimativa global é então utilizada quando um modelo linear é ajustado para cada gene.

Foi ajustado um modelo linear fixo para cada gene considerando apenas o efeito de tratamento, em que os valores da variável resposta são $y_{gij} = \log_2(CsP_{gij} / CsS_{gij})$. Foi considerado o modelo:

$$y = X\alpha + \varepsilon$$

em que y é uma matriz 3×2 de valores M para o gene g , $X = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ é uma matriz delineamento

2×1 , $\alpha_g = (CsP - CsS)$ é um vetor 1×1 de parâmetros desconhecidos e ε é o erro aleatório.

Utilizou-se o comando `lmFit` do *limma* para ajustar um modelo linear para cada gene e, como haviam repetições do gene dentro do *array* foi selecionado o método `gls.series` que estima os parâmetros e o desvio padrão para cada gene utilizando o método de mínimos quadrados generalizados, que leva em conta a correlação entre as repetições dos genes dentro do

array. Considerou-se y como uma matriz 3×2 e $X = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ apenas por questões computacionais, pois, no processo de análise o *software* reajusta y e X (esse processo não é visível) de forma que estes tenham ordem 6×1 .

Como o número de *arrays* utilizados foi pequeno utilizou-se o método bayesiano empírico de Smyth (2004), pois a estimativa de variância da estatística t moderada utilizada neste método é mais estável que a estatística t ordinária.

A estimativa global de correlação foi combinada com o método Bayesiano empírico para calcular as estimativas de variância. Foi também utilizado um ajuste por FDR para controlar a taxa de falsos positivos.

3.2.2.3 Método 3

Foi ajustado um modelo de análise de variância para cada gene:

$$y_{ijk} = G_g + (TG)_{jg} + (AG)_{igk} + \varepsilon_{ijk}$$

em que y_{ijk} é o logaritmo na base 2 do valor de intensidade de expressão observado no *array* i (A_i , $i = 1, 2$), no tratamento j (T_j , $j = 1, 2$), no gene g ($g = 1, \dots, 273$), no *spot* k ($k = 1, 2, 3$), G_g é a média para o gene g , $(TG)_{jg}$ é o efeito do tratamento j no gene g , $(AG)_{igk}$ é o efeito do k -ésimo *spot* do *array* i no gene g , e ε_{ijk} é o resíduo do modelo para o gene g em que $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$. O termo $(TG)_{jg}$ representa o efeito de interesse.

Sejam τ_{1g} e τ_{2g} as verdadeiras médias de $(TG)_{1g}$ e $(TG)_{2g}$, respectivamente. A hipótese a ser testada será $H_0 : \tau_{1g} = \tau_{2g}$. Como pode ser observado na Tabela 3 o número de graus de liberdade do resíduo é igual a cinco. O teste utilizado para testar $H_0 : \tau_{1g} = \tau_{2g}$ foi o teste t com 5 graus de liberdade. Também foi utilizado um ajuste por FDR para controlar a taxa de falsos positivos.

Tabela 3 - Esquema de análise de variância

Causa da variação	GL
Tratamentos	1
<i>Arrays (Spots)</i>	5
Resíduo	5
Total	11

3.2.2.4 Método 4

Foi ajustado um modelo de análise de variância em dois estágios, proposto por Wolfinger et al. (2001), em que todos os efeitos foram considerados fixos. O primeiro modelo (modelo de normalização) foi:

$$y_{ijgk} = \mu + A_i + T_j + (AT)_{ij} + r_{ijgk}$$

em que y_{ijgk} é o logaritmo na base 2 do valor de intensidade de expressão observado no *array* i (A_i , $i = 1, 2$), no tratamento j (T_j , $j = 1, 2$), no gene g ($g = 1, \dots, 273$), no *spot* k ($k = 1, 2, 3$), μ é a média geral, $(AT)_{ij}$ é a interação (tratamento $\times$ *array*) e r_{ijgk} são os resíduos do modelo em que $r_{ijgk} \sim N(0, \sigma_r^2)$. Note que os valores da variável resposta utilizados foram os logaritmos dos valores de intensidade ao invés de logaritmos da razão entre os valores de intensidade das amostras.

Os resíduos normalizados do primeiro modelo são então utilizados como entrada no modelo gênico (modelo de efeitos específicos para cada gene de interesse)

$$r_{gijk} = G_g + (TG)_{jg} + (AG)_{igk} + \varepsilon_{gijk}$$

em que r_{ijgk} são os resíduos do modelo anterior, G_g é a média do gene g , $(TG)_{jg}$ é o efeito do tratamento j no gene g , $(AG)_{igk}$ é o efeito do k -ésimo *spot* do *array* i no gene g , e ε_{gijk} é o

resíduo do modelo para o gene g em que $\varepsilon_{gijk} \sim N(0, \sigma_\varepsilon^2)$. O termo $(TG)_{jg}$ representa o efeito de interesse.

O ajuste do modelo gênico é feito ordenando o banco de dados por gene e então é considerado o intervalo de linhas do banco de dados correspondente ao gene para o qual o modelo será ajustado, ou seja, estão sendo considerados os efeitos envolvidos no modelo dentro de cada gene, assim, o efeito $(TG)_{jg}$ representa o efeito do tratamento j dentro do gene g . No *software R* isso é feito facilmente utilizando um comando para ordenar o banco de dados por gene e uma rotina para ajustar o modelo e armazenar o resultado para cada gene.

O número de graus de liberdade (g.l.) para os testes t será considerado igual ao número de graus de liberdade do resíduo do modelo gênico, o qual é obtido na coluna g.l. da Tabela 4.

Tabela 4 - Esquema de análise de variância do modelo gênico

Causa da variação	g.l.
Tratamentos	1
Arrays (<i>Spots</i>)	5
Resíduo	5
Total	11

Como pode ser observado na Tabela 4 o número de graus de liberdade do resíduo é igual a cinco, então os testes t terão cinco graus de liberdade. Foi utilizado o ajuste por FDR para controlar a taxa de falsos positivos.

3.2.2.5 Método 5

Casella (2008) considera um experimento de *microarray* para identificação de genes com expressão diferencial como um experimento em parcelas subdivididas. Foi ajustado um modelo em parcelas subdivididas, onde o *array* foi considerado como bloco, as amostras de cDNA (CsS e CsP) foram consideradas como os tratamentos aplicados às parcelas e os genes foram considerados como os tratamentos aplicados à subparcela. Cada subparcela foi considerada como os três *spots* contendo as sequências de um mesmo gene, assim a variável resposta foi

considerada como a média dos logaritmos das intensidades de expressão das três repetições de cada gene.

O modelo utilizado foi

$$y_{ijg} = \mu + A_i + T_j + \varepsilon_{ij} + G_g + (AG)_{ig} + (TG)_{jg} + \delta_{ijg}$$

em que y_{ijg} é a média dos logaritmos na base 2 dos valores de intensidade de expressão das três repetições (*spots*) do gene g no array i no tratamento j , μ é o efeito da média geral, A_i é o efeito do array i ($i = 1, 2$), T_j é o efeito do tratamento j ($j = 1, 2$) aplicado à parcela, ε_{ij} é o efeito do erro devido à parcela em que $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, G_g é o efeito do gene g ($g = 1, \dots, 273$), $(AG)_{ig}$ é a interação do array i com o gene g , $(TG)_{jg}$ é a interação entre o tratamento j e o gene g e δ_{ijg} é o efeito do erro devido à subparcela em que $\delta_{ijg} \sim N(0, \sigma_\delta^2)$.

O efeito de interesse novamente é $(TG)_{jg}$. Analisar esse efeito significa comparar as médias do tratamento (CsS versus CsP) dentro de cada gene, ou seja, é equivalente à comparar duas médias do tratamento da parcela em um mesmo nível do tratamento da subparcela. O esquema de análise de variância é mostrado na Tabela 5.

O desdobramento do efeito do tratamento dentro de cada gene é feito como descrito em Gomes (1977). Sejam $\mathbf{a}$ o número de tratamentos aplicados às parcelas, $\mathbf{b}$ o número de tratamentos aplicados às subparcelas e $\mathbf{r}$ o número de repetições (blocos). Sejam V_a e V_b as variâncias residuais da parcela e da subparcela, respectivamente.

Sejam τ_{1g} e τ_{2g} as verdadeiras médias de $(TG)_{1g}$ e $(TG)_{2g}$, respectivamente. A hipótese a ser testada será $H_0 : \tau_{1g} = \tau_{2g}$. Para um contraste $\beta_g = \tau_{2g} - \tau_{1g}$, o erro padrão do contraste será

$$s(\hat{\beta}_g) = \sqrt{2 \frac{(b-1)V_b + V_a}{br}}.$$

Tabela 5 - Esquema de análise de variância para o delineamento em parcelas subdivididas

Causa da variação	GL
<i>Array</i>	1
Tratamento	1
Resíduo (a)	1
(Parcelas)	(3)
Gene	272
Gene $\times$ <i>Array</i>	272
Gene $\times$ Tratamento	272
Resíduo (b)	272
Total	1091

O número de graus de liberdade para usar no teste t é obtido utilizando a aproximação de Satterthwaite

$$n' = \frac{\left[\frac{V_a + (b-1)V_b}{V_a^2 + \frac{(b-1)^2 V_b^2}{n_b}} \right]^2}{\frac{1}{n_a} + \frac{(b-1)^2}{n_b}}$$

O teste t, com n' graus de liberdade, será dado então por

$$t = \frac{\hat{\beta}_g}{s(\hat{\beta}_g)}.$$

Foi utilizado o ajuste por FDR para controlar a taxa de falsos positivos.

3.2.2.6 Método 6

Como os valores de intensidade de alguns genes não seguem uma distribuição normal utilizou-se também um teste não paramétrico para aumentar a confiabilidade dos resultados.

Foi feito o teste da soma de postos de Wilcoxon para comparar as médias das amostras CsS e CsP para cada gene.

4 RESULTADOS E DISCUSSÃO

4.1 Normalização dos dados

Para verificar a eficiência da normalização com a sonda *Overgo* foram consideradas as três repetições do gene 1 (*spots* 1, 2 e 3) no *array* A₁ referentes à amostra CsS (folha de *Citrus sinensis* sadia). Já que os três *spots* contêm as sequências do mesmo gene e foram hibridizados com a mesma amostra era de se esperar que eles apresentassem níveis de expressão bem próximos. Pode-se observar na Figura (34) que antes da normalização com a sonda *Overgo* havia uma grande diferença do nível de expressão em um dos *spots* em relação aos outros dois e que após a divisão dos valores de intensidade do gene 1 pelos respectivos valores da sonda *Overgo* houve uma redução da variação entre os três.

