**Experimentos com *microarrays* -
modelos para a identificação de
genes diferencialmente expressos**

Rejane Augusta de Oliveira Figueiredo

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO GRAU DE MESTRE
EM
ESTATÍSTICA

Área de Concentração: **ESTATÍSTICA**
Orientadora: Profa. Dra. **Júlia Maria Pavan Soler**

— São Paulo, maio de 2006 —

**Experimentos com *microarrays* –
modelos para a identificação de
genes diferencialmente expressos**

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Rejane Augusta de Oliveira Figueiredo
e aprovada pela comissão julgadora.

São Paulo, maio de 2006

Banca examinadora:

- Prof. Dra. Júlia Maria Pavan Soler (Orientadora) (IME - USP)
- Prof. Dr. Eduardo Jordão Neves (IME - USP)
- Luzia Aparecida Trinca (Instituto de Biociências - UNESP)

Resumo

Grandes avanços estão ocorrendo na área de Genética e Genômica. Inicialmente cada gene era analisado separadamente com o intuito de verificar, por exemplo, associações com desenvolvimento de doenças. Recentemente surge a técnica de *microarrays* que permite que milhares de genes sejam avaliados simultaneamente. Esta técnica apresenta diversas vantagens nas aplicações em muitas áreas do conhecimento como, por exemplo, a área médica, porém uma série de dificuldades são encontradas nestes experimentos devido às diversas fontes de variações sistemáticas que podem interferir nas mensurações obtidas, acarretando em resultados falso-positivos. Devido a estas variações e a outros problemas encontrados nestes experimentos, como o problema de múltiplos testes, pois milhares de genes são avaliados num mesmo momento, muitos são os esforços em se encontrar uma melhor abordagem de análise estatística para a identificação de genes diferencialmente expressos (DE). Baseado nestes aspectos, neste estudo são apresentados detalhes dos experimentos com *microarrays*, suas limitações e, frente a isso, possíveis técnicas de análise estatísticas são consideradas como alternativas não paramétricas e paramétricas de normalização de dados e modelos estatísticos de efeitos fixos e modelos mistos na identificação de genes DE. Como aplicação e motivação deste trabalho, algumas técnicas de análise são aplicadas a um conjunto de dados reais com ratos diabéticos.

Abstract

Great advances are occurring in the Genetics and Genomics area. At the beginning each gene was analysed separately with the intention to verify, for example, associations with the development of diseases. Recently the *microarrays* technique appears, that allows that thousands of genes are evaluated simultaneously. This technique presents several advantages in applications in many areas, such as the medical area. However, a series of difficulties are found in these experiments due to several sources of systematic variations that can interfere in the results, and may come to false-positive results. Due to these variations and other problems found in these experiments, such as the multiple tests problem, because thousands of genes are evaluated at the same moment, many are the efforts at finding a better approach of statistical analysis for the identification of differentially expressed genes (DE). Based on these aspects, this study presents details of the experiments with *micoarrays*, its limitations and, face to this, the possible techniques of statistical analysis are considered as nonparametric and parametric alternatives of normalization of statistical models and data of fixed effects and mixed models in the identification of the genes DE. As application and motivation of this work some techniques of analysis are applied to a set of real data with diabetic mice

Agradecimentos

Às mulheres da minha vida: minha filha JÚLIA e minha mãe IZABEL.

À minha filha, simplesmente por ela existir e ser maravilhosa, o amor da minha vida. Esta obra não teria acontecido se não fosse ELA.

Minha mãe pelo incentivo, amor, carinho, minha formação pessoal, apoio ao estudo e sua sempre grande força em todos os momentos da minha vida.

Ao Bá, meu amor e companheiro, por estar sempre do meu lado. Pelas alegrias, tristezas, carinho e pelo apoio nos momentos difíceis

Aos meus irmãos Tronzinha, Her e Vi, e tios Dio (minha tiazinha) e Tito (meu tiozãozinho) que, com tanto amor, contribuíram para o meu crescimento pessoal e de forma direta ou indireta me ajudaram a estar aqui.

À minha especial amiga Val, que eu conheci no início do curso de mestrado e que, desde quando éramos apenas colegas de estudo, me ajudou muito a estar aqui. Virou uma amigona e continua me ajudando na vida, nos estudos, companheirismo e amizade.

A minha amiga Ângela, por todo apoio, força. Não tenho palavras para falar o quanto ela me ajudou, principalmente na reta final. Sem ela, com certeza, não conseguiria estar aqui.

Aos meus grandes amigos Vá, Bob e Rô pelo ombro amigo.

Aos amigos da pós graduação Paula, Gianne e o Alcimar, pela força e trocas.

À turma da Faculdade de Saúde Pública pelo conhecimento adquirido, por eu ter aprendido o que é trabalhar num ambiente agradável, pelo apoio e pelos Chopps: Victor, Rose, Natália, Fátima, Esther, Anderson, Luzia, Rossana.

Ao Júlio Pereira, pela força, ouvido e incentivo.

Sumário

Descrição	Página
1. Introdução.....	1
1.1. Os dados – Motivação.....	12
2. Experimentos com microarrays	21
2.1. Descrição da Técnica de Microarrays.....	21
2.1.1. Lâminas com <i>Oligonucleotídeo</i>	22
2.1.2. Lâminas com cDNA.....	24
2.2. Quantificação da imagem.....	26
2.3. Planejamento de Experimento com microarrays.....	27
2.3.1. Representação gráfica do planejamento do experimento.....	29
2.3.2. Estruturas do Delineamento de Experimentos com <i>Microarrays</i>	31
2.3.3. Delineamentos de Experimentos com um ou mais fatores.....	36
3. Normalizações das Medidas de Intensidade de Expressão Gênica.....	45
3.1. Métodos de normalização não paramétricos.....	50
3.2. Métodos de normalização paramétricos.....	55
4. Métodos Estatísticos para identificação de genes diferencialmente expressos.....	57
4.1. Comparações entre Duas condições de interesse.....	58
4.1.1. Método Empírico através do <i>Fold Change</i>	58
4.1.2. Métodos paramétricos - Estatística <i>t</i>	60
4.2. Análise para comparações com mais de dois tratamentos de interesse.....	66
4.4.1. Modelo com efeitos fixos.....	68
4.4.2. Modelo com efeitos fixos e aleatórios (modelos mistos).....	72

Descrição	Página
4.2.3. Estatística <i>F</i> Penalizada.....	84
4.3. Gráfico Vulcão – Gráfico exploratório para avaliação da significância dos testes.....	86
4.4. O problema dos múltiplos testes.....	88
4.4.1. Controle FWER (Family-Wise Error Rate).....	90
4.4.1. Controle FDR (False Discovery Rate).....	90
5. Aplicação.....	94
5.1. Modelo I – Teste t em dados com normalização não paramétrica....	96
5.2. Modelo II – Anova com modelos mistos (modelo de normalização e genético).....	102
5.3. Modelo III – Anova com efeitos mistos em um único modelo.....	106
5.4. Modelo IV – Anova com modelos de efeitos fixos.....	109
5.5. Comparação entre os modelos.....	112
6. Considerações Finais.....	120
Referências Bibliográficas.....	123
Apêndices.....	132
Apêndice 1 – Exemplo de banco de dados para os 40 primeiros genes (spots) avaliados em um dos 24 animais do estudo.....	133
Apêndice 2 – Códigos SAS para os programas utilizados no estudo.....	134

Lista de Figuras

Descrição	Página
Figura 1.1 – Molécula de DNA com suas bases complementares.....	5
Figura 1.2 – Desenho ilustrativo de transcrição.....	6
Figura 1.3 – Lâmina de <i>microarrays</i> contendo os 32 blocos, com discriminação de um dos blocos.....	13
Figura 1.4 – Histograma das medidas de intensidades de expressão, independente de gene e tratamento, considerando os dados dos ratos do estudo.....	15
Figura 1.5 – Histograma das medidas de intensidades de expressão transformadas, independente de gene e tratamento, considerando os dados dos ratos do estudo.....	16
Figura 2.1 – Desenho esquemático da técnica de <i>microarrays</i> de cDNA.....	25
Figura 2.2 – Ilustração da lâmina e imagem das cores obtidas em experimentos com <i>microarrays</i>	27
Figura 2.3 – Representação de delineamentos dos experimentos com <i>microarrays</i> de cDNA.....	30
Figura 2.4 – Representação gráfica de experimento com comparações diretas e indiretas.....	36
Figura 2.5 – Representação gráfica de delineamentos de experimentos com um fator.....	39
Figura 2.6 – Representação gráfica de um experimento tipo <i>looping</i> e de referência.....	41
Figura 3.1 – Exemplo de gráficos MA-Plot não normalizado e normalizado..	48

Descrição	Página
Figura 4.1 – Representação do Gráfico Vulcão.....	87
Figura 5.1 – Gráficos MA-PLOT para os diferentes tratamentos.....	97
Figura 5.2 – Gráficos MA-PLOT para os diferentes tratamentos com dados normalizados.....	98
Figura 5.3 – Gráfico Vulcão para resultados do Modelo I e <i>Fold Change</i>	101
Figura 5.4 – Gráfico Vulcão para resultados do Modelo II e <i>Fold Change</i>	105
Figura 5.5 – Gráfico Vulcão para resultados do Modelo III e <i>Fold Change</i>	108
Figura 5.6 – Gráfico Vulcão para resultados do Modelo IV e <i>Fold Change</i>	111

Lista de Tabelas

Descrição	Página
Tabela 1.1 - Medidas descritivas para as intensidades de expressão e valores de correção para quatro genes do estudo.....	17
Tabela 1.2 - Medidas descritivas para as intensidades de expressão transformadas em logaritmos e razão entre estas medidas para quatro genes de interesse.....	18
Tabela 1.3 – Estatísticas descritivas para a razão $\log_2 G/ R$ das intensidades de expressão comparadas entre diferentes tratamentos, para quatro genes de interesse.....	20
Tabela 2.1 – Distribuição em Blocos para planejamentos do tipo <i>looping</i> com três tratamentos e duas réplicas.....	42
Tabela 2.2 – Distribuição em Blocos para planejamentos com referência com três tratamentos e duas réplicas.....	42
Tabela 4.1 – Tabela resumo para o problema de múltiplos testes.....	91
Tabela 5.1 – Distribuição dos dados por <i>Flags</i>	95
Tabela 5.2 – Genes considerados DE para cada comparação entre os tratamentos – Modelo I.....	100
Tabela 5.3 – Estimativa da média da diferença de intensidade de expressão (na escala $\log_2$) entre os tratamentos para os genes considerados DE – Modelo I.....	101
Tabela 5.4 – Genes considerados DE para cada comparação entre os tratamentos – Modelo II.....	103
Tabela 5.5 – Estimativa da média da diferença de intensidade de expressão (na escala $\log_2$) entre os tratamentos para os genes considerados DE – Modelo II.....	104

Descrição	Página
Tabela 5.6 – Genes considerados DE para cada comparação entre os tratamentos – Modelo III.....	106
Tabela 5.7 – Estimativa da média da diferença de intensidade de expressão (na escala $\log_2$) entre os tratamentos para os genes considerados DE – Modelo III.....	107
Tabela 5.8 – Genes considerados DE para cada comparação entre os tratamentos – Modelo IV.....	110
Tabela 5.9 – Estimativa da média da diferença de intensidade de expressão (na escala $\log_2$) entre os tratamentos para os genes considerados DE – Modelo IV.....	110
Tabela 5.10– Genes DE para a comparação Tratamento 1 x Tratamento 2..	113
Tabela 5.11 – Genes DE para a comparação Tratamento 1 x Tratamento 3..	113
Tabela 5.12 – Genes DE para a comparação Tratamento 1 x Tratamento 4.	114
Tabela 5.13 – Genes DE para a comparação Tratamento 2 x Tratamento 3.	114
Tabela 5.14 – Genes DE para a comparação Tratamento 2 x Tratamento 4.	115
Tabela 5.15 – Genes DE para a comparação Tratamento 3 x Tratamento 4.	115
Tabela 5.16 – Estimativa da média da diferença de intensidade de expressão (na escala $\log_2$) entre os tratamentos para os genes considerados DE – Modelo II.....	118
Tabela 5.17 – Eficiência relativa entre os modelos.....	118

Capítulo 1

Introdução

O desenvolvimento embrionário e a formação de um ser humano são controlados por genes¹. Sendo assim, conhecer a estrutura genética possibilita trazer informações de características dos indivíduos, animais e vegetais, com forte impacto em áreas de aplicações das ciências biológicas, como a medicina e a agricultura (Filho & Verjovski-Almeida, 2003). Nos dias de hoje, diversas técnicas estão sendo utilizadas para entender a estrutura genética de uma grande variedade de doenças com o objetivo de produzir melhores diagnósticos, prevenção e cura destas doenças (Satagopan & Panageas, 2003).

Para se chegar à fase do conhecimento do genoma em que estamos, conhecer e compreender os genes foi uma etapa primordial. A idéia de analisar a seqüência de DNA² (Ácido Desoxirribonucléico do inglês *Deoxyribonucleic acid*), identificar genes e, com isso, as possíveis funções que eles poderiam exercer no corpo humano é relativamente antiga como, por exemplo, em pesquisas com os sobreviventes da bomba atômica. Neste período o interesse em analisar o DNA surgiu para verificar se havia ocorrido algumas alterações nos DNAs (mutações) destes sobreviventes. No entanto, apenas em meados da década de 1980 é mencionada, publicamente, a possibilidade de sequenciar todo o genoma³ (Bueno, 1997).

¹ Gene – seqüência de bases nitrogenadas da molécula de DNA que codificam proteína, sendo assim considerada a parte funcional do DNA (Farah, 1997).

² DNA - é a molécula orgânica que quando transcrita em RNA, tem a capacidade de codificar proteínas.

³ Genoma – termo utilizado para designar o conjunto total de genes em um indivíduo.

A concretização da identificação dos genes se inicia em 1990 quando surge nos Estados Unidos o Projeto Genoma Humano (PGH) numa proposta de mapeamento⁴ e sequenciamento⁵ do genoma humano. Esse projeto foi financiado pelas principais agências de pesquisa dos EUA, o *National Institute of Health* (NIH) e o Departamento de Energia (DOE), responsável pelo planejamento e desenvolvimento do projeto (ver, www.ghr.nlm.nih.gov e www.genome.gov/10001477). Desta iniciativa, o estudo se estendeu a genomas de outros organismos como plantas e animais, uma vez que a maioria dos organismos vivos apresenta muitos genes que são similares ou homólogos, ou seja, com funções semelhantes (ver, www.ghente.org/ciencia/genoma).

O PGH surgiu inicialmente nos Estados Unidos, mas hoje diversos países estão participando e contribuindo para este estudo, dentre eles o Brasil. A pesquisa genômica no país surgiu com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), responsável pela organização, criação e gerenciamento da Rede ONSA (Organização para Sequenciamento e Análise de Nucleotídeos), que se caracteriza por ser um instituto virtual de genômica formado, no início, por 30 laboratórios ligados à Fapesp (www.fapesp.br). No Brasil, há vários projetos de pesquisa ligados ao genoma financiados pela Fapesp e por outras agências públicas ou particulares, tais como Conselho Nacional de Desenvolvimento Científico (CNPq), Programa de Apoio ao Desenvolvimento Científico e Tecnológico (PADCT) (Bueno, 1997), Instituto Ludwig de Pesquisas, etc. Muitos destes projetos estão em andamento e muitos já foram concluídos.

As informações obtidas nestes projetos, iniciadas e derivadas do Projeto Genoma Humano, que se caracteriza por ser um esforço internacional para elucidar e caracterizar a seqüência completa dos pares de bases do genoma

⁴ Mapeamento – localização do gene no genoma.

⁵ Sequenciamento genético – técnica utilizada para identificar a seqüência de bases (ATCG) do DNA.

humano, vai revolucionar a prática da medicina no presente século por prover instrumentos para determinar o componente genético de todas as doenças (Simpson & Caballero, 2000).

O seqüenciamento do genoma humano já foi concluído (International Human Genome Sequencing Consortium, 2004) e estamos na fase do seqüenciamento do genoma de várias espécies e do genoma funcional, porém anteriormente a este período os mapas genéticos e os mapas físicos proporcionavam a melhor descrição da localização dos genes no organismo. O mapa genético caracteriza-se por ser o primeiro tipo de mapa molecular construído para a identificação da posição de locos⁶ genéticos ao longo dos cromossomos, com as distâncias entre os locos definidas em função da fração de recombinação de material genético entre eles (Huber et al., 2003). Sendo assim, o mapa genético reflete a localização relativa de genes nos cromossomos, apontando, por exemplo, que quanto maior a distância entre dois locos maior a taxa de recombinação entre eles. O mapa físico, por sua vez, diferencia-se do mapa genético, por se referir a uma posição física de um determinado gene dentro do genoma, onde a distância entre os genes é calculada em unidades físicas de comprimento, os números de pares de bases nitrogenadas (Farah, 1997).

Os mapas genéticos, mapas físicos e seqüenciamento tratam do genoma estrutural. A partir das informações obtidas no genoma estrutural, começa o interesse pela genômica funcional que tem como objetivo compreender como e quando os genes são expressos em um organismo, quais suas inter-relações com outros genes e com o ambiente, e como são regulados. Sendo assim, a genômica funcional, usando as informações geradas pela genômica estrutural, leva à completa caracterização do padrão de expressão do conjunto completo dos genes, assim como à investigação sistemática das propriedades funcionais desse conjunto de genes (Simpson & Caballero, 2000).

⁶ Loco – determinada região cromossômica no genoma

É no campo da genômica funcional que surgem as técnicas para avaliar e entender a expressão de alguns genes em determinadas situações. *Microarrays* é uma das técnicas utilizada para medir e investigar a expressão genética, que tem como grande vantagem a possibilidade de avaliar uma enorme quantidade de genes ao mesmo tempo (Draghici, 2003). A análise das informações encontradas, nesta e nas demais técnicas disponíveis, é realizada utilizando-se métodos computacionais e estatísticos intensivos e não triviais. Neste cenário, é crescente a necessidade de desenvolvimento e aperfeiçoamento de ferramentas computacionais específicas, bem como de suas aplicações em novas abordagens experimentais (Lee & Lee, 2000).

Com o avanço biotecnológico, abordagens que avaliavam em um experimento um único gene ficaram ultrapassadas e, assim, foram desenvolvidas novas estratégias capazes de fornecer uma perspectiva global e simultânea da expressão do genoma de um organismo ou tecido por meio de um único experimento (Santos et al., 2001). Dentre estas estratégias está a tecnologia de *Microarrays*, que foi desenvolvida para explorar os dados da sequência de DNA e produzir informações sobre o nível de expressão gênica do genoma (Dudoit et al., 2000). A seguir será apresentada uma breve revisão sobre alguns conceitos de Genética para uma melhor compreensão dos experimentos com *Microarrays*, foco deste estudo.

No núcleo das células está presente a molécula de DNA que contém toda a informação genética de um organismo. O DNA consiste de duas fitas, sendo cada uma o complemento da outra. As fitas são combinações de bases nitrogenadas de 4 tipos: A (Adenina), C (Citosina), G (Guanina), T (Timina) (ver Figura 1.1 para ilustração). A Uracila (U) é também uma base nitrogenada complementar à Adenina, presente na molécula de RNA (Ácido Ribonucléico do

inglês *Ribonucleic acid*) durante o processo de transcrição⁷ de seqüências de DNA (ver Figura 1.2). Desta forma, as bases complementares são: A-T (ou A-U) e C-G. Cada seqüência de 3 bases, irá determinar um aminoácido (unidade de uma proteína). O arranjo dos aminoácidos, por sua vez, irá determinar a função de uma proteína (Satagopan & Panageas, 2003).

Figura 1.1 – Molécula de DNA com suas bases complementares

Os genes são a parte funcional do DNA que codificam proteína, e que será passada de uma geração para a outra. As instruções para fazer proteínas, moléculas responsáveis por controlar a estrutura e função da célula, são conduzidas pelos genes no DNA. Para a confecção da proteína, a informação de uma das fitas do DNA é transcrita para um mRNA (RNA mensageiro) que irá conduzir as instruções genéticas para a confecção das proteínas (ver Figura 1.2). Esse mRNA será uma seqüência complementar ao pedaço de DNA que ele está transcrevendo (Farah, 1997).

A expressão gênica é medida pelo número aproximado de cópias de mRNA para um particular gene na célula. Um RNA em abundância representa um gene que tem alta expressão, ou seja, que está sendo transcrito em abundância. Desta forma, a quantidade de mRNA está relacionada com a

⁷ Transcrição – processo pelo qual a dupla hélice de DNA se desenrola e uma cópia de RNA é sintetizada a partir de uma das fitas complementares de um gene.

quantidade de proteína que o gene produz. Um dos principais objetivos em análises de expressão gênica é revelar padrões presentes na expressão dos genes de, por exemplo, diferentes tecidos de um mesmo organismo como, por exemplo, o tecido adiposo e tecido muscular.

Figura 1.2 – Desenho ilustrativo de transcrição

Sabe-se que os mesmos genes estão presentes nos DNA de todas as células de um mesmo organismo, porém nem todos os genes são expressos da mesma forma em todos os locais, tecidos e fases celulares, ou seja, a

transcrição de mRNA ocorre de diferentes formas nos diferentes tecidos do corpo ao longo da vida de um indivíduo (Farah, 1997).

A técnica de *microarrays* é utilizada para medir a expressão de milhares de genes, através da imobilização (fixação) de fragmentos previamente conhecidos de DNA em uma lâmina. Utilizando-se o princípio da complementaridade descrito anteriormente, o mRNA dos diferentes tratamentos⁸ de interesse irá, por exemplo, se juntar à sequência fixada na lâmina. Há diversos tipos de técnicas, porém as mais utilizadas atualmente são: a técnica de *microarrays* baseada em oligonucleotídeos⁹, onde curtos pedaços da sequência de DNA são pré-fixadas nas lâminas, e a técnica baseada no DNA complementar (cDNA), onde longas sequências de DNA são fixadas na lâmina. Além desta diferença, há outras diferenças entre tais técnicas, a citar: nos experimentos com oligonucleotídeos para cada tratamento de interesse no estudo é utilizada uma diferente lâmina, já nos experimentos com cDNA, em cada lâmina são avaliados dois diferentes tratamentos, simultaneamente (Woo et al., 2004). É difícil julgar qual das tecnologias é superior, porém algumas vantagens e desvantagens podem ser citadas como: nos experimentos com cDNA pode haver uma sequência de DNA fixada na lâmina que seja desconhecida, o que, em geral, não ocorre em experimentos com oligonucleotídeos; pode haver uma maior fonte de variação sistemática nos estudos com cDNA como, por exemplo, variação devido às diferentes colorações usadas para definir os tratamentos dentro de cada lâmina, isto não ocorre nos experimentos com oligonucleotídeos pois em cada lâmina há apenas um tratamento com apenas uma coloração de identificação; uma vez que tenha

⁸ Tratamentos – em geral, textos relacionados à experimentos com *microarrays* utilizam o termo amostras de mRNA, para se referenciar a, por exemplo, amostras (pedaços) de diferentes tecidos. No presente estudo utilizaremos o termo “tratamento” para se referir a estas amostras de mRNA para não haver confusão com o termo amostra utilizado em estudos estatísticos.

⁹ Oligonucleotídeo – fragmento sintético de DNA composto por somente poucas sequências (em torno de 25 pares de bases nitrogenadas).

sido feito um bom planejamento do experimento e que haja interesse na comparação de apenas dois tratamentos a análise com cDNA é mais simples; o custo das lâminas de oligonucleotídeos são também mais caras (ver Draghici, 2003; Coe, 2003).

Nos estudos com oligonucleotídeos há diversos modelos de lâminas utilizadas para os experimentos, tais como o modelo desenvolvido pela empresa Affimetrix (ver, www.affimetrix.com) e o modelo da Motorola CodeLink (Ramakrishnan et al., 2002). Nestes experimentos, para cada gene a ser pesquisado são utilizados em torno de 11 a 20 oligonucleotídeos (Speed, 2003), cada um contendo cerca de 25 pares de bases nitrogenadas. Os oligonucleotídeos são fixados na lâmina aos pares, um dos pares correspondendo à verdadeira seqüência do gene em estudo, denominada PM (do inglês *perfect-match*), e outro par correspondendo a uma seqüência controle, denominada MM (do inglês *mismatch*). Um PM é o complementar de uma subsequência do gene de interesse e MM é exatamente o mesmo complementar com apenas uma das bases na posição central alterada. Em geral, é calculada a diferença entre as respostas de expressão gênica da seqüência PM e da seqüência MM (PM-MM) para cada um dos oligonucleotídeos. Considerando todos os pares para avaliar um específico gene (em torno de 11 a 20), a média das diferenças para todos os pares é utilizada como o índice de expressão do gene em questão (Satagopan & Panageas, 2003).