Figura 34 - Gráficos de barras dos valores de intensidades de expressão das três repetições do Gene 1 no *array* A₁ referentes à amostra CsS (folha de *Citrus sinensis* sadia) antes (a) e após (b) a normalização com a sonda *Overgo*

Na Figura 35 podem ser observados os gráficos MA antes (a) e após (b) a normalização dependente da intensidade de expressão dentro de cada *array* utilizando regressão não paramétrica (*Loess*). Na Figura 36a pode-se observar pelos *boxplots* dos valores *M* dos dois *arrays* após a normalização dentro do *array* devida ao viés da intensidade de expressão que existe uma diferença na escala de variação entre os *arrays*. Na Figura 36b observa-se que após a normalização entre os *arrays* as distâncias interquartílicas dos *boxplots* ficaram mais semelhantes.

Figura 35 - Gráficos MA dos *arrays* A_1 e A_2 antes (A e B, respectivamente) e após (C e D) a normalização dentro do *array* utilizando regressão não paramétrica (*loess*)

Figura 36 - *Boxplot* dos *arrays* A_1 e A_2 após a normalização dentro do *array* (a) e após a normalização entre *arrays* (b)

4.2 Análise dos dados

4.2.1 Análise utilizando o método 1

Dos 273 genes analisados 119 genes apresentaram expressão diferencial significativa ao nível $\alpha = 5\%$ de probabilidade. Após a aplicação do procedimento FDR ($\alpha^* = 5\%$) 91 genes apresentaram expressão diferencial significativa, onde desses 91 genes, 66 se expressaram duas vezes mais em uma amostra do que na outra. Na Figura 37 pode-se observar o gráfico *Volcano* da análise com o método 1 em que os pontos acima e sobre a linha horizontal representam genes com p-valores menores ou iguais ao ponto de corte calculado pelo procedimento FDR $P_{(i)} = 0,01659938$, ou seja, genes com expressão diferencial significativa a um nível de significância conjunto $\alpha^* = 5\%$, pontos à esquerda da linha vertical da esquerda representam genes que estão se expressando duas vezes mais na amostra CsS em relação à amostra CsP e pontos à direita da linha vertical da direita representam genes que estão se expressando duas vezes mais na amostra CsP em relação à amostra CsS.

Figura 37 - Gráfico *Volcano* - Método 1

Na Tabela 6 pode ser observado o *rank* dos 10 genes com os menores *p*-valores (mais diferencialmente expressos) e *rank* dos 10 genes com maiores *fold-changes* para a análise com o método 1. Note que o gene 79, por exemplo, tem um grande valor do logaritmo da *fold-change*, porém o *p*-valor correspondente é maior do que o ponto de corte FDR, assim, apesar de a

diferença de expressão entre as duas amostras ser grande, estatisticamente essa diferença não foi significativa.

Tabela 6 - *Rank* dos 10 genes com menores *p*-valores e *rank* dos 10 genes com maiores *fold-changes* para a análise com o Método 1

Rank por <i>p</i>-valor			Rank por <i>fold-change</i>		
Gene	<i>p</i>-valor	LogFC	Gene	<i>p</i>-valor	LogFC
109	0,00000006	-2,45157600	141	0,00544636	4,10833184
249	0,00000060	-1,44073900	241	0,00281827	3,93956044
108	0,00000093	-1,96434700	79	0,02607396	3,76504797
201	0,00000332	-0,55989900	102	0,00194363	3,74821688
205	0,00000333	-2,74566500	288	0,01212249	3,68165380
83	0,00000503	-1,51406400	103	0,01224884	3,56900426
127	0,00000825	2,25998500	26	0,00096492	3,46200845
130	0,00000834	-1,95421500	172	0,00020639	2,99648692
155	0,00000859	-2,26768800	225	0,00036833	-2,94819261
244	0,00001162	2,79654900	244	0,00001162	2,79654886

4.2.2 Análise utilizando o método 2

Dos 273 genes analisados 60 genes apresentaram expressão diferencial significativa ao nível $\alpha = 5\%$ de probabilidade. Após a aplicação do procedimento FDR ($\alpha^* = 5\%$) 20 genes apresentaram expressão diferencial significativa, onde os 20 se expressaram duas vezes mais em uma amostra do que na outra. Na Tabela 7 pode ser observado o *rank* dos 10 genes com menores *p*-valores (mais diferencialmente expressos) e *rank* dos 10 genes com maiores *fold-changes* para a análise com o método 2.

Tabela 7 - *Rank* dos 10 genes com menores *p*-valores e *rank* dos 10 genes com maiores *fold-changes* para a análise com o Método 2

Rank por <i>p</i>-valor			Rank por <i>fold-change</i>		
Gene	<i>p</i>-valor	LogFC	Gene	<i>p</i>-valor	LogFC
283	0,00011824	2,74097316	141	0,00088698	4,10833184
109	0,00013652	-2,45157641	241	0,00067869	3,93956044
172	0,00016375	2,99648692	79	0,38301498	3,76504797
88	0,00021342	2,43045466	102	0,00032914	3,74821688
205	0,00024978	-2,74566490	288	0,00985102	3,68165380
102	0,00032914	3,74821688	103	0,00356513	3,56900426

(continua)

Tabela 7 - *Rank* dos 10 genes com menores *p*-valores e *rank* dos 10 genes com maiores *fold-changes* para a análise com o Método 2

(conclusão)					
<i>Rank por p-valor</i>			<i>Rank por fold-change</i>		
Gene	<i>p</i> -valor	LogFC	Gene	<i>p</i> -valor	LogFC
241	0,000679	3,93956	26	0,000707	3,462008
26	0,000707	3,462008	172	0,000164	2,996487
108	0,000707	-1,96435	225	0,000871	-2,94819
130	0,000741	-1,95421	244	0,002377	2,796549

Na Figura 38 pode-se observar o gráfico *Volcano* da análise com o método 2. O ponto de corte calculado pelo procedimento FDR foi $P_{(i)} = 0,003565126$.

Figura 38 - Gráfico *Volcano* - Método 2

4.2.3 Análise utilizando o método 3

Dos 273 genes analisados 127 genes apresentaram expressão diferencial significativa ao nível $\alpha = 5\%$ de probabilidade. Após a aplicação do procedimento FDR ($\alpha^* = 5\%$) 99 genes apresentaram expressão diferencial significativa, onde desses 99 genes 74 se expressaram duas vezes mais em uma amostra do que na outra. Na Figura 39 pode-se observar o gráfico *Volcano* da análise com o método 3. O ponto de corte calculado pelo procedimento FDR foi $P_{(i)} = 0,01625002$.

Na Tabela 8 pode ser observado o *Rank* dos 10 genes com menores *p*-valores (mais diferencialmente expressos) e *rank* dos 10 genes com maiores *fold-changes* para a análise com o método 3.

Figura 39 - Gráfico *Volcano* - Método 3

Tabela 8 - *Rank* dos 10 genes com menores *p*-valores e *rank* dos 10 genes com maiores *fold-changes* para a análise com o Método 3

<i>Rank por p-valor</i>			<i>Rank por fold-change</i>		
Gene	<i>p</i> -valor	LogFC	Gene	<i>p</i> -valor	LogFC
109	6E-08	-2,45158	141	0,005422	4,108332
249	6E-07	-1,44074	241	0,002773	3,93956
108	9,3E-07	-1,96435	79	0,002788	3,765048
83	1,13E-06	-1,51406	102	0,001941	3,748217
205	1,64E-06	-2,74567	288	0,011479	3,681654
201	2,22E-06	-0,5599	103	0,012164	3,569004
244	3,08E-06	2,796549	26	0,000937	3,462008
92	3,45E-06	0,878026	172	0,000205	2,996487
132	4,08E-06	1,870617	225	0,000367	-2,94819
273	5,04E-06	-1,35338	244	3,08E-06	2,796549

4.2.4 Análise utilizando o método 4

Dos 273 genes analisados 122 genes apresentaram expressão diferencial significativa ao nível $\alpha = 5\%$ de probabilidade. Após a aplicação do procedimento FDR ($\alpha^* = 5\%$) 95 genes

apresentaram expressão diferencial significativa, onde desses 95 genes 80 se expressaram duas vezes mais em uma amostra do que na outra. Na Figura 40 pode-se observar o gráfico *Volcano* da análise com o método 4. O ponto de corte calculado pelo procedimento FDR foi $P_{(i)} = 0,01561277$. Na Tabela 9 pode ser observado o *Rank* dos 10 genes com menores *p*-valores (mais diferencialmente expressos) e *rank* dos 10 genes com maiores *fold-changes* para a análise com o método 4.

Figura 40 - Gráfico *Volcano* - Método 4

Tabela 9 - *Rank* dos 10 genes com menores *p*-valores e *rank* dos 10 genes com maiores *fold-changes* para a análise com o Método 4

<i>Rank por p-valor</i>			<i>Rank por fold-change</i>		
Gene	<i>p</i> -valor	LogFC	Gene	<i>p</i> -valor	LogFC
109	1,81627E-08	-2,550745	141	0,004467587	4,009163
108	6,43968E-08	-2,063516	241	0,002123496	3,840391
205	8,07132E-08	-2,844834	79	0,002870524	3,665879
273	2,69235E-07	-1,452552	102	0,001404036	3,649048
155	8,4226E-07	-2,366857	288	0,009917045	3,582485
217	1,20899E-06	1,344641	103	0,01048601	3,469835
83	1,76912E-06	-1,613233	26	0,000602392	3,362839
59	2,05185E-06	-1,193818	225	0,000642171	-3,047362
184	2,14369E-06	-1,056228	172	8,79763E-05	2,897318
244	3,52613E-06	2,697380	205	8,07132E-08	-2,844834

4.2.5 Análise utilizando o método 5

Pode-se observar na Tabela 10 que a interação a interação foi “Gene x Amostra de cDNA” foi significativa. Foi feito então o desdobramento do efeito do tratamento principal dentro de cada nível do tratamento secundário. O erro padrão para cada contraste $\beta_g = \tau_{2g} - \tau_{1g}$ foi $s(\hat{\beta}_g) = 1,008097$ e os testes t para cada contraste foram realizados com $n' = 259,5375$ graus de liberdade. Observe que $\beta_g = \tau_{2g} - \tau_{1g}$ é igual ao logaritmo da *fold-change*. Dos 273 genes analisados 26 apresentaram expressão diferencial significativa ao nível $\alpha = 5\%$ de probabilidade. Após a aplicação do procedimento FDR ($\alpha^* = 5\%$) apenas 7 genes apresentaram expressão diferencial significativa, onde os 7 se expressaram duas vezes mais em uma amostra do que na outra.