Em estudos com cDNA, cada gene é representado por uma grande e única seqüência de bases. Além disso, como citado anteriormente, para identificação de genes diferencialmente expressos, em cada lâmina, amostras de mRNA de dois tratamentos são utilizadas para a análise. As duas amostras são marcadas com fluorescências de diferentes cores, em geral, verde (cianina 3 - denotado por Cy3) e vermelha (cianina 5 - denotado por Cy5), e depois

hibridizadas¹⁰ na lâmina, onde as seqüências de cDNA estão fixadas. Nas análises descritivas ou em análises com técnicas não paramétricas, a medida de expressão gênica diferencial para dois tratamentos pode ser avaliada em uma mesma lâmina através da diferença (na escala logarítmica) entre as intensidades de expressão dos dois tratamentos. O presente estudo será baseado nesta técnica de cDNA, com duas amostras (tratamentos) em cada lâmina. Maiores detalhes serão apresentados nos capítulos seguintes.

Para as duas técnicas, tanto com oligonucleotídeos quanto com cDNA, uma seqüência de uma única fita do DNA complementar de um gene de interesse é fixada nas cavidades (do inglês *spots*¹¹) presentes nas lâminas de materiais como vidro, nylon ou folha de quartzo (Huber et al., 2003). As amostras de mRNA para cada tratamento de interesse são marcadas com uma determinada fluorescência. Em experimentos com oligonucleotídeos, por conterem estes apenas uma lâmina para cada tratamento avaliado, haverá uma única cor de fluorescência e, em experimentos com cDNA, duas cores de fluorescência, uma para cada tratamento em uma mesma lâmina. Essas amostras de mRNA de interesse, previamente marcadas com as fluorescências, são hibridizadas nas seqüências pré-fixadas das lâminas. Após a hibridização, será medida a intensidade de fluorescência presente em cada *spot* com o auxílio de um aparelho de *scanner*. Esta intensidade corresponderá à quantidade de mRNA em cada *spot*, medindo-se assim a expressão gênica daquela seqüência. Desta forma, podemos dizer que um gene é mais expresso em um determinado tratamento quando há predominância da sua coloração em um *spot* específico.

¹⁰ Hibridizar – é a ligação de seqüências de DNA e/ou RNA através do pareamento das bases nitrogenadas complementares para formar uma cadeia de dupla fita.

¹¹ Spots – por haver uma grande dificuldade na tradução de alguns termos utilizados em experimentos com microarrays e com o intuito de manter este material coerente com outras publicações existentes sobre o tema, para representar as cavidades existentes nas lâminas, ao longo deste estudo será utilizado esse termo *spot*, não sendo traduzido para o português.

Embora estes dois tipos de lâminas (*arrays*), com oligonucleotídeos e com cDNA, produzam expressão gênica usando diferentes métodos, ambas podem ser dirigidas para uma pesquisa com a mesma finalidade em questão e técnicas estatísticas similares de análise dos dados podem ser aplicadas (Satagopan & Panageas, 2003).

As medidas de expressão gênica obtidas por meio das técnicas de *microarrays* podem apresentar diversas fontes de variação, como problemas na preparação das amostras de mRNA de cada tratamento antes de serem hibridizadas, erros devido às propriedades de hibridização da lâmina, falhas na detecção da fluorescência e até mesmo sujeira na lâmina (Holder et al., 2001). Como exemplo, em estudos com mais de uma lâmina, pode haver variação na qualidade de uma lâmina para outra e, assim, devido a condições inconsistentes do experimento, pode-se aumentar ou reduzir a eficiência de hibridização de um cDNA, fazendo com que as diferenças de intensidade da expressão gênica ocorram devido às diferentes lâminas e não aos diferentes tratamentos. Além disso, uma das colorações pode estar mais brilhante do que a outra por propriedades físicas ou químicas. Desta forma, um gene pode emitir um sinal de fluorescência mais alto (ou mais baixo) do que outros genes por apresentar um maior nível de expressão, ou devido a estas variações na eficiência da hibridização e/ou na eficiência da coloração das amostras de mRNA. Outro grande problema encontrado neste tipo de estudo é que uma enorme quantidade de genes são avaliados, porém com uma baixa quantidade de unidades experimentais (Draghici, 2003), isso devido, principalmente, ao alto custo das lâminas. Como exemplo, os dados utilizados neste estudo contêm informação de apenas 24 animais e para cada animal foram avaliados 12.768 genes. Vários são os esforços realizados em experimentos com *microarrays*, visando minimizar as variações encontradas nestes experimentos devido a fatores externos e indesejáveis. Para se garantir a qualidade das medidas de expressão, há os procedimentos de controle *a priori*, realizados por meio de um cuidadoso planejamento do experimento; e os procedimentos de controle *a*

posteriori, como as técnicas de normalizações de dados, que são utilizadas após medida a intensidade da expressão gênica (Schuchhardt et al., 2000). Algumas formas de ajuste para estes possíveis problemas, tanto *a priori* como *a posteriori*, serão tratados neste trabalho.

O presente estudo tem como objetivo apresentar alguns métodos de análise estatística de dados de *microarrays* de cDNA para identificar genes diferencialmente expressos. São consideradas possíveis formas de normalizar dados, frente a diversas possibilidades de variações encontradas neste tipo de experimento. Além dos aspectos acima, serão abordados possíveis delineamentos de experimentos para estudos com *microarrays* de cDNA, alternativas de análise paramétricas e não paramétricas, e, também, serão citadas algumas estratégias para problemas de múltiplos testes, por haver comparações para um número muito grande de genes.

Visando introduzir os pontos citados acima, os capítulos deste trabalho estão distribuídos da seguinte forma:

- No Capítulo 2 é apresentada uma descrição dos experimentos com *microarrays*, suas principais características e estruturas de planejamento envolvidas. Maior ênfase é dada para experimentos com cDNA, por ser o foco deste estudo;
- No Capítulo 3 são abordados os principais problemas da técnica de *microarrays*, as possíveis fontes de variação sistemáticas presentes nestes experimentos e algumas formas de contornar, ou minimizar estas possíveis variações por meio das chamadas normalizações de dados;
- No Capítulo 4 são apresentadas as técnicas de análise estatística para identificação de genes diferencialmente expressos como testes *t*, análise de variância clássica, modelos mistos e as alternativas para o problema de múltiplos testes;

- O Capítulo 5 traz a aplicação das metodologias de análise estatística consideradas neste trabalho a um conjunto de dados reais obtidos em um experimento com *Microarrays*;
- Finalmente apresentamos as considerações finais e projetos futuros.

A seguir é apresentada a descrição dos dados que serão avaliados no presente estudo, facilitando e motivando a compreensão do problema.

1.1. Os Dados – Motivação

Os dados utilizados neste trabalho são referentes a um experimento realizado na Southwest Foundation for Biomedical Research (San Antonio, TX, USA). O estudo é de interesse e tem aplicação na área de Fisiologia do Exercício, visando identificar genes diferencialmente expressos comparados entre quatro diferentes tratamentos realizados com ratos. Considerando os quatro diferentes tratamentos, foram avaliados 24 animais subdivididos da seguinte forma:

- Tratamento 1 – animais diabéticos, não exercitados (n=6 animais);
- Tratamento 2 – animais diabéticos, treinados por 3 semanas, respondendo bem ao tratamento e que não eram mais diabéticos quando foram sacrificados (n=6 animais);
- Tratamento 3 – animais diabéticos, treinados por 3 semanas e que não responderam bem ao tratamento e permaneceram diabéticos (n=8 animais);
- Tratamento 4 – animais não diabéticos e não exercitados (n=4 animais).

Para o estudo foram utilizadas 24 lâminas de *microarrays* de cDNA, uma para cada animal e, em cada uma delas, cada tratamento foi comparado com um tratamento de referência (TR). As amostras experimentais de cada

tratamento receberam fluorescência verde (Cy3) e as amostras de referência de cada tratamento receberam a fluorescência vermelha (Cy5).

No experimento, para cada animal foram avaliadas 12.768 seqüências de bases nitrogenadas, representando os genes de interesse. Estas seqüências estavam dispostas nos *spots* das lâminas em 32 blocos contendo 19 linhas e 21 colunas cada um deles, como apresentado na Figura 1.3.

Figura 1.3 – Lâmina de *microarrays* contendo 32 blocos, com discriminação de um dos blocos.

O banco de dados, que parte dele está apresentado no Apêndice 1, contém as informações da medida de intensidade de expressão gênica para os tratamentos de interesse, rotulados com Cy3 (verde), e para as amostras de referência, preparadas com Cy5 (vermelha). A medição foi feita utilizando o GenePix Pro 3.0 (www.axon.com/gn_Genomics.html). As fontes de

fluorescência, Cy3 e Cy5, são relativamente instáveis. Por conta disso, elas podem influenciar a eficiência durante a rotulação da amostra e, assim, as intensidades detectadas por *scanners* podem ser medidas com diferentes eficiências (Speed, 2003). Baseado nestas informações, o *software* apresenta, além da medida de intensidade avaliada em cada *spot* (para cada gene), uma correção da intensidade medida também para cada *spot* (correção pelo *background*). Desta forma, para cada gene há informações da intensidade mediana¹² para o tratamento de interesse (Cy3) e para a medida de referência (Cy5), e os respectivos valores de correção das intensidades verdes e vermelhas. Isto posto, para avaliar a intensidade da coloração verde (Cy3) ou da coloração vermelha (Cy5), que neste estudo serão denotadas por G (do inglês *Green*) e R (do inglês *Red*), respectivamente, deverá ser calculado o valor da intensidade de uma determinada cor menos o seu fator de correção em cada *spot*. Em muitas análises e estudos exploratórios, para se verificar o comportamento dos genes, a identificação de genes diferencialmente expressos se faz através da utilização da relação entre as medidas de intensidade avaliada para cada tratamento, já com as devidas correções pelo *background*¹³.

Nas análises de experimentos com *microarrays* há a possibilidade de encontrar valores muito altos e ordens de grandezas diferentes entre as intensidades. Além disso, ao observarmos as medidas de intensidade avaliadas para a fluorescência verde e vermelha para o experimento como um todo,

¹² Intensidade mediana – em cada *spot* a medição da imagem se faz por meio de diversos *pixels* que são sinais luminosos de um corante, ou seja, pontos luminosos emitidos pelas moléculas (Dantas, 2004). A intensidade de cada *spot* é dada por medidas resumo obtidas destes *pixels*. De acordo com o *software* utilizado neste estudo a medida utilizada é a mediana dos valores.

¹³ *Background* – região no *spot* onde é medida a intensidade das cores, porém não há uma sequência de bases nitrogenadas, ou seja, não há uma hibridização (Dantas, 2004; Wu & Yan, 2003). Desta forma, é uma mancha ao redor do local onde está alocada a sequência de interesse dentro de um determinado *spot*. A medida de intensidade desta região não é de interesse ao se avaliar a medida de intensidade de expressão de um determinado gene, devendo desta forma ser corrigida.

independente de gene e de tratamento (ver Figura 1.4 a e b), nota-se que os dados se distribuem de forma assimétrica e, no ponto mais alto da curva, muito afunilada (curtose positiva). Para atenuar estas discrepâncias e obter uma distribuição aproximadamente Normal dos dados é utilizada a transformação logarítmica na base 2 para as intensidades medidas nos estudos com *microarrays* (Speed, 2000). Na Figura 1.5, considerando os dados dos ratos do estudo, está ilustrado o gráfico com as intensidades de fluorescências verdes e vermelhas transformadas .

Figura 1.4 – Histograma das medidas de intensidade de expressão, independente de gene e tratamento, considerando os dados dos ratos do estudo.

Figura 1.5 – Histograma das medidas de intensidade de expressão transformadas, independente de gene e tratamento, considerando os dados dos ratos do estudo

A Tabela 1.1 traz as estatísticas descritivas dos valores obtidos pelo *software* de leitura da imagem de fluorescência, que são as medidas de intensidade verde e vermelha, e suas respectivas correções pelo *background*, sendo:

G: medida de expressão gênica para os grupos experimentais;

R: medida de expressão gênica para os grupos de referência.

Estes resultados apenas mostram exemplos de como são os valores iniciais medidos em um experimento com *microarrays*. As medições iniciais das intensidades R e G devem ser ajustadas pelas correções do *background* e somente depois deve-se proceder com a transformação logarítmica na base 2, citadas anteriormente. Para ilustração da variação das medidas entre unidades experimentais, a seguir estão apresentadas as estatísticas descritivas apenas para quatro genes do estudo com diferentes comportamentos.

Tabela 1.1 - Medidas descritivas para as intensidades de expressão e valores de correção para quatro genes do estudo.

		Gene			
		25	6320	6808	10690
Intensidade - G	Média	1939,9	2492,7	2883	1950,9
	Mediana	1590	1334	2421	1554,5
	Desvio Padrão	1455,7	2798,7	1773,5	1250,2
	Mínimo	626	561	909	444
	Máximo	7615	10649	8725	5450
	CV ¹⁴	0,75	1,12	0,62	0,64
Correção Background - G	Média	221,1	335,4	163,4	134,1
	Mediana	151	307	96	94
	Desvio Padrão	225,6	281,4	197,3	171,4
	Mínimo	25	15	34	14
	Máximo	868	904	910	769
	CV	1,02	0,84	1,21	1,28
Intensidade - R	Média	2071,9	990	2132,7	1938,6
	Mediana	1611	731	1628,5	1586
	Desvio Padrão	1435,9	1002,1	1608	1076,7
	Mínimo	584	90	504	509
	Máximo	6468	4081	6721	4170
	CV	0,69	1,01	0,75	0,56
Correção Background - R	Média	192,6	235,5	187,9	160,9
	Mediana	104	154	91,5	63,5
	Desvio Padrão	196,8	249,6	225,3	216
	Mínimo	0	0	0	0
	Máximo	642	740	840	759
	CV	1,02	1,06	1,20	1,34

Na Tabela 1.2 estão dispostos os valores já corrigidos e transformados (na base logarítmica) dos quatro genes anteriormente citados e, também, a medida da relação entre essas intensidades $\log_2(G/R)$, independente do tratamento. A análise descritiva, bem como algumas comparações em experimentos com *microarrays* se faz através desta medida de relação entre as intensidades das duas colorações, ou seja, a diferença entre as intensidades de dois tratamentos na escala logarítmica. No presente estudo o interesse reside em comparar o comportamento dos genes que se expressam diferencialmente, quando comparados entre os quatro tratamentos. Os dados das Tabelas 1.1 e 1.2 não estão segmentados entre os quatro tratamentos e podem ser utilizados

¹⁴ CV – coeficiente de variação

para descrever a distribuição geral das respostas de intensidade. Desta forma, estes dados servem para uma melhor compreensão do padrão de variação das intensidades de expressão dos genes, bem como exemplificar que diferentes genes podem apresentar comportamentos diferenciados entre si. Pelos dados da Tabela 1.2 pode-se observar que o gene 6320 apresenta maior valor da razão $\log_2(G/R)$, sugerindo que, entre os quatro genes considerados, em média, é o que mais se expressa nas amostras tratadas com G relativamente a R, contudo para o conjunto das 24 amostras sua variabilidade é maior (desvio-padrão de 2.12 unidades de medida).

Tabela 1.2 - Medidas descritivas para as intensidades de expressão transformadas em logaritmos e razão entre estas medidas para quatro genes de interesse

	Estatísticas (n=24)	Gene			
		25	6320	6808	10690
Log ₂ G	Média	10,5	10,4	11,2	10,5
	Mediana	10,4	10,0	11,2	10,5
	Desvio Padrão	0,8	1,3	0,9	1,1
	Mínimo	9,1	8,9	9,4	8,4
	Máximo	12,7	13,3	13,1	12,4
Log ₂ R	Média	10,6	9,1	10,4	10,6
	Mediana	10,5	9,3	10,4	10,4
	Desvio Padrão	0,9	1,4	1,4	0,8
	Mínimo	8,8	5,9	7,4	9,0
	Máximo	12,5	11,7	12,7	12,0
Log ₂ (G/R)	Média	-0,10	1,25	0,78	-0,08
	Mediana	-0,23	0,63	0,40	-0,02
	Desvio Padrão	0,49	2,12	1,71	0,97
	Mínimo	-0,95	-0,96	-2,73	-2,82
	Máximo	1,07	5,95	5,66	1,56

Como vimos anteriormente, em geral, a análise de *microarrays* de cDNA é feita comparando-se dois tratamentos em uma mesma lâmina, porém em muitas situações, quando mais de dois tratamentos estão sendo estudados, há a necessidade de se comparar resultados entre diferentes lâminas. Este é o caso do presente estudo, onde há o interesse em se comparar genes diferencialmente expressos entre quatro tratamentos. Diversas possibilidades de análise e planejamento do experimento são possíveis nestas situações. Como

neste estudo, em uma mesma lâmina, todos os tratamentos foram comparados com um mesmo tratamento de referência, as comparações entre os tratamentos podem ser feitas de forma indireta ou relativa. Para esta situação, a interpretação empírica e descritiva das comparações se dá da seguinte forma:

- $\text{Log}_2 (G/R) > 0 \Rightarrow$ maior expressão no tratamento experimental;
- $\text{Log}_2 (G/R) < 0 \Rightarrow$ maior expressão no tratamento de referência (TR);
- $\text{Log}_2 (G/R) = 0 \Rightarrow$ mesma expressão.

Como os mesmos genes foram avaliados para cada um dos 24 animais do estudo, pode ser feita uma comparação descritiva de cada gene entre os quatro tratamentos. Na Tabela 1.3 estão apresentadas as estatísticas descritivas das razões para os diferentes tratamentos, separadamente, para cada um dos 4 genes avaliados anteriormente.

Observando os resultados da Tabela 1.3, podemos ver que, para o gene 25, não há fortes evidências de diferenças entre os quatro tratamentos. O gene 6320 se destaca para o tratamento 2, em que a média de sua expressão neste tratamento é 3,72 vezes maior que sua expressão no tratamento de referência, sugerindo ser diferencialmente expresso para animais diabéticos e exercitados que responderam bem ao tratamento. Para este mesmo gene, pode-se notar que ele apresenta alta variabilidade no tratamento 2 (desvio-padrão de 3,06 unidades de medida), maior do que nos demais tratamentos. O gene 6808 sugere ser bastante expresso em dois tratamentos (tratamento 1 e 4), porém apresenta maior intensidade média e variabilidade no tratamento 4. O gene 10690, também apresenta comportamento diferenciado dos demais, sugerindo destaque apenas para o tratamento 1, onde ele se expressa mais no tratamento de referência (2,38 vezes mais) do que nos demais tratamentos.

A análise apresentada na Tabela 1.3 é apenas descritiva, visando ilustrar as escalas de medidas das respostas de intensidade da expressão gênica. Os resultados foram considerados apenas para quatro genes, visando dar uma idéia

geral e simplificada da análise e o grau de dificuldade que se tem ao estender a análise para uma grande quantidade de genes, como é o caso dos estudos em *microarrays* (neste caso, tem-se 12.768 avaliados). Deve-se destacar a alta variação presente nos dados e na estatística $\log_2 G/R$, o que pode ser observado pelos valores dos coeficientes de variação. Tal comportamento pode ser explicado pelo efeito de muitas fontes de variação que são esperadas atuarem nestes experimentos. A necessidade de normalização é imprescindível, como uma alternativa de obter inferências mais precisas. Ao longo deste estudo serão apresentados detalhes da técnica de *microarrays* e alternativas de análise estatística para identificação de genes diferencialmente expressos nestes experimentos.

Tabela 1.3 – Estatísticas descritivas para a razão $\log_2 G/R$ das intensidades de expressão comparadas entre diferentes tratamentos, para quatro genes de interesse

Log ₂ (G/ R)	Estatísticas (n=24)	Tratamento			
		1	2	3	4
Gene 25	Média	-0,13	0,19	-0,47	0,27
	Mediana	0,01	0,12	-0,46	0,17
	Desvio Padrão	0,29	0,51	0,31	0,53
	Mínimo	-0,53	-0,32	-0,95	-0,26
	Máximo	0,2	1,07	0,02	1
	CV	-2,23	2,68	0,66	1,96
Gene 6320	Média	0,46	3,72	1,18	0,51
	Mediana	0,46	3,72	0,3	0,49
	Desvio Padrão	0,74	3,06	2,5	0,73
	Mínimo	-0,06	1,56	-0,96	-0,35
	Máximo	0,98	5,89	5,95	1,4
	CV	1,61	0,82	2,12	1,43
Gene 6808	Média	1,31	0,87	-0,58	2,81
	Mediana	0,97	0,64	-0,22	2,19
	Desvio Padrão	1,31	1,01	1,01	1,96
	Mínimo	0,25	-0,24	-2,73	1,21
	Máximo	3,06	2,74	0,21	5,66
	CV	1,00	1,16	-1,74	0,70
Gene 10690	Média	-2,38	0,06	0,14	0,39
	Mediana	-2,38	-0,07	0,07	0,1
	Desvio Padrão	0,62	0,53	0,53	0,82
	Mínimo	-2,82	-0,44	-0,62	-0,21
	Máximo	-1,94	0,72	0,83	1,56
	CV	0,26	8,83	3,79	2,10

Capítulo 2

Experimentos com *Microarrays*

Neste capítulo são apresentados detalhes da técnica de *microarrays* para obtenção das medidas de expressão gênica. Além disso, aspectos importantes do planejamento de experimentos com *microarrays* são também considerados.

2.1. Descrição da Técnica de *Microarrays*

Como visto no Capítulo 1, *Microarrays* é uma tecnologia de análise laboratorial para avaliar a expressão gênica sob diferentes condições experimentais, e que tem como grande vantagem a possibilidade de avaliar uma enorme quantidade de genes ao mesmo tempo em uma mesma lâmina. Estes genes avaliados estão dispostos nas lâminas de *microarrays* nos diversos *spots*.

Um dos principais objetivos na análise de dados de expressão gênica com *microarrays* é a identificação de genes diferencialmente expressos na comparação de diversos tratamentos ou condições experimentais e, por se tratar de um experimento com uma grande quantidade de dados, a análise exige intensa aplicação computacional e de técnicas estatísticas.

A seguir é apresentada a descrição de como é medido o nível de expressão gênica das duas principais técnicas de análise com lâminas de *microarrays*: a técnica utilizando oligonucleotídeos e a técnica utilizando cDNA.

2.1.1. Lâminas com *Oligonucleotídeo*

Nesta abordagem, para cada gene a ser pesquisado são utilizados em torno de 11 a 20 oligonucleotídeos (Speed, 2003), cada um contendo cerca de 25 pares de bases nitrogenadas. Como citado no capítulo anterior, os oligonucleotídeos são fixados aos pares, um correspondendo à sequência do gene de interesse (PM) e outro de sequência controle (MM), sendo que o MM serve de controle para a especificidade de hibridização (Santos e outros, 2001). A média das diferenças entre as intensidades de expressão obtida pelas sequências PM e MM de cada oligonucleotídeo é uma das alternativas utilizada como índice de expressão de um gene corrigido para possíveis variações indesejáveis dentro da lâmina. Além da média das diferenças entre PM e MM, há outras formas de se obter a medida de expressão de um gene dentro da lâmina em experimentos com oligonucleotídeos, como apresentado em Irizarry et al. (2003), que propôs diferentes ponderações para o cálculo da medida resumo das intensidades do gene. Neste trabalho, para melhor compreensão e exemplificação deste tipo de experimento, será utilizada a diferença (PM - MM).

Em linhas gerais, a intensidade de expressão gênica para este tipo de experimento é avaliada da seguinte forma (Holder et al., 2001):

- Suponha um número I ($I \geq 1$) de condições experimentais (tratamentos). Os tratamentos serão indexados por i , $i=1, \dots, I$;
- Uma ou mais lâminas poderão ser utilizadas para cada tratamento. As réplicas de lâminas serão indexadas por j , sendo $j = 1, \dots, J_i$ onde J_i é o número de lâminas para o i -ésimo tratamento;
- Em cada lâmina há G genes fixados, ou seja, G conjuntos de pequenos oligonucleotídeos ($g = 1, \dots, G$);
- O número de oligonucleotídeos para o g -ésimo gene é dado por L_g e são indexados por l , tal que $l = 1, \dots, L_g$, lembrando que L_g estará em torno de 11 a 20 ;

Sendo assim, o nível de expressão corrigido é avaliado por:

$$y_{ijg} = PM_{ijg} - MM_{ijg}$$

onde:

y_{ijg} : é a medida de intensidade de expressão do l -ésimo oligonucleotídeo do g -ésimo gene, para i -ésimo tratamento hibridizado na j -ésima lâmina;

PM_{ijg} : é a medida de intensidade de expressão da l -ésima sequência do oligonucleotídeo de interesse do g -ésimo gene, para o i -ésimo tratamento hibridizado na j -ésima lâmina;

MM_{ijg} : é a medida de intensidade de expressão da l -ésima sequência controle, do g -ésimo gene, para o i -ésimo tratamento hibridizado na j -ésima lâmina.