Tabela 10 - Análise de Variância -Método 5

Causa da variação	g.l.	S.Q.	Q.M.	F	Pr(>F)
Array	1	0,4137	0,4137	-	-
Amostra de cDNA (<i>Trat. Principal</i>) A	1	2,6848	2,6848	0,5555	0,5923 ^{NS}
Resíduo (a)	1	4,8335	4,8335	-	-
(Parcelas)	(3)	7,9320	-	-	-
Gene (<i>Trat. Secundário</i>) B	272	4264,4	15,7	15,6431	< 2,2e-16 ^{**}
Gene x Amostra de cDNA	272	389,0	1,4	1,4268	0,001745 ^{**}
Gene x Array	272	125,5	0,5	0,4604	1 ^{NS}
Resíduo (b)	272	272,6	1	-	-
Total	1091	5059,4320	-	-	-

Nota – Sinais convencionais utilizados:

* Significativo a 5%.

** Significativo a 1%.

NS Não significativo.

Na Figura 41 pode-se observar o gráfico *Volcano* da análise com o método 5, onde os pontos acima da linha horizontal pontilhada representam genes com p -valores menores que $\alpha = 5\%$ (sem o ajuste FDR) e pontos acima e sobre a linha horizontal sólida representam genes com p -valores menores ou iguais ao ponto de corte calculado pelo procedimento FDR. Observe que este gráfico difere dos gráficos dos outros métodos quanto à dispersão dos pontos devido ao fato de que para o cálculo das estatísticas t de todos os genes foi utilizado o mesmo erro padrão dos contrastes. O ponto de corte calculado pelo procedimento FDR foi $P_{(i)} = 0,0006916602$.

Figura 41 - Gráfico *Volcano* - Método 5

Na Tabela 11 pode-se observar os 7 genes com diferenças de expressão significativas pela análise com o método 5. Note que foram considerados diferencialmente expressos apenas genes com valores grandes dos logaritmos da *fold-change*. A estatística $t = \hat{\beta}_g / s(\hat{\beta}_g)$ calculada para cada gene utilizou o mesmo erro padrão do contraste, assim, diferenças significativas são devidas a contrastes com valores altos, ou seja, *fold-change* com valores altos. Quando a análise é feita ajustando um modelo para cada gene separadamente então cada gene terá uma estimativa de variância diferente e, um grande valor da estatística t pode ser devido a uma variância pequena.

Tabela 11 - *Rank* dos 7 genes com menores p -valores e *rank* dos 7 genes com maiores *fold-changes* para a análise com o Método 5

Rank por p-valor			Rank por <i>fold-change</i>		
Gene	p-valor	LogFC	Gene	p-valor	LogFC
141	0,000061	4,108331841	141	0,000061	4,108331841
241	0,000119	3,939560435	241	0,000119	3,939560435
79	0,000231	3,765047966	79	0,000231	3,765047966
102	0,000246	3,748216878	102	0,000246	3,748216878
288	0,000315	3,681653798	288	0,000315	3,681653798
103	0,000473	3,569004256	103	0,000473	3,569004256
26	0,000692	3,462008455	26	0,000692	3,462008455

4.2.6 Análise utilizando o método 6

Dos 273 genes analisados 57 não apresentaram distribuição normal pelo teste de Shapiro-Wilk. Foi feito um teste de Wilcoxon para todos os genes para comparar com os resultados das outras análises, porém quando os dados seguem distribuição normal o teste t proporciona maior poder à análise do que o teste de Wilcoxon (Wit e McClure, 2004).

Dos 273 genes analisados 109 genes apresentaram expressão diferencial significativa ao nível $\alpha = 5\%$ de probabilidade. Após a aplicação do procedimento FDR ($\alpha^* = 5\%$) 92 genes apresentaram expressão diferencial significativa, onde desses 92 genes 77 genes se expressaram duas vezes mais em uma amostra do que na outra. Na Figura 42 pode-se observar o gráfico *Volcano* para a análise com o método 6. O ponto de corte calculado pelo procedimento FDR foi $P_{(i)} = 0,01515152$. Os 62 genes mais diferencialmente expressos por esta análise apresentaram p-valor = 0,002164502.

Figura 42 - Gráfico *Volcano* - Método 6

Dos 273 genes analisados 109 genes apresentaram expressão diferencial significativa ao nível $\alpha = 5\%$ de probabilidade. Após a aplicação do procedimento FDR ($\alpha^* = 5\%$) 92 genes apresentaram expressão diferencial significativa, onde desses 92 genes 77 genes se expressaram

duas vezes mais em uma amostra do que na outra. Na Figura 42 pode-se observar o gráfico *Volcano* para a análise com o método 2. O ponto de corte calculado pelo procedimento FDR foi $P_{(i)} = 0,01515152$.

4.3 Comparação dos resultados

A Tabela 12 apresenta o número de genes identificados como diferencialmente expressos em dois métodos simultaneamente.

Tabela 12 - Número de genes diferencialmente expressos simultaneamente

Métodos	1	2	3	4	5	6
1	91	20	91	80	06	77
2	-	20	20	20	05	20
3	-	-	99	85	07	77
4	-	-	-	94	07	72
5	-	-	-	-	07	06
6	-	-	-	-	-	92

Nota – Sinais utilizados:

- Mesmo valor do elemento simétrico a este em relação à diagonal principal.

Note que cada valor da diagonal da Tabela 12 representa o número de genes identificados como diferencialmente expressos em um único método, assim pode-se notar que os métodos 1, 3, 4 e 6 identificaram aproximadamente a mesma quantidade de genes como diferencialmente expressos. A Tabela 13 apresenta os resultados das análises para cada gene em cada método quanto à significância da diferença de expressão. Genes com expressão diferencial significativa em um determinado método foram representados por “*” e genes com expressão diferencial não significativa foram representados por “NS”. Os genes que não atenderam a suposição de normalidade foram representados pelo símbolo “♣”.

Tabela 13 - Significância da diferença de expressão para cada gene em cada método

(continua)

Gene	Método						
	1	2	3	4	5	6	
1	*	NS	*	*	NS	*	♣
2	*	NS	*	*	NS	*	♣
3	NS	NS	NS	NS	NS	NS	
5	*	NS	*	*	NS	*	
7	*	NS	*	*	NS	*	♣
8	NS	NS	NS	NS	NS	NS	
9	NS	NS	NS	NS	NS	NS	♣
10	*	NS	*	*	NS	NS	
11	NS	NS	NS	NS	NS	NS	♣
12	NS	NS	NS	NS	NS	NS	
13	NS	NS	NS	NS	NS	NS	
14	*	NS	*	*	NS	*	
15	NS	NS	NS	NS	NS	NS	♣
16	*	NS	*	*	NS	*	
17	NS	NS	NS	*	NS	NS	
18	NS	NS	NS	NS	NS	NS	
19	NS	NS	NS	NS	NS	NS	
20	NS	NS	NS	NS	NS	NS	
21	NS	NS	NS	NS	NS	NS	
22	NS	NS	NS	NS	NS	NS	
23	NS	NS	*	NS	NS	NS	
24	NS	NS	NS	NS	NS	*	♣
25	*	NS	*	*	NS	*	
26	*	*	*	*	*	*	♣
27	NS	NS	NS	NS	NS	NS	
28	NS	NS	NS	NS	NS	NS	
29	NS	NS	NS	NS	NS	NS	
30	NS	NS	NS	NS	NS	NS	
31	NS	NS	NS	NS	NS	NS	
32	NS	NS	NS	NS	NS	NS	
33	*	NS	*	*	NS	*	
34	NS	NS	NS	NS	NS	NS	
35	*	NS	*	*	NS	NS	
36	NS	NS	NS	NS	NS	NS	
37	NS	NS	NS	*	NS	NS	
38	*	*	*	*	NS	*	
39	NS	NS	NS	NS	NS	NS	♣
40	NS	NS	NS	*	NS	*	
41	NS	NS	NS	NS	NS	NS	
42	NS	NS	NS	NS	NS	NS	
43	NS	NS	NS	NS	NS	NS	♣

Gene	Método						
	1	2	3	4	5	6	
44	NS	NS	NS	NS	NS	NS	
45	NS	NS	NS	NS	NS	NS	
46	*	NS	*	*	NS	NS	
47	NS	NS	NS	NS	NS	NS	
48	*	NS	*	*	NS	*	
49	NS	NS	NS	NS	NS	NS	
50	NS	NS	NS	NS	NS	NS	
51	NS	NS	NS	NS	NS	NS	
52	*	NS	*	*	NS	*	
53	*	NS	*	NS	NS	*	
54	*	NS	*	NS	NS	*	
55	NS	NS	NS	NS	NS	NS	
56	NS	NS	NS	NS	NS	NS	
57	NS	NS	NS	NS	NS	NS	
58	*	NS	*	*	NS	*	
59	*	NS	*	*	NS	*	
60	*	NS	*	*	NS	*	♣
61	NS	NS	NS	NS	NS	NS	
62	NS	NS	NS	NS	NS	NS	♣
63	*	NS	*	*	NS	*	
64	*	NS	*	*	NS	*	♣
65	*	NS	*	*	NS	*	
67	NS	NS	NS	NS	NS	NS	
68	NS	NS	NS	NS	NS	NS	♣
69	NS	NS	NS	NS	NS	NS	♣
71	NS	NS	NS	NS	NS	NS	
72	*	NS	*	NS	NS	*	♣
73	NS	NS	NS	NS	NS	*	
74	NS	NS	NS	NS	NS	NS	
75	NS	NS	NS	NS	NS	NS	♣
76	NS	NS	NS	NS	NS	NS	
77	*	NS	*	NS	NS	*	
78	NS	NS	NS	NS	NS	NS	
79	NS	NS	*	*	*	NS	
80	NS	NS	NS	NS	NS	NS	
81	NS	NS	NS	NS	NS	NS	
83	*	NS	*	*	NS	*	
85	*	NS	*	*	NS	NS	
87	NS	NS	NS	NS	NS	NS	
88	*	*	*	*	NS	*	
89	NS	NS	NS	NS	NS	*	♣

Tabela 13 - Significância da diferença de expressão para cada gene em cada método

(continuação)

Gene	Método						Gene	Método					
	1	2	3	4	5	6		1	2	3	4	5	6
90	NS	NS	NS	NS	NS	NS	133	NS	NS	NS	*	NS	* ♣
91	NS	NS	NS	NS	NS	NS	134	NS	NS	NS	NS	NS	NS
92	*	NS	*	*	NS	NS	135	NS	NS	NS	NS	NS	NS
93	NS	NS	NS	NS	NS	NS	136	NS	NS	NS	NS	NS	NS
94	NS	NS	NS	NS	NS	NS	137	NS	NS	NS	NS	NS	NS
95	NS	NS	NS	NS	NS	NS	138	NS	NS	NS	NS	NS	NS
96	NS	NS	NS	NS	NS	NS	139	NS	NS	NS	NS	NS	NS
97	NS	NS	NS	NS	NS	NS	140	NS	NS	NS	NS	NS	NS
98	*	NS	*	*	NS	*	141	*	*	*	*	*	*
99	*	NS	*	*	NS	*	142	NS	NS	NS	NS	NS	NS ♣
100	NS	NS	NS	NS	NS	*	143	NS	NS	NS	NS	NS	NS
101	NS	NS	NS	NS	NS	NS	144	*	NS	*	*	NS	*
102	*	*	*	*	*	*	145	*	NS	*	*	NS	*
103	*	*	*	*	*	*	146	NS	NS	NS	NS	NS	NS
104	NS	NS	NS	NS	NS	NS	147	NS	NS	NS	NS	NS	NS
105	NS	NS	NS	NS	NS	NS	148	*	NS	*	*	NS	* ♣
106	NS	NS	NS	NS	NS	NS	149	NS	NS	NS	NS	NS	NS
107	*	NS	*	*	NS	NS	150	NS	NS	NS	NS	NS	NS
108	*	*	*	*	NS	*	151	NS	NS	NS	NS	NS	NS
109	*	*	*	*	NS	*	152	NS	NS	NS	NS	NS	NS
110	NS	NS	NS	NS	NS	NS	153	NS	NS	NS	NS	NS	NS
111	NS	NS	*	NS	NS	NS	154	NS	NS	NS	NS	NS	NS
112	NS	NS	NS	NS	NS	NS	155	*	*	*	*	NS	*
113	*	NS	*	*	NS	*	156	NS	NS	NS	NS	NS	*
114	NS	NS	NS	NS	NS	NS	157	NS	NS	NS	NS	NS	NS ♣
115	*	NS	*	*	NS	*	158	NS	NS	NS	NS	NS	NS
116	*	NS	*	NS	NS	*	159	NS	NS	NS	NS	NS	NS
117	NS	NS	NS	NS	NS	NS	160	NS	NS	*	*	NS	NS
118	NS	NS	NS	NS	NS	NS	161	NS	NS	NS	NS	NS	NS
119	NS	NS	NS	NS	NS	NS	162	*	NS	*	NS	NS	* ♣
121	NS	NS	NS	NS	NS	NS	163	NS	NS	NS	NS	NS	NS
122	*	NS	*	*	NS	*	164	*	NS	*	*	NS	NS ♣
123	NS	NS	NS	NS	NS	NS	165	*	NS	*	*	NS	*
124	NS	NS	NS	NS	NS	NS	166	NS	NS	*	*	NS	NS
125	NS	NS	*	*	NS	NS	167	*	NS	*	*	NS	* ♣
126	*	NS	*	NS	NS	NS	168	*	NS	*	*	NS	* ♣
127	*	*	*	*	NS	*	169	*	NS	*	NS	NS	NS
129	NS	NS	NS	NS	NS	NS	170	*	NS	*	*	NS	*
130	*	*	*	*	NS	*	171	NS	NS	NS	NS	NS	NS
131	NS	NS	NS	NS	NS	NS	172	*	*	*	*	NS	* ♣
132	*	*	*	*	NS	*	173	*	NS	*	*	NS	*

Tabela 13 - Significância da diferença de expressão para cada gene em cada método

(continuação)

Gene	Método						Gene	Método					
	1	2	3	4	5	6		1	2	3	4	5	6
175	NS	NS	NS	NS	NS	NS	217	*	NS	*	*	NS	*
176	NS	NS	NS	NS	NS	NS	218	NS	NS	NS	NS	NS	NS
177	*	NS	*	*	NS	NS	219	*	NS	*	*	NS	*
178	NS	NS	NS	NS	NS	NS	220	NS	NS	NS	NS	NS	NS
179	*	NS	*	*	NS	*	221	NS	NS	*	NS	NS	NS
180	NS	NS	NS	NS	NS	NS	222	*	NS	*	NS	NS	*
181	NS	NS	NS	*	NS	*	223	NS	NS	NS	NS	NS	NS
182	NS	NS	NS	NS	NS	*	224	*	NS	*	NS	NS	NS
183	NS	NS	NS	NS	NS	NS	225	*	*	*	*	NS	*
184	*	NS	*	*	NS	*	226	NS	NS	NS	NS	NS	*
185	*	*	*	*	NS	*	227	*	NS	*	*	NS	*
186	NS	NS	NS	NS	NS	NS	228	*	NS	*	*	NS	*
187	NS	NS	NS	NS	NS	NS	229	NS	NS	NS	NS	NS	*
188	NS	NS	NS	NS	NS	NS	230	NS	NS	NS	NS	NS	NS
189	NS	NS	NS	NS	NS	NS	232	*	NS	*	*	NS	*
190	NS	NS	NS	NS	NS	NS	233	NS	NS	NS	*	NS	NS
191	NS	NS	NS	NS	NS	NS	234	NS	NS	NS	NS	NS	NS
192	NS	NS	NS	NS	NS	NS	236	NS	NS	NS	NS	NS	NS
193	NS	NS	NS	NS	NS	NS	238	NS	NS	NS	NS	NS	NS
194	NS	NS	NS	NS	NS	NS	239	NS	NS	NS	NS	NS	NS
195	NS	NS	NS	NS	NS	NS	240	NS	NS	NS	NS	NS	NS
196	NS	NS	NS	NS	NS	NS	241	*	*	*	*	*	*
197	NS	NS	NS	NS	NS	*	242	NS	NS	NS	NS	NS	NS
198	NS	NS	NS	NS	NS	*	243	NS	NS	NS	NS	NS	NS
199	NS	NS	NS	NS	NS	NS	244	*	*	*	*	NS	*
200	*	NS	*	*	NS	*	245	NS	NS	NS	NS	NS	NS
201	*	NS	*	*	NS	NS	246	NS	NS	NS	NS	NS	NS
202	NS	NS	NS	NS	NS	NS	247	NS	NS	NS	NS	NS	NS
204	NS	NS	NS	NS	NS	NS	248	*	NS	*	*	NS	*
205	*	*	*	*	NS	*	249	*	NS	*	*	NS	*
206	NS	NS	NS	NS	NS	NS	250	*	NS	*	*	NS	*
207	NS	NS	NS	NS	NS	NS	251	NS	NS	NS	NS	NS	*
208	*	NS	*	*	NS	*	252	NS	NS	NS	NS	NS	NS
209	NS	NS	NS	NS	NS	NS	253	NS	NS	NS	NS	NS	NS
210	*	NS	*	*	NS	*	254	*	NS	*	NS	NS	*
211	NS	NS	*	*	NS	NS	255	NS	NS	NS	NS	NS	NS
212	NS	NS	NS	NS	NS	NS	256	NS	NS	NS	NS	NS	NS
213	NS	NS	NS	NS	NS	NS	257	NS	NS	NS	NS	NS	NS
214	NS	NS	NS	NS	NS	NS	258	NS	NS	NS	NS	NS	*
215	NS	NS	NS	NS	NS	NS	259	NS	NS	NS	NS	NS	NS
216	*	NS	*	*	NS	*	260	*	NS	*	*	NS	*

Tabela 13 - Significância da diferença de expressão para cada gene em cada método

							(conclusão)						
Gene	Método						Gene	Método					
	1	2	3	4	5	6		1	2	3	4	5	6
262	NS	NS	NS	*	NS	NS	277	NS	NS	NS	NS	NS	NS ♣
263	NS	NS	NS	NS	NS	NS	278	NS	NS	NS	NS	NS	NS
264	NS	NS	NS	NS	NS	NS	279	NS	NS	NS	NS	NS	NS
265	NS	NS	NS	NS	NS	NS	280	NS	NS	NS	*	NS	NS
266	*	*	*	*	NS	*	281	NS	NS	NS	NS	NS	NS ♣
267	NS	NS	NS	NS	NS	NS	282	NS	NS	NS	NS	NS	NS
269	NS	NS	NS	NS	NS	NS	283	*	*	*	*	NS	*
271	*	NS	*	*	NS	*	284	*	NS	*	*	NS	NS
272	NS	NS	NS	NS	NS	NS	285	NS	NS	NS	NS	NS	NS
273	*	NS	*	*	NS	*	286	NS	NS	NS	NS	NS	NS
274	*	NS	*	*	NS	NS	287	NS	NS	NS	*	NS	NS
275	*	NS	*	*	NS	*	288	*	NS	*	*	*	*
276	NS	NS	NS	NS	NS	NS							

Nota – Sinais utilizados:

* Significativo ao nível conjunto de 5% pelo procedimento FDR.

NS Não significativo ao nível conjunto de 5% pelo procedimento FDR.

♣ Não atendeu a suposição de normalidade pelo teste de Shapiro-Wilk.

Dos 273 genes analisados 153 não apresentaram diferenças de expressão significativas em nenhum dos métodos, 120 apresentaram diferenças de expressão significativas em pelo menos um dos métodos, 99 apresentaram diferenças de expressão significativas em pelo menos dois métodos, 89 apresentaram diferenças de expressão significativas em pelo menos três métodos, 69 apresentaram diferenças de expressão significativas em pelo menos quatro métodos, 21 apresentaram diferenças de expressão significativas em pelo menos cinco métodos e apenas 5 genes apresentaram diferenças de expressão significativas nos seis métodos.