Desta forma, para o g -ésimo gene na j -ésima lâmina a medida de intensidade de expressão no i -ésimo tratamento se dá pela média das diferenças:

$$\bar{Y}_{ijg} = \frac{\sum_{l=1}^{L_g} Y_{ijgl}}{L_g}$$

onde,

$\bar{Y}_{ijg}$: é a media aritmética da intensidade de expressão do g -ésimo gene, para i -ésimo tratamento hibridizado na j -ésima lâmina.

Note que a medida de intensidade de cada oligonucleotídeo representante de um gene na lâmina é corrigida para efeitos residuais dentro da lâmina que possam conduzir a um falso nível de expressão (medido por MM). A literatura é polêmica sobre o ganho em precisão e o controle de fontes de variação obtidas por meio dos experimentos com oligonucleotídeo nos quais um fragmento de DNA sob estudo é particionado em pequenos fragmentos, cada um com seu correspondente controle dentro da lâmina.

2.1.2. Lâminas com cDNA

A técnica de *microarrays* de cDNA se baseia, inicialmente, em selecionar quantidades de mRNA nas condições experimentais de interesse, prepará-las e, assim, avaliar o nível de expressão genético de dois diferentes tratamentos em uma mesma lâmina. Podemos exemplificar a técnica com a comparação de dois tratamentos, sendo que em um deles as amostras de mRNA são extraídas de tecidos de pacientes com alguma doença (tumor) e o outro de tecidos de pacientes sem a doença. Cada uma destas amostras de mRNA são preparadas com fluorescentes de cores diferentes, em geral, uma com verde (G) e a outra com vermelho (R). As amostras com os corantes são misturadas e derramadas nas lâminas já preparadas com as seqüências de DNA previamente conhecidas para que, com estas seqüências, ocorra a hibridização (ver Figura 2.1). Através da leitura de um *scanner* a laser são verificadas as intensidades das cores em cada *spot*. Como cada *spot* representa um gene, para cada um deles verifica-se qual intensidade de cor é mais evidente, ou seja, para qual tratamento o gene está mais expresso.

Assim, a expressão genética em cada tratamento é avaliada a partir da intensidade de fluorescência (cores) medida nos *spots* que representam um gene de interesse. Lembrando que, como introduzido no capítulo anterior, a intensidade de expressão gênica é avaliada através da transformação logarítmica da medida de intensidade inicialmente obtida. Como em muitos estudos de *microarrays* a análise se faz pela comparação das medidas de intensidade das duas colorações em cada *spot* após a normalização dos dados, temos que:

$$M_{jg} = \log_2 (G_{jg}/R_{jg}) = \log_2 G_{jg} - \log_2 R_{jg}$$

onde,

M_{jg} : $\log_2$ da medida de intensidade da razão G/R do g -ésimo gene na j -ésima lâmina;

R_{ij} : $\log_2$ da medida de intensidade do tratamento marcado com a coloração vermelha do g -ésimo gene na j -ésima lâmina;

G_{ij} : $\log_2$ da medida de intensidade do tratamento marcado com a coloração verde do g -ésimo gene na j -ésima lâmina.

Sendo que:

- Em cada lâmina há G genes a serem analisados ($g = 1, \dots, G$);
- Haverá interesse em examinar I diferentes tratamentos, onde $i = 1, \dots, I$, sendo que em cada lâmina são considerados dois tratamentos.

Comparando dois tratamentos em uma mesma lâmina, pode-se interpretar empírica e descritivamente os valores de M_{ij} da seguinte forma:

- $M_{ij} > 0 \Rightarrow$ maior expressão no tratamento experimental corado com G;
- $M_{ij} < 0 \Rightarrow$ maior expressão no tratamento de referência (TR) corado com R;
- $M_{ij} = 0 \Rightarrow$ mesma expressão.

Hibridização com duas cores em experimento cDNA

Figura 2.1 – Desenho esquemático da técnica de *microarrays* de cDNA (Hirata, 2005)

2.2. Quantificação da Imagem

Após a hibridização, é feita a leitura da imagem das cores apresentadas na lâmina, por meio de um *scanner* específico com alta resolução espacial. Após realizada a leitura da imagem o grande esforço é quantificar a intensidade desta imagem. O processamento da imagem é realizado por meios computacionais. Neste estudo o software utilizado foi o Genepix, porém há diversos softwares disponíveis para a quantificação e análise das imagens (ver Speed, 2003). Há diversas opções para a quantificação da imagem e há ainda diversas pesquisas nesta área de análise da imagem visando obter melhores resultados e estimação da intensidade da cor (ver, por exemplo, Speed, 2003; Yang et al., 2002c; Chen et al. 2004), porém maiores detalhes não serão abordados neste trabalho por estar focado na abordagem estatística após a obtenção da medida de intensidade da imagem.

As intensidades de cada uma das cores medidas pelo *scanner* podem variar numa ordem de grandeza de 0 a 2^{16} (65.535). Para obter a medida de intensidade de expressão, cada grupo de *spots* precisa ser identificado e assim a intensidade quantificada. Vale citar que em cada *spot* a leitura de intensidade de imagem é feita em uma resolução de cerca de 2.500 (50x50) pontos (*pixels*) e a medida de intensidade do *spot* é dada por uma medida resumo destes pontos, em geral, a mediana (Cristo, 2003).

A medida da quantidade de cada fluorescente em cada *spot* mostra uma intensidade que está correlacionada à abundância do correspondente RNA transcrito (Huber et al., 2003). Através da avaliação da intensidade de cada uma das fluorescências, pode-se, então, verificar a qual amostra aquele determinado gene (seqüência), alocado num determinado *spot*, pode estar associado (ver Figura 2.2). Desta forma, o nível de expressão gênico é medido pela intensidade de fluorescência obtida para as respectivas cores Verde (G) e Vermelho (R), podendo-se obter, de forma geral, as seguintes colorações em cada *spot*:

- - Gene diferencialmente expresso para o tratamento colorido com G;
- - Gene diferencialmente expresso para o tratamento colorido com R;
- - Gene sem expressão em nenhum dos tratamentos;
- - Gene que não pode ser avaliado (má qualidade da imagem).

Figura 2.2 – Ilustração da lâmina e imagem das cores obtidas em experimentos com *microarrays* (ver: http://en.wikipedia.org/wiki/CDNA_microarray)

2.3. Planejamento de Experimento com *Microarrays*

É de extrema importância o bom planejamento do estudo que será realizado, visando a melhora da eficiência e da confiabilidade dos resultados. No planejamento, vários aspectos devem ser abordados:

- Aspecto científico do estudo: entender o objetivo do experimento e questões específicas a serem respondidas. Aqui vale lembrar que é de grande importância a interação do pesquisador com o estatístico, para melhor esclarecimento dos objetivos do estudo;

- Aspecto prático: definir os tipos e quantidades de amostras de mRNA que farão parte do estudo (controles, amostra de referência, diferentes tratamentos);
- Aspectos específicos sobre o experimento: descrever como será isolada a amostra de mRNA dos indivíduos ou animais selecionados para a pesquisa, como será feita a extração de mRNA, qual a fluorescência utilizada, qual o tipo de lâmina a ser utilizada, quantidade de lâminas, qual plataforma de *microarrays* será usada, entre outros.

Deve-se levar em consideração, também, que diferentes tipos de comparações da expressão gênica podem ser realizadas considerando o local de extração do mRNA. Rosa & Tempelman (2003) discutem estes aspectos:

- Diferentes tecidos de um mesmo organismo (como no cérebro e fígado de um mesmo indivíduo);
- Mesmo tecido em um mesmo organismo, porém pontos diferenciados (no pescoço, parte do tecido com tumor e parte do tecido sem tumor);
- Mesmo tecido, diferentes organismos (como o fragmento da tiróide de um paciente com câncer e o fragmento da tiróide de um paciente controle selecionado em um mesmo hospital);
- Experimentos ao longo do tempo, que são avaliados em diferentes momentos (como uma amostra do tecido de um paciente obeso e depois de um determinado período uma amostra do mesmo tecido deste paciente, após ter feito uma dieta).

Muitas vezes o próprio objetivo do experimento conduz diretamente a um melhor ajuste do planejamento, mas muitas outras vezes, diversas possibilidades de planejamento parecem ser igualmente ajustáveis, o que dificulta a decisão pelo delineamento ótimo. No caso de experimentos com *microarrays*, vários aspectos, que serão discutidos a seguir, devem ser levados em consideração para a escolha do melhor delineamento.

2.3.1. Representação Gráfica do Experimento com *Microarrays* de cDNA

A seguir serão utilizadas representações gráficas para discriminar um experimento de *microarrays* em questão (Yang & Speed, 2002). Considerando um experimento com *microarrays* de cDNA, na Figura 2.3, as caixas indicam as amostras de mRNA de interesse, e a linha interior indica o caminho da comparação que será feita entre dois tratamentos, apontando quais serão comparados diretamente. Outra convenção utilizada é que a parte final da flecha indica que será utilizada para tal amostra a coloração verde (G) e a parte da origem da flecha indica a coloração vermelha (R). Considere as seguintes situações:

- A Figura 2.3a está representando que as amostras de mRNA dos tratamentos A e B serão hibridizadas e comparadas em uma mesma lâmina de cDNA. Para a amostra A será utilizada a coloração verde (G) e para a amostra B a coloração vermelha (R);
- Na Figura 2.3b, as amostras de A e B serão hibridizadas e serão realizadas 5 réplicas de hibridizações. As réplicas correspondem às diferentes lâminas dos mesmos tratamentos, porém, serão discutidas mais adiante;
- A Figura 2.3c mostra que há 3 grupos de tratamentos a serem comparados (A, B e C) com relação às suas expressões gênicas. As amostras de mRNA correspondentes serão hibridizadas e comparadas duas a duas. Partindo de A, este tratamento inicialmente utilizará a coloração G e será comparado com o tratamento B que, por sua vez, receberá a coloração R. A próxima comparação será entre os tratamentos B e C que, pelo diagrama, desta vez, a amostra B receberá coloração verde (G) e a amostra C receberá coloração vermelha (R). A mesma

descrição deverá ser feita para a terceira lâmina que irá comparar os tratamentos C e A;

- A Figura 2.3d representa o delineamento do experimento com ratos, que será analisado posteriormente neste trabalho. Vimos no capítulo anterior que serão analisados quatro tratamentos aleatorizados em 24 animais e cada animal será avaliado em uma diferente lâmina. As comparações que serão feitas em cada lâmina, sempre usarão o mesmo tratamento de referência (TR). Para cada lâmina haverá um número diferente de réplicas, por exemplo, na comparação do Tratamento 1 com o tratamento de referência seis réplicas foram avaliadas, e assim por diante.

Figura 2.3 – Representação de delineamentos de experimentos com *microarrays* de cDNA

2.3.2. Estruturas do Delineamento de Experimentos com *Microarrays*

Os aspectos abordados a seguir são específicos para análises com cDNA, que é o foco deste estudo, porém, considerações relacionadas às réplicas, *pooling* (misturas) de amostras de mRNA e a robustez do experimento servem para experimentos tanto com lâminas com oligonucleotídeos, como para lâminas com cDNA.

Réplicas

A utilização de réplicas, repetições dos experimentos, permite que seja obtida uma estimativa de erro experimental. Utilizando estas estimativas pode-se obter conclusões de diferenças estatisticamente significantes ou não entre as diversas comparações de efeito de tratamentos (Draghici, 2003). Na análise de variância (ANOVA) é realizada a comparação da variação dentro de tratamento (erro experimental) com a variação entre tratamentos (efeito de interesse). Ao utilizar réplicas em experimentos com *microarrays*, deve-se levar em consideração que diferentes repetições de viés ou fontes de variação podem ocorrer paralelamente. Isto se dá pois, muitas vezes, as réplicas de hibridização são realizadas por amostras de um mesmo indivíduo, utilizando-se o mesmo equipamento. Desta forma, se houver algum viés desta amostra, este se repetirá, fazendo com que se percam vantagens em utilizar tais réplicas. Abaixo serão descritas algumas formas de réplicas, porém maiores detalhes poderão ser encontrados em Speed (2003).