Na Tabela 14 pode-se observar os 21 genes que se expressaram em pelo menos 5 métodos. Note que alguns desses genes não satisfazem a suposição de normalidade, o teste de Wilcoxon pode então auxiliar a tirar conclusões sobre estes genes. Como os genes que não atenderam a suposição de normalidade foram identificados como diferencialmente expressos pelo teste de Wilcoxon e pela maioria dos outros métodos pode-se dizer com maior confiabilidade que eles são realmente diferencialmente expressos.

Tabela 14 - Genes que se expressaram diferencialmente em pelo menos cinco métodos

Gene	Métodos						
	1	2	3	4	5	6	
26	*	*	*	*	*	*	♣
38	*	*	*	*	NS	*	
88	*	*	*	*	NS	*	
102	*	*	*	*	*	*	♣
103	*	*	*	*	*	*	
108	*	*	*	*	NS	*	♣
109	*	*	*	*	NS	*	
127	*	*	*	*	NS	*	♣
130	*	*	*	*	NS	*	
132	*	*	*	*	NS	*	♣
141	*	*	*	*	*	*	
155	*	*	*	*	NS	*	
172	*	*	*	*	NS	*	♣
185	*	*	*	*	NS	*	
205	*	*	*	*	NS	*	♣
225	*	*	*	*	NS	*	
241	*	*	*	*	*	*	
244	*	*	*	*	NS	*	
266	*	*	*	*	NS	*	♣
283	*	*	*	*	NS	*	
288	*	NS	*	*	*	*	

Segundo Haddad (2007) um problema de experimentos de *microarray* é que estimativa da variância para muitos genes pode ser subestimada e quando a variância de um gene é muito pequena, basta uma pequena diferença de expressão entre as duas amostras para que haja um valor grande da estatística *t* e, isso ocorre principalmente quando se tem poucas repetições.

A Figura 43 ilustra a distribuição dos denominadores da estatística *t* ordinária (Figura 43a) calculadas no método 1 e a distribuição dos denominadores da estatística *t* moderada (Figura 43b) calculadas no método 2. A maior parte dos valores dos denominadores da estatística *t* ordinária está concentrada entre 0 e 0,5 enquanto os valores dos denominadores da estatística *t* moderada estão concentrados em valores um pouco maiores.

Na Figura 44 pode-se observar as distribuições dos valores de probabilidade obtidos da estatística *t* ordinária (Figura 44a) e da estatística *t* moderada (Figura 44b).

Note que o fato de grande parte dos *p*-valores obtidos da estatística *t* ordinária estar

concentrada próxima de zero pode ser devido a um grande número de genes com pouca variabilidade (Figura 43a) e não necessariamente devido às diferenças nos níveis de expressão entre as duas amostras.

Figura 43 - Distribuição dos denominadores da estatística t ordinária (a) da estatística t moderada (b) para os 273 genes

Figura 44 - Distribuição dos valores de probabilidade obtidos das estatísticas t ordinárias (a) das estatísticas t moderadas (b)

Comparando os métodos 1 e 3 pode-se ver que todos os 91 genes identificados como diferencialmente expressos pelo método 1 também foram identificados como diferencialmente expressos pelo método 3, que além desses 91 genes identificou mais 8 como diferencialmente expressos. A inclusão do efeito de *spots* no método 3 isolou esse efeito e essa inclusão modificou a classificação dos genes considerando um *rank* de acordo com os *p*-valores, como pode ser observado nas Tabelas 6 e 8.

A diferença dos métodos 3 e 4 foi que o modelo 4 foi analisado em dois estágios em que o primeiro estágio utilizou um modelo considerando os fatores de variação globais, assim, foi feita uma correção dos valores da variável resposta quanto a essas fontes de variação, porém, devido às etapas de normalização que foram realizadas antes da análise propriamente dita essa correção não alterou muito os resultados, visto que os métodos 3 e 4 identificaram 85 genes simultaneamente como diferencialmente expressos.

O método 5 identificou 7 genes como diferencialmente expressos. Desses 7 genes, 6 foram identificados como diferencialmente expressos no método 1 e 5 foram identificados como diferencialmente expressos pelo método 2. Nos outros métodos todos os 7 genes também foram identificados como diferencialmente expressos. O método 5 considerou o mesmo erro padrão para todos os genes então estes 7 genes foram identificados como diferencialmente expressos devido ao grande valor da diferença entre as duas amostras. Na análise com o método 5 sem o procedimento FDR foram identificados 26 genes como diferencialmente expressos. Comparando com os outros métodos desses 26 genes 16 foram identificados como diferencialmente expressos pelo método 2 e nos outros métodos 23 desses 26 genes foram identificados como diferencialmente expressos.

Comparando os resultados obtidos no método 6 com os métodos 1, 3 e 4 observa-se que aproximadamente 70 genes foram identificados como diferencialmente expressos nesses quatro métodos simultaneamente. Como o método 6 identificou 92 genes como diferencialmente expressos então alguns desses genes não apresentaram diferenças significativas nos outros métodos, e, como alguns dos genes identificados como diferencialmente expressos apenas pelo método 6 não seguem distribuição normal talvez seja o caso de dar uma atenção especial para esses genes. Dos 20 genes identificados como diferencialmente expressos pelo método 2 constatou-se que 8 genes apresentavam distribuição não normal e, comparando os resultados para

esses 8 genes com os do método 6 verificou-se que os 8 genes tiveram diferença de expressão significativa pelo teste de Wilcoxon. Quando os dados seguirem uma distribuição normal é preferível utilizar o teste t, pois este proporciona mais poder à análise do que o teste de Wilcoxon quando atendida a suposição de normalidade.

5 CONCLUSÕES

Quando foi utilizada a estatística *t* ordinária como estatística de teste e os parâmetros foram estimados por mínimos quadrados as diferenças nos resultados da análise com o modelo linear considerando apenas o efeito de tratamento apresentou resultados muito semelhantes à análise com o modelo incluindo efeitos de média, *spots* e de tratamento.

A análise em dois estágios com um modelo de normalização e um modelo gênico, talvez devido às etapas de normalização realizadas antes da análise, também apresentou resultados semelhantes aos das análises com o modelo linear considerando apenas o efeito de tratamento e a análise com o modelo de análise de variância. Análises futuras podem considerar os efeitos de *array* e de *spot* como aleatórios.

A análise considerando um delineamento em parcelas subdivididas reduziu substancialmente o número de genes classificados como diferencialmente expressos, porém essa análise não considerou a variância individual de cada gene, fazendo com que fossem identificados como diferencialmente expressos apenas genes com grande diferença de expressão entre as amostras e, comparando com os outros métodos, alguns dos genes identificados como diferencialmente expressos por essa análise tinham grandes variâncias que se tivessem sido consideradas reduziria o valor da estatística de teste fazendo com que a diferença de expressão não fosse significativa.

A análise pelo teste de Wilcoxon pode auxiliar a tirar conclusões sobre genes cuja suposição de normalidade não foi atendida dando maior credibilidade à análise.

Verificou-se que grande parte dos genes identificados como diferencialmente expressos na maioria das análises foi devida à pequena variabilidade dentro dos genes e, a análise utilizando o método bayesiano empírico proporcionou estimativas de variâncias que amenizaram esse problema. Essa análise também levou em consideração a correlação entre as repetições dos genes. A análise com o método bayesiano empírico reduziu substancialmente o número de genes identificados como diferencialmente expressos, o que é muito interessante para um experimento de *macroarray*, tendo em vista que esta é uma análise preliminar para selecionar alguns genes que posteriormente serão estudados individualmente.

REFERÊNCIAS

- BANZATTO, D.A.; KRONKA, S. N. **Experimentação agrícola**. 3. ed. Jaboticabal: FUNEP, 1995. 247 p.
- BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society**, London, v. 57, n. 1, p. 289-300, 1995.
- BOLSTAD B.M.; IRIZARRY R.A.; ASTRAND M; SPEED T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. **Bioinformatics**, Oxford, v. 19, n. 2, 2003. Disponível em: <<http://dx.doi.org/10.1093/bioinformatics/19.2.185>>. Acesso em: 20 out. 2009.
- BORGES, L.C.; FERREIRA, D.F. Comparação de duas aproximações do teste t com variâncias heterogêneas através de simulação. **Ciência e Agrotecnologia**, Lavras, v.23, n.2, 1999. Disponível em: <http://www.editora.ufla.br/revista/23_2/art20.pdf> Acesso em: 01 out. 2009.
- CARAZZOLLE, M.F. **Análises de microarranjos de DNA**. Curso de extensão – Bioinformática na Prática I: Genômica e Transcriptômica, 2008, Universidade Estadual de Campinas, Campinas. Notas de Aula. Disponível em: <<http://www.lge.ibi.unicamp.br/lgeextensao2008/extsup/microarray.pdf>>. Acesso em: 01 maio 2009.
- CASELLA, G. **Statistical Design**. New York: Springer, 2008. 307 p.
- COFFEY, C.S.; COFIELD, S.S. Parametric Linear Models. In: ALLISON, D.B. et al. **DNA microarrays and related genomics techniques: design, analysis, and interpretation of experiments**. Boca Raton: Chapman & Hall/CRC, 2006. chap. 12, p. 223-243.
- DUDOIT, S.; YANG, Y.H.; CALLOW, M.J.; SPEED, T.P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. **Statistica Sinica**, Taipei, v. 12, n. 1, 2002. Disponível em: <<http://www.stat.berkeley.edu/users/sandrine/Docs/Papers/sinica.final.pdf>>. Acesso em: 01 nov. 2009.
- EISEN, M. **ScanAlyze User Manual**. Stanford University, Stanford, 1998. Disponível em: <<http://rana.lbl.gov/EisenSoftware.htm>> Acesso em: ago. de 2008.
- ESTEVES, G.H. **Validação de procedimentos para medida de expressão gênica a partir de imagens de cDNA Microarray**. 2002. Dissertação (Mestrado em Ciências – Área de concentração: Oncologia) – Instituto Ludwig de Pesquisas sobre o Câncer, São Paulo, 2001. Disponível em: <http://www.vision.ime.usp.br/~gesteves/documents/esteves_2002.pdf>. Acesso em: 05 jun. 2008.

ESTEVEES, G.H. **Métodos estatísticos para a análise de dados de cDNA microarray em um ambiente computacional integrado.** 2007. Tese (Doutorado em Bioinformática) – Bioinformática (IME/IFSC/ESALQ/IQ/IB/ICB/FMVZ/FCFRP) – Universidade de São Paulo, São Paulo, 2007. Disponível em: <http://www.teses.usp.br/teses/disponiveis/95/95131/tde-03062007-210232/publico/tese_final.pdf>. Acesso em: 01 jun. 2009.

FUJITA, A. **Análise de dados de expressão gênica: normalização de *microarrays* e modelagem de redes regulatórias.** 2007. Tese (Doutorado em Ciências – Área de concentração: Bioinformática) – Bioinformática (IME/IFSC/ESALQ/IQ/IB/ICB/FMVZ/FCFRP) – Universidade de São Paulo, São Paulo, 2007. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/95/95131/tde-14092007-173758/publico/FujitaTeseBiblioteca.pdf>>. Acesso em: 22 Jul. 2008.

GENTLEMAN, R.C.; CAREY, V.J.; BATES, D.M.; BOLSTAD, B.; DETTLING, M.; DUDOIT, S.; ELLIS, B.; GAUTIER, L.; GE, Y.; GENTRY, J.; HORNIK, K.; HOTHORN, T.; HUBER, W.; IACUS, S.; IRIZARRY, R.; LEISCH, F.; LI, C.; MAECHLER, M.; ROSSINI, A.J.; SAWITZKI, G.; SMITH, C.; SMYTH, G.; TIERNEY, L.; YANG, J.Y.; ZHANG, J. Bioconductor: Open software development for computational biology and bioinformatics. **Genome Biology**, London, v. 5, p. R80, 2004. Disponível em: <<http://genomebiology.com/2004/5/10/R80>>. Acesso em: 15 Jul. 2008.

GOMES, F. P. **Curso de Estatística Experimental.** 7.ed. Piracicaba, Nobel, 1977. 430 p.

GUIDETTI-GONZALEZ, S. **Identificação de genes envolvidos na defesa contra patógenos no banco de dados do CitEST e em macroarranjos da interação *Citrus sinensis*-*Guignardia citricarpa*.** 2009. Tese (Doutorado em Ciências – Área de concentração: Bioquímica de Plantas) – Escola Superior de Agricultura “Luiz de Queiroz” – Universidade de São Paulo, Piracicaba, 2008. Disponível em: <http://www.teses.usp.br/teses/disponiveis/11/11144/tde-19052009-092239/publico/Simone_Gonzalez.pdf>. Acesso em: 30 maio 2009.

HADDAD, S. R. **Aplicação de modelos lineares para análise de expressão gênica em experimentos de microarrays.** 2007. Dissertação (Mestrado em Zootecnia – Área de concentração: Nutrição e Produção Animal) – Faculdade de Medicina Veterinária e Zootecnia – Universidade Estadual Paulista, Botucatu, 2007. Disponível em: <http://www.fmvz.unesp.br/PosGraduacao/Zootec/Dissertacoes_Teses/pg_aviso.php?nome=Samia_Haddad.pdf>. Acesso em: 15 abr. 2009.

HUBER, W.; HEYDEBRECK, A. von; VINGRON, M. Analysis of Microarray Gene Expression Data. In: Balding, D.J.; Bishop, M.; Cannings, C. **Handbook of statistical genetics**, West Sussex: John Wiley & Sons, 2003. chap. 6, p. 162-187.

JUNQUEIRA, L.C.; CARNEIRO, J. **Biologia Celular e Molecular.** 6.ed. Rio de Janeiro: Guanabara Koogan, 1997. 299 p.

KERR, M.K.; Churchill, G.A. Experimental design for gene expression microarrays. **Biostatistics**, Great Britain, v. 2, n. 2, p. 183-201, 2001.

KERR, M.K.; Afshari, C.A.; Bennett, L.; Bushel, P.; Martinez, J.; Walker, N.J.; Churchill, G.A. A statistical analysis of a gene expression microarray experiment with replication. **Statistica Sinica**, Taipei, v. 12, n. 2, p. 203-217, 2002. Disponível em: <<http://www3.stat.sinica.edu.tw/statistica/oldpdf/A12n110.pdf>>. Acesso em: 17 mar. 2009.

KERR, M.K. Linear Models for Microarray Data Analysis: Hidden Similarities and Differences. **Journal of Computational Biology**, New York, v. 10, n. 6, p. 891-901, 2003. Disponível em: <<http://www.bepress.com/cgi/viewcontent.cgi?article=1013&context=uwbiostat>>. Acesso em: 07 out. 2009.

LÖNNSTEDT, I. ; SPEED, T. Replicated microarray data. **Statistica Sinica**, Taipei, v. 12, n. 2, p. 31-46, 2002. Disponível em: <<http://www3.stat.sinica.edu.tw/statistica/j12n1/j12n12/j12n12.htm>> Acesso em: 13 out. 2009.

LOPES, F.C.; PAIS, H.L.M. **Exploração, filtragem e análise de dados laboratoriais**. Aveiro: Universidade de Aveiro. 2006. 79 p. Relatório de Projecto de Final de Curso. Disponível em: <<http://analismicroarray.no.sapo.pt/relatorio/Relatorio-Final.pdf>>. Acesso em: 20 jul. 2008.

PEREIRA, R.N. **Controle do erro do tipo I em um experimento de Microarrays com eucalipto**. 2008. 56 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2008.

R Development Core Team (2009). **R: A Language and Environment for Statistical Computing**. Vienna. Disponível em: <<http://www.R-project.org>>. Acesso em: 01 nov. 2009.

ROSA, F.H.F.P.; SOLER, J.M.P. **Relatório sobre o Projeto "Utilização de Métodos Estatísticos e desenvolvimento de ferramentas de bioinformática para análise de cDNA utilizando microarrays"**. Projeto Fapesp 03/02074-0. São Paulo. 2004. Disponível em: <<http://www.feferraz.net/br/documentos.php?numdoc=2>> . Acesso em: 03 mar. 2009.

SCHENA, M. et al., (Jul. 1998). Microarrays: biotechnology's discovery platform for functional genomics. **Trends in biotechnology**, Amsterdam, v. 16, n. 7, p. 301-306, 1998.

SMYTH, G.K.; Ritchie, M; Thorne, M.; Wettenhall, J. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. **Statistical Applications in Genetics and Molecular Biology**, Berkeley, v. 3, n. 1, 2004. Disponível em: <<http://www.statsci.org/smyth/pubs/ebayes.pdf>>. Acesso em: 11 nov. 2008.

SMYTH, G.K. Limma: linear models for microarray data. In: R. Gentleman et al. **Bioinformatics and Computational Biology Solutions using R and Bioconductor**, Springer, New York, 2005, chap. 23, p 397-420.

SMYTH, G.K.; Michaud, J.; Scott, H.S. The use of within-array replicate spots for assessing differential expression in microarray experiments. **Bioinformatics**, Oxford, v. 21, n. 9, p. 2067-2075, 2005. Disponível em: <<http://bioinformatics.oxfordjournals.org/cgi/content/short/21/9/2067>>. Acesso em: 09 jul. 2009.

SMYTH, G.K.; MICHAUD, J.; SCOTT, H. **limma**: Linear Models for Microarray Data User's Guide. Melbourne: The Walter and Eliza Hall Institute of Medical Research Melbourne. 2008. 95 p.

STAFFORD, P. **Methods in Microarray Normalization**. Boca Raton; CRC Press, 2008. 304p.

THORNE, N.; **Understanding the design matrix in linear models for microarray experiments**. Cambridge University - The Hutchison/MRC Research Center. Cambridge, 2006. Disponível em: <<http://www.damtp.cam.ac.uk/user/npt22/npt22webpages/designMatrixLects.pdf>>. Acesso em: 11 nov. 2008.

WIT, E.; MCCLURE, J. **Statistics for Microarrays**: design, analysis and inference. West Sussex: John Wiley & Sons, 2004. 265 p.

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. **Journal of Computational Biology**, New York, v. 8, n. 6, p. 625-637, 2001. Disponível em: <<http://statgen.ncsu.edu/ggibson/Publications/WGetc.pdf>>. Acesso em: 15 set. 2009.

YANG, Y.H.; DUDOIT, S.; LUU, P.; LIN, D.M.; PENG, V.; NGAI, J.; SPEED, T.P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. **Nucleic Acids Research**, Oxford, v. 30 n. 4, 2002. Disponível em: <<http://nar.oxfordjournals.org/cgi/reprint/30/4/e15>>. Acesso em: 23 maio 2008.

ANEXOS

ANEXO A – ENTRADA DOS DADOS NO R

Para ler os dados no R o banco de dados deve estar organizado de forma que as hibridizações estejam nas colunas e os genes estejam nas linhas. Como, no banco de dados analisados nesse trabalho, os genes foram repetidos três vezes dentro do *array*, então cada linha representará uma repetição do gene, de forma que cada gene estará representado em três linhas do banco de dados. Pode-se observar na Tabela 14 a estrutura do banco de dados para leitura dos dados no R. A coluna 1 representa a amostra Overgo hibridizada no *array* A₁, a coluna 2 representa a amostra CsS hibridizada no *array* A₁ e assim por diante. A linha 1 representa o gene 1 na repetição (*spot*) 1, a linha 2 representa o gene 1 na repetição 2, a linha 3 representa o gene 1 na repetição 3, a linha 4 representa o gene 2 na repetição 1 e assim por diante.