As réplicas mais utilizadas são as de duplicação dos *spots* dentro das lâminas, réplicas da técnica e réplicas biológicas. Nas chamadas duplicação dos *spots*, os cDNAs são duplicados em uma mesma lâmina e, freqüentemente,

fixados em posições adjacentes (Speed, 2003). Réplicas da técnica é o termo utilizado para denotar réplicas de hibridizações onde o mRNA dos tratamentos de interesse é de um mesmo *pool* de genes, isto é, é obtido numa mesma coleta de mRNA de um mesmo organismo ou da mesma mistura de mRNA's. Este tipo de réplica envolve menor grau de variação do que réplicas biológicas, porém não produzem independência dos dados, podendo, assim, compartilhar erros sistemáticos que podem ocorrer. As réplicas biológicas referem-se a hibridizações que envolvem mRNA de diferentes extrações biológicas, como exemplo, o mRNA dos tratamentos de interesse são extraídos de diferentes organismos, ou de diferentes locais de um mesmo tecido. Em muitos casos é a melhor forma de réplica, pois amostras obtidas de diferentes organismos apresentam maior independência entre os dados. Contudo, este tipo de réplica biológica pode apresentar maior variabilidade das respostas de intensidade de expressão (ver Speed, 2003; Yang & Speed, 2002).

Em geral, em um experimento poderá haver tanto réplicas da técnica, como réplicas biológicas. As réplicas biológicas servirão para obter estatísticas a partir de dados independentes e para validar as conclusões. Para complementar o estudo, haverá réplicas da técnica para reduzir a variabilidade do estudo (Yang & Speed, 2002).

Experimentos com Inversão da Coloração (*Dye-Swaps*)

Na maioria dos experimentos as colorações verde (G) e vermelha (R) podem sofrer diferenças sistemáticas nas medidas de intensidades, exigindo que sejam feitas correções nas medidas de expressão obtidas. Alguns destes ajustes podem ser feitos por meio de métodos de normalização, que serão discutidos posteriormente, porém há situações onde os métodos de normalização não ajustam corretamente estas diferenças. Desta forma, experimentos conhecidos como *Dye-Swap* podem ser utilizados onde,

inicialmente, numa primeira lâmina, em um dos tratamentos é utilizada a coloração R e no outro tratamento a coloração G. Num segundo momento, ou seja, numa segunda lâmina, é feita uma troca da coloração inicial destas amostras, de forma que a amostra que iniciou com a coloração verde nesta nova hibridização utilizará a coloração vermelha. O diagrama representando este tipo de experimento pode ser visto na Figura 2.4a.

Robustez

Um experimento é robusto quando a eficiência com a qual os efeitos de interesse são estimados não é muito afetada quando uma pequena alteração no planejamento é realizada (Paula, 2004). Pode-se citar situações como a perda de um número de observações devido a, por exemplo, falhas na reação de hibridização, fazendo com que diminua a quantidade de observações utilizadas na estimação dos resultados. Como em experimentos com *microarrays* o tamanho da amostra já é pequeno e a variabilidade entre lâminas é grande, a perda de poucas observações diminui a eficiência dos testes. Deve-se evitar, então, que a perda de observações ocorram e que sejam feitas por meio de uma única e particular hibridização. Desta forma alguns cuidados devem ser tomados para que, dentre as diversas hibridizações existentes, seja possível efetuar todas as estimações de interesse como, por exemplo, garantindo repetições das hibridizações de interesse (réplicas), ou que num mesmo experimento haja possibilidade de comparações diretas e indiretas entre os tratamentos, ponto que será abordado posteriormente.

Mistura de Amostras de mRNA (*Pooling*)

Em experimentos com *microarrays* o número de animais ou indivíduos selecionados para o estudo, em geral, é muito pequeno. Devido a este problema

de dimensionalidade, há duas situações onde é utilizado um *pool* de amostras de mRNA (Speed, 2003):

- Quando a quantidade de mRNA da amostra de um tratamento não é suficiente para a hibridização em uma lâmina;
- Na tentativa de diminuir a variabilidade existente entre as amostras de mRNA selecionadas para o estudo, podendo ser de diferentes seres vivos ou de diferentes locais de um mesmo tecido de um mesmo organismo.

Uma grande vantagem em se utilizar um *pool* de amostras de mRNA é que em um experimento pode-se utilizar um maior número de indivíduos, utilizando poucas lâminas e, assim, diminuir o custo do estudo, uma vez que o custo das lâminas de *microarrays* é bastante alto. Outra vantagem de amostras *pool* é minimizar diferenças devido à variação de indivíduo para indivíduo, ou seja, eliminar possíveis resultados alterados devido a comportamentos aberrantes de alguns indivíduos, uma vez que poucos elementos são utilizados nos experimentos. Estas vantagens são bem atrativas, com resultados equivalentes aos obtidos com experimentos sem o *pool* de amostras, quando nestes experimentos são utilizadas poucas lâminas, em torno de 2 ou 3 (Kendzioriski et al. , 2004).

Ao se utilizar o *pool* de amostras, as inferências sobre a expressão gênica é dada para esta população de *pool*. Este procedimento pode ser vantajoso em situações onde há populações homogêneas geneticamente como, por exemplo, em uma população de animais congênitos. Para a população humana a utilização de amostras *pool* é questionada, devido ao elevado grau de heterogeneidade genética existente entre os indivíduos. Não é claro como entender a expressão gênica neste caso.

Comparação Direta e Indireta

A comparação entre dois tratamentos pode ser direta (ver Figura 2.4a) quando ambos os tratamentos são hibridizados em uma mesma lâmina, ou indireta (ver Figura 2.4b) quando a hibridização em diferentes lâminas é feita relativamente a um mesmo tratamento de referência. Supondo que há interesse em comparar dois diferentes tratamentos, A e B, nas comparações diretas entre estes dois tratamentos as hibridizações ocorrem em uma mesma lâmina. Assim a relação das intensidades de expressão de interesse é obtida diretamente a partir dos dados de uma mesma lâmina (por exemplo, $\log_2(A/B)$). Já em experimentos com comparações indiretas, cada amostra de mRNA do tratamento de interesse é hibridizada em lâminas diferentes. Em cada uma das lâminas, o tratamento de interesse será comparado com um tratamento de referência (TR) em uma mesma lâmina, para depois ser feita a comparação entre os dois tratamentos de interesse. É comum utilizar para os tratamentos de referência, amostras de mRNA oriundas de organismos denominados *wild-type*, sendo um tecido de controle ou um *pool* de todas as amostras de interesse (Speed, 2003). Na comparação indireta podemos chegar na relação entre A e B a partir da equação a seguir:

$$\log_2(A/B) = \log_2(A/ TR) - \log_2(B/ TR).$$

No exemplo apresentado onde o interesse é comparar duas amostras de dois tratamentos (A e B) não há vantagens em se utilizar a comparação indireta (assumindo lâminas de cDNA), uma vez que esta apresenta maior variabilidade. Porém há estudos onde o interesse é comparar mais de dois tratamentos e não há uma quantidade suficiente de amostras de mRNA para que sejam feitas todas as possíveis comparações duas a duas entre os tratamentos e, nestas situações, é vantagem se utilizar a comparação indireta (Yang & Speed, 2002).

Figura 2.4 – Representação gráfica de experimentos com comparações diretas e indiretas

2.3.3. Delineamento de Experimentos com um ou mais Fatores

Quando se trata de experimentos de *microarrays* com cDNA com mais do que dois tratamentos de interesse, após a definição dos aspectos apontados anteriormente, o próximo passo para se ter um bom experimento com alta precisão dos resultados é definir como as amostras de mRNA serão alocadas nas lâminas, duas a duas.

Quando o interesse é identificar genes diferencialmente expressos entre dois tratamentos, que é o tipo mais simples de experimento de *microarrays*, é comum alocar os dois tratamentos em uma mesma lâmina, e alternativas como réplicas e experimento *Dye-Swap* poderão ser utilizados. Para estas comparações, temos que a análise de cada gene será feita para um planejamento considerando um fator de interesse (tratamento) com dois níveis (A e B).

Pode-se comparar experimentos com *microarrays* com experimentos em blocos, onde cada lâmina é considerada um bloco de tamanho 2. Em cada

bloco, há a possibilidade de se avaliar apenas dois tratamentos. Em casos onde há mais do que dois tratamentos de interesse, não há como se avaliar todos em uma mesma lâmina, e desta forma o delineamento é considerado um planejamento em blocos incompletos. Nestas situações, onde há interesse em se comparar mais do que dois níveis de um fator (mais do que dois tratamentos) ou, até mesmo, experimentos mais complexos como experimentos com dois ou mais fatores, muitos autores (ver, por exemplo: Kerr & Churchill, 2001b; Draghici, 2003; Yang & Speed, 2002; Rosa et al., 2005) exploram as propriedades dos experimentos com uma estrutura de blocos incompletos, uma vez que em uma mesma lâmina de *microarrays* de cDNA apenas 2 tratamentos podem ser comparados.

A estrutura em bloco é usada para aumentar a precisão quando diferentes espécies ou tratamentos são avaliadas em um experimento. Desta forma, espera-se que dentro de um mesmo bloco as condições experimentais sejam homogêneas (Draghici, 2003). Partindo do conceito de blocos incompletos, nem todos os tratamentos serão comparados em um mesmo bloco. Dentro de um mesmo bloco as condições experimentais serão mais homogêneas e, possivelmente, heterogêneas entre os blocos. Considerando os experimentos com *microarrays*, as lâminas são consideradas como blocos e os *spots*, representando os genes de interesse, são as unidades experimentais que recebem os tratamentos (Kerr & Churchill, 2001b).

Nos experimentos com mais de dois tratamentos de interesse, há uma alternativa de delineamento que é introduzir no experimento um tratamento adicional, o tratamento de referência, já apresentado nas seções anteriores. O tratamento de referência será comparado com todos os outros tratamentos de interesse do estudo. Esses experimentos são chamados de delineamentos de referência (Rosa et al., 2005). Vale lembrar que para estes experimentos há a possibilidade de se fazer comparações diretas e indiretas. Nas alternativas diretas, as comparações serão feitas entre tratamentos avaliados em um mesmo

bloco, nas indiretas as comparações serão realizadas entre tratamentos hibridizados em diferentes blocos.

Para exemplificar algumas alternativas de delineamentos, a seguir serão apresentados experimentos com um fator com três níveis (três tratamentos: A, B e C), seguindo exemplos sugeridos por Yang & Speed (2002), onde se quer comparar três amostras de mRNA. Será suposto que todas as comparações duas a duas são de igual importância. Algumas escolhas de planejamento são possíveis para este tipo de estudo.

Seguem algumas situações:

- Delineamento I - TR é uma referência comum – neste estudo são utilizados 3 lâminas para as seguintes comparações: A e TR, B e TR, C e TR. Este planejamento é indicado quando há uma quantidade ilimitada de material de referência, mas uma quantidade limitada das amostras de interesse para comparação (ver Figura 2.5a). Note que trata-se de um delineamento aleatorizado em blocos incompletos desbalanceados, já que todos os tratamentos (A, B e C) são comparados com TR que, por esta razão, tem um maior número de réplicas (três neste caso) versus uma para os demais tratamentos.
- Delineamento II - TR é uma referência comum, com réplicas – como apresentado na Figura 2.5b. Neste estudo são utilizadas 6 lâminas para realizar as mesmas comparações anteriores, porém para cada comparação haverá duas réplicas. Como no caso anterior, trata-se de um delineamento aleatorizado em blocos incompletos e, neste caso, o desbalanceamento envolve seis réplicas do tratamento de referência e duas réplicas para cada um dos demais tratamentos.
- Delineamento III - Não há uma referência comum – neste estudo são utilizadas 3 lâminas para realizar as comparações entre os tratamentos A,

B e C (Figura 2.5c). Para este desenho experimental, também chamado de delineamento *looping*, o tratamento A é comparado com B e C, B é comparado com A e C e, por sua vez, C é comparado com A e B. O número de amostras de mRNA neste estudo é semelhante ao tamanho necessário para o estudo anterior (duas em cada caso), porém com uma menor quantidade de lâminas (3 lâminas). Neste caso, trata-se de um delineamento aleatorizado em blocos incompletos balanceados, em que cada comparação entre dois tratamentos ocorre com o mesmo número de vezes (1 neste caso) e cada tratamento individualmente tem o mesmo número de réplicas (2 neste caso).

Delineamento	Número de lâminas	Numero de amostras por tratamento (réplicas)	Variância média
a) Delineamento I – comparações indiretas	3	$A=B=C=1$	2,0
b) Delineamento II – comparações indiretas	6	$A=B=C=2$	1,0
c) Delineamento III – comparações diretas	3	$A=B=C=2$	0,67

Figura 2.5 – Representação gráfica de delineamentos de experimentos com um fator (Yang & Speed, 2002).