Tabela 15 - Estrutura do banco de dados para leitura dos dados no R

<i>Overgo</i> A ₁	CsS A ₁	CsP A ₁	<i>Overgo</i> A ₂	CsS A ₂	CsP A ₂
1.428	1.944	6.089	2.111	0.601	5.1
2.45	4.068	13.702	2.717	0.844	9.64
1.603	2.056	8.527	1.885	0.567	5.093
1.492	2.514	7.297	2.294	1.484	10.262
2.241	3.77	13.389	2.104	1.419	9.769
1.926	3.128	10.495	2.431	1.805	12.872

Primeiramente, deve-se salvar o banco de dados como um arquivo de texto delimitado por tabulações. Observe que no lugar de vírgula são utilizados pontos. O banco de dados foi salvo com o nome Dados.txt.

```
#####
#Ler os dados no R
#####
```

```
dados<-read.table("C:/Microarray/Dados.txt", header=TRUE)
```

ANEXO B – ROTINAS PARA PRÉ-PROCESSAMENTO DOS DADOS

```
#####
#Fazer uma matriz contendo apenas os dados das quatro
#hibridizações
#####
dados.hib <- cbind(dados[,2:3],dados[5:6])
dados.hib <- as.matrix(dados.hib,nrow=864, ncol=4)

#####
#Primeira normalização (OVERGO)
#####
overgo.1 <- dados.hib[,1:2]/dados[,1]
overgo.2 <- dados.hib[,3:4]/dados[,4]
dados.norm1 <- cbind(overgo.1, overgo.2)

# Os arquivos Genes.txt e Tipo.txt contém os nomes dos genes
# (Gene 1,...,Gene 288) e o tipo de gene (EST, ACTINA, GAPDH)

#####
#Indexar Genes e Tipo
#####
Genes <- read.table("C:/Microarray/Genes.txt",header=TRUE)
Tipo <- read.table("C:/Microarray/Tipo.txt",header=TRUE)
dados.hib <- cbind(Genes,dados.norm1,Tipo)

#####
#removendo os genes controle
#####
dados.hib <- dados.hib[dados.hib$Tipo!="GAPDH",]
dados.hib <- dados.hib[dados.hib$Tipo!="ACTINA",]
dados.hib2 <- dados.hib[,2:5]
dados.hib2 <- as.matrix(dados.hib2)
nomes <- dados.hib$Gene
rownames(dados.hib2) <- nomes

#####
#Criar o objeto RG
#####
R <- dados.hib2[,c(2,4)]
G <- dados.hib2[,c(1,3)]
RG <- list(R=R,G=G)
```

```
#####
#Criar o objeto MA
#####
library(limma)
MA <- normalizeWithinArrays(RG,method="none")

#####
#Gráficos MA
#####
par(mfrow=c(2,2))
plot(MA$A[,1],MA$M[,1],lwd=2,cex=0.1,xlab="A",ylab="M",
main="(A)")
lines(lowess(MA$A[,1],MA$M[,1]),col="blue",lwd=2)
plot(MA$A[,2],MA$M[,2],lwd=2,cex=0.1,xlab="A",ylab="M",
main="(B)")
lines(lowess(MA$A[,2],MA$M[,2]),col="blue",lwd=2)

#####
#Normalização dentro do array
#####
MA2 <- normalizeWithinArrays(RG,method="loess")
colnames(MA2$M) <- c("Array1","Array2")
colnames(MA2$A) <- c("Array1","Array2")

#####
#Gráficos MA e Boxplot
#####
plot(MA2$A[,1],MA2$M[,1],lwd=2,cex=0.1,xlab="A",ylab="M",
main="(C)")
lines(lowess(MA2$A[,1],MA2$M[,1]),col="blue",lwd=2)
plot(MA2$A[,2],MA2$M[,2],lwd=2,cex=0.1,xlab="A",ylab="M",
main="(D)")
lines(lowess(MA2$A[,2],MA2$M[,2]),col="blue",lwd=2)

boxplot(MA2$M[,1],MA2$M[,2],col="lightgray",
names=c(expression(A[1]),expression(A[2])))

#####
#Normalização entre arrays
#####
MA3 <- normalizeBetweenArrays(MA2,method="scale")
boxplot(MA3$M[,1],MA3$M[,2],col="lightgray",
names=c(expression(A[1]),expression(A[2])))
```

ANEXO C – ROTINAS PARA A ANÁLISE COM O MÉTODO 1

```
#####
#Criar um objeto MA em que cada linha de M contenha
#todos os valores M de um único gene, e cada linha de
#A contenha todos os valores de A de um único gene.
#####
nomes2 <- nomes[seq(1,819,3)]
M4 <- unwrapdups(MA3$M,ndups=3)
A4 <- unwrapdups(MA3$A,ndups=3)
MA4 <- list(M=M4,A=A4,genes=nomes2)

#####
#Ajustando um modelo linear com a função lm()
#####
Resultado <- matrix(NA,ncol=4,nrow=273)
colnames(Resultado) <- c("Gene","P.Valor","LogFC")
Resultado <- as.data.frame(Resultado)

for(i in seq(1:273)){
  fm <- lm(MA4$M[i,]~1)
  Resultado[i,1] <- nomes2[i]
  Resultado[i,2] <- summary(fm)$coefficients[1,4]
  Resultado[i,3] <- mean(MA4$M[i,])
}

Resultado
Resultado[Resultado$P.Valor < 0.05,]

#####
#Ponto de corte FDR
#####
Pval <- sort(Resultado $P.Valor)
M <- length(pval)
fdr.cutoff <- NA
for(i in 1:m){
  q <- m*pval[i]/i
  if(q<0.05){
    fdr.cutoff <- pval[i]
  }
}
fdr.cutoff

Resultado[Resultado$P.Valor <= fdr.cutoff,]
```

```
#####  
#Gráfico Volcano  
#####  
plot(Resultado$LogFC,-log(Resultado $P.Valor,10),cex=0.1, lwd=3,  
xlab=expression(log[2]("fold change")),  
ylab=expression(-log[10]("p valor")),main="")  
abline(v=c(-1,1),col="blue")  
abline(h=-log(fdr.cutoff,10),col="red")
```


ANEXO D – ROTINAS PARA A ANÁLISE COM O MÉTODO 2

```
#####
#Calcular a correlação entre as réplicas dos genes
#dentro do array.
#####
design <- c(1,1)
corfit <- duplicateCorrelation(MA3, design, ndups = 3,
spacing = 1,block=NULL,weights=NULL)
corfit$consensus.correlation
boxplot(tanh(corfit$atanh.correlations),col="lightgray")

#####
#Ajustando um modelo linear
#####
fit <- lmFit(MA3,design,ndups=3,correlation=corfit$consensus)

#####
#Calculando as estatísticas t moderadas pelo método
#Bayesiano Empírico.
#####
fit <- eBayes(fit)
names(fit)

#####
#Resultado
#####
Resultado <- topTable(fit,number=288,adjust="BH")
Resultado

#####
#Ponto de corte FDR
#####
Pval <- sort(Resultado$P.Value)
m <- length(pval)
fdr.cutoff <- NA
for(i in 1:m){
q <- m*pval[i]/i
if(q<0.05){
fdr.cutoff <- pval[i]
}
}
fdr.cutoff
```

```
#####
#Gráfico Volcano
#####
plot(Resultado[,2],-log(Resultado[,5],10), cex=0.4,pch=19,
xlim=c(-4,5),main="",
xlab=expression(log[2]("fold change")),
ylab=expression(-log[10]("p valor")))
abline(v=c(-1,1),col="blue")
abline(h=-log(fdr.cutoff,10),col="red")
```

ANEXO E – ROTINAS PARA A ANÁLISE COM O MÉTODO 3

```
#####
#Obtendo os valores de log(Sadia*) e log(Doente*)
#####
LogDoente <- MA3$A + MA3$M/2
LogSadia <- MA3$A - MA3$M/2
dados.norm2 <- cbind(LogSadia[,1],LogDoente[,1],
LogSadia[,2],LogDoente[,2])
colnames(dados.norm2) <- c("Sadial","Doente1","Sadia2",
"Doente2")

#####
#Ajustar o banco de dados
#####

#Vetor dos logaritmos das intensidades:
v1 <- c(dados.norm2[,1],dados.norm2[,3],
dados.norm2[,2],dados.norm2[,4])
length(v1)

#Vetor dos Genes
v2 <- rep(nomes,4)
length(v2)

#Vetor dos tratamentos
v3 <- factor(c(rep("sadio",1638),rep("doente",1638)))
length(v3)

#Vetor dos arrays
v4 <- factor(c(rep(1,273*3),rep(2,273*3),
rep(1,273*3),rep(2,273*3)))
length(v4)

dados.final <- data.frame(v1,v2,v3,v4)
colnames(dados.final) <- c("Y","Gene","Trat","Array")
rownames(dados.final) <- NULL
dim(dados.final)
dados.final[1:20,]
attach(dados.final)
```

```
#####
#Anova para cada gene
#####

nomes2 <- nomes[seq(1,819,3)]
Resultado <- matrix(NA,ncol=3,nrow=273)
colnames(Resultado) <- c("Gene","P.Valor","LogFC")

for(i in 1:273){

dados.gene <- tapply(Y,data.frame(Array,Trat,Gene),
sort)[c(1,2),c(1,2),i]
dados.gene <- as.data.frame(dados.gene)

dados.gene2 <- c(dados.gene$sadio$`1`,dados.gene$sadio$`2`,
dados.gene$doente$`1`,dados.gene$doente$`2`)

trat.gene <- c(rep("sadio",6),rep("doente",6))
trat.gene <- as.factor(trat.gene)

spot.gene <- rep(c(11,12,13,21,22,23),2)
spot.gene <- as.factor(spot.gene)

fm <- lm(dados.gene2 ~ spot.gene + trat.gene)
logFC <- mean(dados.gene2[7:12])-mean(dados.gene2[1:6])

Resultado[i,1] <- nomes2[i]
Resultado[i,2] <- summary(fm)$"coefficients"[7,4]
Resultado[i,3] <- logFC

}

Resultado <- as.data.frame(Resultado)

#####
#Ponto de corte FDR
#####
Pval <- sort(Resultado$P.Valor)
M <- length(pval)
fdr.cutoff <- NA
for(i in 1:m){
q <- m*pval[i]/i
if(q<0.05){
fdr.cutoff <- pval[i]
}
}
}
```

fdr.cutoff

```
#####  
#Volcano Plot  
#####  
plot(Resultado$LogFC,-log(Resultado$P.Valor,10),cex=0.1,lwd=3,  
xlab=expression(log[2]("fold change")),  
ylab=expression(-log[10]("p valor")),main="")  
abline(v=c(-1,1),col="blue")  
abline(h=-log(fdr.cutoff,10),col="red")
```

ANEXO F – ROTINAS PARA A ANÁLISE COM O MÉTODO 4

```
#####
#Obtendo os valores de log(Sadia*) e log(Doente*)
#####
LogDoente <- MA3$A + MA3$M/2
LogSadia <- MA3$A - MA3$M/2
dados.norm2 <- cbind(LogSadia[,1],LogDoente[,1],
LogSadia[,2],LogDoente[,2])
colnames(dados.norm2) <- c("Sadia1","Doente1","Sadia2",
"Doente2")

#####
#Ajustar o banco de dados
#####

#Vetor dos logaritmos das intensidades:
v1 <- c(dados.norm2[,1],dados.norm2[,3],
dados.norm2[,2],dados.norm2[,4])
length(v1)

#Vetor dos Genes
v2 <- rep(nomes,4)
length(v2)

#Vetor dos tratamentos
v3 <- factor(c(rep("sadio",1638),rep("doente",1638)))
length(v3)

#Vetor dos arrays
v4 <- factor(c(rep(1,273*3),rep(2,273*3),
rep(1,273*3),rep(2,273*3)))
length(v4)

dados.final <- data.frame(v1,v2,v3,v4)
colnames(dados.final) <- c("Y","Gene","Trat","Array")
rownames(dados.final) <- NULL
dim(dados.final)
dados.final[1:20,]
attach(dados.final)

#####
#Modelo de Normalização
#####
fm1 <- aov(Y ~ Trat + Array + Trat:Array)
```

```

summary(fm1)
residuo <- resid(fm1)

#####
#Modelo Genico
#####

nomes2 <- nomes[seq(1,819,3)]
Resultado <- matrix(NA,ncol=3,nrow=273)
colnames(Resultado) <- c("Gene", "P.Valor", "LogFC")

trat.gene <- c(rep("sadio",6),rep("doente",6))
trat.gene <- as.factor(trat.gene)

spot.gene <- rep(c(11,12,13,21,22,23),2)
spot.gene <- as.factor(spot.gene)

for(i in 1:273){

dados.gene <- tapply(residuo,data.frame(Array,Trat,Gene),
sort)[c(1,2),c(1,2),i]
dados.gene <- as.data.frame(dados.gene)

resid.gene <- c(dados.gene$sadio$`1`,dados.gene$sadio$`2`,
dados.gene$doente$`1`,dados.gene$doente$`2`)

fm2 <- lm(resid.gene ~ spot.gene + trat.gene)
logFC <- mean(resid.gene[7:12])-mean(resid.gene[1:6])

Resultado[i,1] <- nomes2[i]
Resultado[i,2] <- (summary(fm2)$"coefficients"[7,4])
Resultado[i,3] <- logFC

}

Resultado <- as.data.frame(Resultado)

#####
#Ponto de corte FDR
#####
pval<-sort(Resultado$P.Valor)
m<-length(pval)
fdr.cutoff<-NA
for(i in 1:m){

```

```

q<-m*pval[i]/i
if(q<0.05){
fdr.cutoff<-pval[i]
}
}
fdr.cutoff

```

```

#####
#Volcano Plot
#####
plot(Resultado$LogFC,-log(Resultado$P.Valor,10),cex=0.1,lwd=3,
xlab=expression(log[2]("fold change")),
ylab=expression(-log[10]("p valor")),main="")
abline(v=c(-1,1),col="blue")
abline(h=-log(fdr.cutoff,10),col="red")

```


ANEXO G – ROTINAS PARA A ANÁLISE COM O MÉTODO 5

```
#####
#Obtendo os valores de log(Sadia*) e log(Doente*)
#####
LogDoente <- MA3$A + MA3$M/2
LogSadia <- MA3$A - MA3$M/2
dados.norm2 <- cbind(LogSadia[,1],LogDoente[,1],
LogSadia[,2],LogDoente[,2])
colnames(dados.norm2) <- c("Sadia1","Doente1","Sadia2",
"Doente2")

#####
#Calcular a media de cada gene
#####
dim(dados.norm2)
dados.media <- matrix(NA,ncol=4,nrow=273)
colnames(dados.media) <- c("Sadio1","Doente1",
"Sadio2","Doente2")

nomes2 <- nomes[seq(1,819,3)]
nomes2 <- as.vector(nomes2)
rownames(dados.media)<-paste("Gene",nomes2)

j=1
for(i in seq(1,nrow(dados.norm2),3)){
dados.media[j,] <- (dados.norm2[i,]+dados.norm2[i+1,]
+dados.norm2[i+2,])/3
j=j+1
}

#####
#Ajustar o banco de dados
#####

#Vetor dos logaritmos das intensidades:
v1 <- c(dados.media[,1], dados.media[,3],
dados.media[,2], dados.media[,4])
length(v1)

#Vetor dos genes
v2 <- rep(nomes2,4)
length(v2)
```

```

#Vetor dos tratamentos
v3 <- factor(c(rep("sadio",546),rep("doente",546)))
length(v3)

#Vetor dos arrays
v4 <- factor(c(rep(1,273),rep(2,273),rep(1,273),rep(2,273)))
length(v4)

dados.final <- data.frame(v1,v2,v3,v4)
colnames(dados.final) <- c("Y", "Gene", "Trat", "Array")
rownames(dados.final) <- NULL

#####
#Parcelas Subdivididas
#####
Gene <- as.factor(dados.final$Gene)
Trat <- as.factor(dados.final$Trat)
Array <- as.factor(dados.final$Array)
Y <- dados.final$Y

fml <- aov(Y ~ Array + Trat*Gene + Gene:Array +
Error(Array/Trat/Gene))
summary(fml)

#####
#Desdobramento Tratamento dentro de Gene
#####

#número de tratamentos (Parcela)
a=length(levels(Trat))
a

#número de tratamentos (Sub-parcela)
b=length(levels(Gene))
b

#número de repetições
r=length(Y)/(a*b)
r

#Anova Parcela
aov.P1<-summary(fml)$"Error: Array:Trat"
aov.P2<-unlist(aov.P1)
aov.P3<-as.data.frame(aov.P2)
aov.P4<-aov.P3[rownames(aov.P3)=="Mean Sq2",]

```

```

aov.P1
aov.P4

#QM resíduo (Parcela)
va=aov.P4
va

#Anova Sub-parcela
aov.SP1<-summary(fm1)$"Error: Array:Trat:Gene"
aov.SP2<-unlist(aov.SP1)
aov.SP3<-as.data.frame(aov.SP2)
aov.SP4<-aov.SP3[rownames(aov.SP3)=="Mean Sq4",]
aov.SP1
aov.SP4

#QM resíduo (Sub-parcela)
vb=aov.SP4
vb

#Graus de liberdade do resíduo (Parcela)
na=(a-1)*(r-1)
na

#Graus de liberdade do resíduo (Sub-parcela)
nb=(a-1)*(b-1)*(r-1)
nb

#Média de cada tratamento em cada gene
media<-tapply(Y,list(Gene,Trat),mean)
media[1:10,]
dim(media)

#Erro padrão do contraste
EPCont<-sqrt(2*((b-1)*vb+va)/(b*r))
EPCont

#Aproximação de Satterthwaite
nlinha<-((va+(b-1)*vb)^2)/(((va^2)/na)+(((b-1)^2)*(vb^2))/nb)
nlinha

Resultado<-matrix(NA,ncol=3,nrow=273)
colnames(Resultado)<-c("Gene", "LogFC", "P.Valor")

for(i in 1:273){
Contraste<-media[i,1] - media[i,2]
tcalc <- Contraste/EPCont

```

```

pval <- 2*pt(abs(tcalc),(nlinha),lower.tail=F)
Resultado[i,]<-c(nomes2[i],Contraste,pval)
}
Resultado<-as.data.frame(Resultado)

#####
#Ponto de corte FDR
#####
pval<-sort(Resultado$P.Valor)
m<-length(pval)
fdr.cutoff<-NA
for(i in 1:m){
q<-m*pval[i]/i
if(q<0.05){
fdr.cutoff<-pval[i]
}
}
fdr.cutoff

#####
#Gráfico Volcano
#####
plot(Resultado$LogFC,-log(Resultado$P.Valor,10),cex=0.4,pch=19,
xlab=expression(log[2]("fold change")),
ylab=expression(-log[10]("p valor")),main="")
abline(v=c(-1,1),col="blue")
abline(h=-log(fdr.cutoff,10),col="red")

```

ANEXO H – ROTINAS PARA A ANÁLISE COM O MÉTODO 6

```
#####
#Obtendo os valores de log(Sadia*) e log(Doente*)
#####
LogDoente <- MA3$A + MA3$M/2
LogSadia <- MA3$A - MA3$M/2
dados.norm2 <- cbind(LogSadia[,1],LogDoente[,1],
LogSadia[,2],LogDoente[,2])
colnames(dados.norm2) <- c("Sadia1","Doente1","Sadia2",
"Doente2")

#####
#Teste de Shapiro-Wilk
#####
normalidade.dados <- matrix(NA,ncol=2,nrow=273)
colnames(normalidade.dados) <- c("Gene","P.Valor")

j=1
for(i in seq(1,819,3)){
Y <- c(c(dados.norm2[i:(i+2),1]),c(dados.norm2[i:(i+2),2]),
c(dados.norm2[i:(i+2),3]),c(dados.norm2[i:(i+2),4]))
normalidade.dados[j,1] <- nomes3[j]
normalidade.dados[j,2] <-
as.data.frame(unlist(shapiro.test(Y)[2]))[1,1]
j=j+1
}

normalidade.dados <- as.data.frame(normalidade.dados)
NaoNormal <- normalidade.dados[normalidade.dados$P.Valor<0.05,]
NaoNormal

#####
#Teste de Wilcoxon
#####
Trat <- factor(c(rep("sadia",3),rep("doente",3),rep("sadia",3),
rep("doente",3)))
nomes2 <- nomes[seq(1,819,3)]
wilcoxon <- matrix(NA,ncol=3,nrow=273)
colnames(wilcoxon) <- c("Gene","P.Valor","LogFC")
wilcoxon <- as.data.frame(wilcoxon)

j=1
for(i in seq(1,819,3)){
```

```

Y <- c(c(dados.norm2[i:(i+2),1]),c(dados.norm2[i:(i+2),2]),
c(dados.norm2[i:(i+2),3]),c(dados.norm2[i:(i+2),4]))
wilcox <- wilcox.test(Y~Trat)
wilcoxon[j,2] <- wilcox$p.value
wilcoxon[j,1] <- nomes2[j]
wilcoxon[j,3] <- mean(Y[c(4:6,10:12)])-mean(Y[c(1:3,7:9)])
j=j+1
}

dim(wilcoxon[wilcoxon$pvalor<0.05,])
wilcoxon

#####
#Ponto de corte FDR
#####
pval<-sort(wilcoxon$P.Valor)
m<-length(pval)
fdr.cutoff<-NA
for(i in 1:m){
q<-m*pval[i]/i
if(q<0.05){
fdr.cutoff<-pval[i]
}
}
fdr.cutoff

#####
#Gráfico Volcano
#####
names(wilcoxon)
plot(wilcoxon$LogFC,-log(wilcoxon$P.Valor,10),cex=0.1,lwd=3,
xlab=expression(log[2]("fold change")),
ylab=expression(-log[10]("p valor")),main="")
abline(v=c(-1,1),col="blue")
abline(h=-log(0.01515152,10),col="red")

```