Nos experimentos com *microarrays*, vale enfatizar que a atribuição da específica combinação de dois tratamentos que serão hibridizados juntos em uma lâmina deve ser aleatória, isto é, aleatorização restrita dos tratamentos dentro dos blocos (lâminas). Tal procedimento pode “proteger” as estimativas do efeito de tratamento da influência de possíveis variações devido à lâmina (fornecedor, sujeira, etc.).

Para mais de 3 tipos de tratamentos sob comparação a situação é mais complexa. Para comparar tratamentos com um tratamento de referência o planejamento é semelhante aos delineamentos I e II. Porém, um planejamento semelhante ao delineamento III para um número grande de tratamentos não é desejável devido à grande quantidade de amostras que serão necessárias.

Diversos autores avaliam propriedades destes planejamentos (ver, por exemplo, Kerr & Churchill, 2001a; Yang & Speed, 2002; Churchill, 2002) bem como os planejamentos tipo *Dye-Swap* (ver Figura 2.4a), *looping* (ver Figura 2.5c) e o de referência (ver Figuras 2.5a ou 2.5b). O planejamento de referência é muito utilizado nos estudos com *microarrays* e, por este motivo, serão apresentados alguns aspectos deste delineamento.

Nos planejamentos de referência, em geral, os tratamentos de interesse são rotulados com uma mesma coloração (por exemplo, G) e os tratamentos de referência sempre com a mesma cor (por exemplo, R). Desta forma, há uma tentativa de eliminar uma das fontes de variação sistemática em experimentos com *microarrays* devido ao canal de coloração. Nestes experimentos, em que a comparação envolve I tratamentos de interesse com n genes avaliados, o número de observações aumenta por conter metade das observações do experimento relacionada ao tratamento de referência (TR), pois cada amostra de mRNA dos tratamentos de interesse será hibridizada com uma amostra de mRNA do TR. Como citado anteriormente, uma das grandes dificuldades nos experimentos com *microarrays* é o alto custo das lâminas, sendo este muitas vezes um fator determinante do tipo de planejamento a ser adotado. Há delineamentos em que a quantidade de lâminas utilizadas é a mesma utilizada em planejamentos com tratamento de referência, como são os experimentos tipo *looping*, que é um tipo de planejamento em que as comparações entre os tratamentos são de forma direta e apresenta resultados mais precisos (Kerr & Churchill, 2001a; Landgrebe et al., 2004). Mesmo com a possibilidade alternativa

de utilizar *looping*, por outras razões, a opção por um planejamento com referência tem sido comum em experimentos com *microarrays*.

Muitas vezes há dificuldade em se obter a quantidade de mRNA suficiente para a hibridização em uma lâmina (Speed, 2003). Por sua vez, em geral, a amostra que é utilizada para o TR é oriunda de um pool de amostras e é mais fácil de se obter. Comparando um experimento *looping* e um experimento com referência com uma mesma quantidade de lâminas, pode-se notar que a quantidade de amostras necessárias em um experimento *looping* é maior. Pelo exemplo de experimento *looping* ilustrado na Figura 2.6a, e de experimento com referência ilustrado na Figuras 2.6b, pode-se notar que a quantidade de lâminas é a mesma, porém ao considerar as Tabelas 2.1 e 2.2, observa-se que o número de ocorrências de um determinado tratamento em um planejamento do tipo *looping* é 4 amostras de mRNA, aumentando o custo e tempo do experimento (Draghici, 2003), enquanto que em um experimento com referência este número é 2, ou seja, este último utiliza uma menor quantidade de amostras de mRNA.

Figura 2.6 – Representação gráfica de um experimento tipo *looping* e de referência

Como citado anteriormente, na figura 2.5 pode-se verificar os exemplos de delineamentos de experimento em que se tem o interesse em comparar três tratamentos: A, B e C (Yang & Speed, 2002). No exemplo, são comparados os delineamentos em relação à quantidade de lâminas utilizadas, à quantidade de

amostras por tratamento e ao grau de precisão em cada situação. Pode-se observar situações em que se pode utilizar a mesma quantidade de lâminas com diferentes quantidades de amostras por tratamento, a mesma quantidade de amostras por tratamento e ser necessário uma quantidade de lâminas diferente. Para cada situação a precisão, expressa em termos da média das variâncias obtidas para cada uma das três comparações de interesse (A-B, A-C e B-C), é diferente. Para avaliar a precisão nos três delineamentos, assumiu-se que a variância para o $\log_2$ da razão dentro de uma lâmina para cada gene é σ^2 , sendo $\sigma^2=1$. Observando os resultados deste exemplo, em geral, a melhor precisão se tem quando é realizado delineamentos com comparações diretas, com um maior número de réplicas. O delineamento com comparações indiretas é o que apresenta menor precisão, porém esta precisão pode aumentar com o aumento do número de réplicas.

Tabela 2.1 – Distribuição em Blocos para planejamentos do tipo *looping* com três tratamentos e duas réplicas

<i>Lâminas (Blocos)</i>	<i>Tratamentos</i>	
	<i>G</i>	<i>R</i>
1	B	A
2	B	A
3	A	C
4	A	C
5	C	B
6	C	B

Tabela 2.2 – Distribuição em Blocos para planejamentos com referência com três tratamentos e duas réplicas

<i>Lâminas (Blocos)</i>	<i>Tratamentos</i>	
	<i>G</i>	<i>R</i>
1	A	TR
2	A	TR
3	B	TR
4	B	TR
5	C	TR
6	C	TR

Desta forma, apesar de não apresentar menor precisão, o planejamento com referência tem sido escolhido e apresenta algumas vantagens práticas em sua utilização, como a facilidade de obtenção da amostra de mRNA para o TR, a facilidade em aplicar experimentos com referência, e o controle da variação decorrente do canal de cor, pois em experimentos com *microarrays* são grandes os esforços em minimizar as fontes de variação sistemáticas existentes.

Considerando as situações anteriores, tem-se que, na escolha de um experimento, é claro que o ideal é que o delineamento mais preciso seja escolhido, contudo sempre se deve levar em consideração a disponibilidade dos elementos necessários para o experimento, como: a quantidade de amostras de mRNA de interesse disponível, a quantidade de lâminas disponíveis e seu custo, amostras de referência, etc.

Além do fator tratamento, podemos ter situações onde as amostras de um mesmo indivíduo são comparadas em diferentes momentos (experimentos longitudinais), ou seja, introduz-se na análise mais um fator de interesse no estudo. Diversos planejamentos podem ser realizados onde a comparação das intensidades de expressão gênica entre os tratamentos pode ser feita nos diferentes momentos do estudo, por exemplo: será que os genes que se expressam em dois diferentes tecidos dependem da fase de desenvolvimento do indivíduo?

Deste modo, há situações onde se pode estar interessado em mais de um fator de interesse como, por exemplo, a avaliação de dois tecidos (normal e com tumor) em dois diferentes momentos. Neste caso temos uma análise com dois fatores (tecido e tempo) ambos em dois níveis, isto é, um delineamento fatorial 2x2, necessitando de experimentos mais complexos do que os anteriormente citados (ver Yang & Speed, 2002).

Há também outras situações em que, além dos tratamentos de interesse, os elementos presentes na tecnologia de *microarrays* podem se apresentar como fatores de análise no experimento, por exemplo, os diferentes genes (*spots*) avaliados, as diferentes colorações, as diferentes lâminas (quando houver mais do que uma lâmina para o experimento). Desta forma podemos identificar quatro tipos de fatores de experimento: variedade de amostra de mRNA (tratamentos), coloração, spots e genes (Kerr & Churchill, 2001).

Supondo que todas as comparações entre os tratamentos em um experimento sejam do mesmo interesse, Kerr & Churchill (2001) apresentam um critério de delineamento ótimo para avaliar diversos experimentos, bem como citam algumas considerações a serem feitas na escolha de um bom delineamento:

- planejamentos em que os tratamentos podem ser balanceados com relação ao canal de coloração (*dye*) evitam que haja viés do gene com a coloração, principalmente quando os genes não são replicados na lâmina;
- o bom planejamento é aquele balanceado e com réplica. Balanceado para garantir que o efeito de interesse não seja confundido com outras fontes de variação do experimento, e a réplica para melhorar a precisão das estimativas das medidas de expressão e apresentar graus de liberdade para a estimação dos efeitos de interesse.

Capítulo 3

Normalizações das Medidas de Intensidade de Expressão Gênica

Há muitas fontes de variação que podem ocorrer em um experimento com *microarrays* que devem ser controladas para se prosseguir com a análise. Entende-se por normalização possíveis formas analíticas de ajustar as variações sistemáticas que ocorrem neste tipo de experimento e que afetam a medida do nível de expressão dos genes avaliados. Desta forma, é necessário que a normalização das intensidades medidas seja feita antes de iniciar a análise do nível de expressão gênica (Smyth et al., 2003).

Alguns exemplos de variações sistemáticas que podem ocorrer estão descritas a seguir (ver, por exemplo, Hartemink et al., 2001; Speed, 2003; Draghici, 2003):

- Problemas que podem ocorrer durante a preparação da amostra biológica como a extração e isolamento da amostra de mRNA, variação na introdução da fluorescência. Estes problemas podem ocorrer devido a, por exemplo, erro no material utilizado (pipeta) e flutuação de temperatura;
- Lâminas de má qualidade podem reduzir ou influenciar a eficiência das hibridizações;
- Propriedades técnicas de leitura do *scanner*;
- Em algumas situações a intensidade de fluorescência vermelha (R) tende a ser mais fraca do que a intensidade verde, ou um dos canais de coloração (*dye*) pode brilhar mais do que o outro, interferindo na relação entre as duas intensidades de cores.

A necessidade de normalizações pode ser mais claramente observada em um tipo de experimento onde amostras idênticas de mRNA são rotuladas com diferentes colorações e hibridizadas em uma mesma lâmina. Estes são denominados de experimentos *self-self* (Dudoid et al., 2002). O que se espera é que, mesmo rotuladas com cores diferentes, as intensidades sejam as mesmas, porém frequentemente a intensidade de fluorescência vermelha (R) tende a ser mais fraca do que a intensidade verde (G). Também é comum não haver balanceamento entre a relação das cores R e G entre os genes avaliados dentro de uma mesma lâmina, ou até mesmo entre diferentes lâminas, podendo tais variações ocorrer devido à localização do gene, ou seja, posição dos *spots* na lâmina, devido à origem da lâmina e por outras causas (Smyth et al., 2003).

Em um experimento de *microarrays* utilizando-se cDNA, a normalização dos canais de coloração possibilita que seja feito um balanceamento entre as intensidades dos rótulos de coloração (verde e vermelho), permitindo a comparação dos níveis de expressão entre tratamentos. Porém, como visto anteriormente, podem ocorrer variações devido a outras condições do experimento (Yang et al., 2002b). Desta forma, as normalizações podem ser feitas entre lâminas, dentro de lâminas e para os *spots* de cada lâmina.

Em geral as normalizações são feitas por meio de métodos não paramétricos como o Ajuste Global e técnicas como a de Lowess e Spline. Para estes métodos não paramétricos, as normalizações são realizadas a partir dos gráficos MA-PLOT¹⁵ (Yang et al., 2002b). Este é um gráfico de dispersão das variáveis $M \times A$, podendo ser construído para o total de dados, onde temos que:

- $M_{ig} = \log_2(G_{ig}/R_{ig})$, é o logaritmo na base 2 da razão das medidas de intensidade do g -ésimo gene hibridizado para os tratamentos corados

¹⁵ MA-Plot – trata-se um gráfico de dispersão $M \times A$ mas, por questões de padronização dos termos utilizados em experimentos com *microarrays* e com o intuito de manter este material concordante com outras publicações, este gráfico será citado como MA-PLOT.

com G e R, respectivamente, na j -ésima lâmina. Tal valor representa a variabilidade existente entre ambos os tratamentos;

- $A_{jg} = (\log_2 R_{jg} + \log_2 G_{jg})/2$, é a média do logaritmo das intensidades de expressão do gene g na lâmina j dos dois tratamentos.

Por meio do gráfico MA-PLOT pode-se verificar empiricamente candidatos a genes diferencialmente expressos (DE), ou mesmo diagnosticar alguma tendência nos dados. Quando o interesse é a investigação de genes DE, os genes candidatos são aqueles com altos valores médios de intensidade (A) e com alta variabilidade para o valor do módulo de M . Quando o interesse é identificar tendências, o diagnóstico poderá ser feito observando o comportamento das respostas de intensidade, sabendo que, nestes gráficos espera-se que no eixo das ordenadas os dados estejam dispostos de forma aleatória e ao redor do ponto 0 (Chen et al., 2004) e, desta forma, não apontando para nenhuma tendência.

Havendo necessidade de normalização dos dados, as transformações são feitas baseadas nos valores de M e, a partir deles, novos valores de M são obtidos, denotados por M^* . Após a transformação, novos gráficos são construídos, os denominados M^*A -PLOT. Na Figura 3.1a podemos observar um gráfico não normalizado e na Figura 3.1b um exemplo de gráfico já normalizado. Note que a tendência crescente diagnosticada no gráfico indica que quanto maior a média de expressão do gene (representado por A), maior é a expressão do gene para o tratamento G, e isto ocorre independente de qualquer padrão na classificação dos genes. Por acreditar-se que tal tendência é devida a fontes de variação indesejáveis, comuns a este tipo de experimento, a correção dos dados via normalização, é aconselhável.

Prosseguindo, após obtermos os valores M^* pela aplicação de algum método de normalização, precisamos chegar em G^* e R^* que são as medidas de intensidade já padronizadas que serão utilizadas na análise estatística para

identificação dos genes diferencialmente expressos. Assim, a partir das fórmulas utilizadas para se construir a ordenada e abscissa do MA-Plot, os logaritmos das intensidades normalizadas serão obtidos da seguinte forma (por exemplo, Rosa, 2003):

- $\log_2 G^*_{jg} = A_{jg} + M^*_{jg}/2;$
- $\log_2 R^*_{jg} = A_{jg} - M^*_{jg}/2;$

onde,

$\log_2 G^*_{jg}$ – $\log_2$ da medida de intensidade G padronizada do g -ésimo gene avaliado na j -ésima lâmina;

$\log_2 R^*_{jg}$ – $\log_2$ da medida de intensidade R padronizada para o g -ésimo gene avaliado na j -ésima lâmina;

M^*_{jg} – $\log_2$ da medida de intensidade da razão G^*/R^* do g -ésimo gene avaliado na j -ésima lâmina.

Figura 3.1 – Exemplo de gráficos MA-Plot não normalizado e normalizado (Yang et al., 2001a)

Para a utilização dos métodos de normalização que serão apresentados a seguir, em geral, há uma escolha do conjunto de genes que serão utilizados como base para os cálculos. Três abordagens são geralmente utilizadas para escolha do conjunto de genes (Yang et al., 2002b):

- Todos os genes da lâmina – é usado em situações em que uma pequena quantidade de genes apresenta-se diferencialmente expressa, fato que, frequentemente ocorre em estudos de *microarrays*;
- Genes com expressão constante – usa-se apenas um pequeno conjunto de genes que apresentam nível de expressão que não varia significativamente, mesmo sob condições variadas. Estes são os chamados genes *Housekeeping*¹⁶. É difícil a identificação destes genes que apresentam expressão constante em qualquer experimento, mas é possível localizar genes considerados *Housekeeping temporários*, que não mudam a expressão quando expostos a algumas situações específicas. Uma limitação dos *Housekeeping* é que muitas vezes eles são altamente expressos, e assim não podem ser representativos de outros genes de interesse;
- Elementos de controle – nesta abordagem são utilizadas seqüências sintéticas de DNA ou seqüências de DNA de organismos diferentes das amostras de mRNA que estão sendo estudadas, ou seja, das amostras diferentes daquelas alocadas nos *spots*. Estas seqüências são incluídas em igual quantidade, nos dois diferentes tratamentos do estudo e são colocadas nas lâminas para a hibridização. Como se espera que estes elementos de controle tenham iguais intensidades de R e G, havendo desigualdade entre as intensidades, deverão ser feitas correções, e as normalizações para todos os genes serão feitas a partir das informações obtidas destes elementos controles.

¹⁶ *Housekeeping* - genes responsáveis pela manutenção do metabolismo celular que se mantêm ativos em todas as células do organismo (Zatz, 2004).

Após identificarmos o conjunto de genes que serão utilizados na normalização dos dados, a seguir são apresentados métodos de normalização mais comumente adotados para diversas situações de experimentos de *microarrays*.

3.1. Métodos de Normalização Não Paramétricos

Normalização global

Para este método, que é bastante utilizado em experimentos com *microarrays*, assume-se que o viés sistemático causado pelos diferentes canais de coloração (*dyes*) é o mesmo para todos os genes utilizados na normalização de cada lâmina do experimento. Sendo assim, a relação entre as intensidades de expressão R e G é uma constante (c) e a normalização se faz subtraindo esta constante dos valores M para cada lâmina:

$$M_{jg}^* = M_{jg} - c_j$$

onde,

j ($j=1, \dots, J$), é o número de lâminas do experimento

g ($g=1, \dots, G_n$) é o número de genes do experimento

c_j : uma constante referente à j -ésima lâmina calculada pela média ou a mediana do logaritmo da relação G/R considerando o conjunto dos genes g avaliados na lâmina j (Smyth et al., 2003). Sendo assim o valor de c_j pode ser estimado por:

$$c_j = \frac{\sum_{g=1}^{G_n} \log_2(G_{jg} / R_{jg})}{G_n}$$

É bastante comum utilizar-se a média dos logaritmos das razões (Yang et al., 2001; Smyth et al., 2003), porém a mediana é um estimador mais robusto utilizado em situações em que há presença de valores aberrantes (*outliers*).

Normalização dependente da intensidade – ajuste pela Função de Lowess

Nas situações onde em cada lâmina há a necessidade de normalizações dependentes da intensidade de expressão, o padrão de variação é dependente dos *spots*. Desta forma, a correção em M é feita permitindo a variação entre *spots* e esta variação pode ser observada em um gráfico MA-PLOT. Os valores padronizados de M , isto é M^* , obtidos da normalização, são decorrentes de ajustes baseados em uma função dependente dos valores de A . Estes valores utilizados na normalização serão subtraídos do valor de M , como é mostrado a seguir:

$$M_{jg}^* = M_{jg} - c(A_{jg})$$

onde $j=1,\dots,J$; $g=1,\dots,G$.

Um dos métodos utilizados para esse tipo de normalização é através da regressão de Lowess (Smyth, 2003), isto é, a função $c(A_{jg})$ é estimada através do ajuste por Lowess (Cleveland & Loader, 1995). Esta função de Lowess trata de um suavizador de dados em um diagrama de dispersão que realiza um ajuste robusto e não é afetado pela pequena quantidade de genes diferencialmente expressos que aparecem como valores aberrantes (*outliers*) em um MA-PLOT. Os ajustes são realizados em pequenos subconjuntos de dados, sendo cada um deles formado por frações (f) do conjunto total de dados, sendo $0 < f < 1$ (Yang et al., 2002b). Não há um procedimento ótimo para a escolha do parâmetro de suavização f , sendo geralmente escolhido de forma empírica.

Normalização dependente da intensidade – ajuste pela Função Spline

O método de ajuste pela função *Spline* é uma alternativa ao método de Lowess, também utilizado em situações onde há a necessidade de normalizações dependentes da intensidade em cada lâmina. De modo geral, a

técnica de *Spline* consiste em dividir os valores da variável preditiva (A) em intervalos pré-definidos, seguido por ajustes polinomiais para cada um dos intervalos, de forma que os polinômios se juntem a partir dos nós¹⁷ (do inglês *knot*) de forma suave (Simas, 2003). Neste método, a escolha do local dos nós é controlada por um parâmetro de suavização. Os métodos mais populares para a escolha do melhor parâmetro de suavização são os de Validação Cruzada e Validação Cruzada Generalizada, porém novos métodos estão sendo propostos como o de Máxima Verossimilhança Generalizada e REACT (*Risk Estimation After Coordinate Transform*), como apresentado por Irizarry (2002). Os valores utilizados na normalização dependente da intensidade, isto é, $c(A_{ij})$, são neste caso estimados por meio do ajuste por *Spline*.

Normalização dependente das agulhas

A fixação dos fragmentos de cDNA na lâmina é feita por um grupo de agulhas (do inglês *print-tip*) e é realizada por partes nos diversos sub-grupos de *spots* da lâmina. Há diversos tipos de grupos de agulhas como, por exemplo, grupos com quatro agulhas no formato 2x2, outros com 16 agulhas no formato 4x4. Estas agulhas podem apresentar alguns vieses sistemáticos como diferenças de comprimento das agulhas, diferenças nas aberturas pontas das agulhas, ou deformações ocorridas pelo tempo de impressão (desgaste). Desta forma, pode-se proceder com a normalização dependente das agulhas, que permite ajustar diferenças físicas existentes entre as pontas de impressão destas agulhas, bem como as variações sistemáticas que ocorrem ao longo da lâmina. (Smyth et al., 2003). A forma de ajuste é semelhante às normalizações dependente da intensidade, variando para cada *spot*. Esta normalização também pode ser feita utilizando a metodologia de *Lowess*, porém o ajuste é realizado levando em consideração o subgrupo de agulhas:

¹⁷ Nós - pontos onde os diversos polinômios ajustados irão se juntar para formar a curva de ajuste

$$M_{jpg}^* = M_{jpg} - c(A_{jpg})$$

onde,

$$j=1,\dots,J; \quad g=1,\dots,G \quad \text{e} \quad p=1,\dots,P;$$

$c(A_{jpg})$: é o valor ajustado da medida da intensidade do g -ésimo gene, fixado pela p -ésima agulha na j -ésima lâmina; e P é o número total de agulhas.

Normalização da escala para os grupos de agulha

Após a normalização dentro de cada grupo de agulhas, em cada lâmina, os valores M^* ajustados, estão centralizados ao redor de 0, porém pode haver variabilidade entre os grupos de agulha, sugerindo que seja feito um ajuste de escala para padronizar esta variabilidade.

Uma abordagem utilizada por Yang et al. (2002b) é assumir que as razões G/R de uma determinada lâmina para os *spots* fixados pela p -ésima agulha tenha distribuição Normal com média 0 e variância $a_p^2 \sigma^2$, onde σ^2 é a variância das razões G/R e a_p^2 é o fator da escala para a p -ésima agulha. Para realizar a normalização da escala, os valores de a_p^2 são estimados através do método de máxima verossimilhança, que tem como restrição que $\sum_{p=1}^P \log_2 a_p^2 = 0$, e é apresentado a seguir:

$$\hat{a}_p = \frac{\sum_{z=1}^{n_p} M_{pz}^2}{\sqrt[p]{\prod_{p=1}^P \sum_{z=1}^{n_p} M_{pz}^2}}$$

onde,

M_{pz} é a z -ésima razão do $\log_2(G_{pz}/R_{pz})$ do z -ésimo *spot* fixado pela p -ésima agulha; $z= 1, \dots, n_p$ e $\sum_{p=1}^P n_p = G$, sendo n_p o número de *spots* que são preenchidos pela p -ésima agulha, e G o número total de genes fixados na lâmina.

Uma alternativa robusta é utilizar o desvio da mediana (Md), dada por:

$$\hat{a}_p = \frac{Md_p}{\sqrt[p]{\prod_{p=1}^P Md_p}}$$

onde,

$$Md_p = \text{mediana}_p \{ | M_{pz} - \text{mediana}_p (M_{pz}) | \}$$

Para este procedimento assume-se que uma pequena quantidade de genes serão diferencialmente expressos, assim não sendo afetado por valores aberrantes, e que a distribuição é a mesma ao longo dos grupos de agulhas.

Normalização para estudos Dye-Swap

Um outro tipo de normalização considera experimentos *Dye-Swap*. Para este tipo de normalização utiliza-se os valores encontrados nas duas lâminas (invertidas na atribuição dos corantes), onde as mesmas amostras, em lâminas diferentes, são rotuladas com colorações diferentes. Considere a normalização

$$M_{jg}^* = M_{jg} - c_{jg};$$

onde,

$j=1, \dots, J$; $g=1, \dots, G$ e o valor de c_{jg} se dá por $c_{jg} = (M_{jg} + M'_{jg})/2$, e M_{jg} é calculado através das intensidades da primeira lâmina e M' é calculado através da segunda lâmina que possui as amostras com colorações inversas à primeira lâmina.

Neste caso, pode-se ter também a normalização dependente da intensidade em que $c_{ig}=c(A_{ig})$, sendo $c(A_{ig})$ o valor de normalização estimado pelo ajuste de Lowess para o gráfico $\frac{1}{2}(M + M')$ versus $\frac{1}{2}(A+A')$, utilizando-se todos os genes do experimento (Yang et al., 2002b).

Normalização de várias lâminas

Para possíveis variabilidades existentes entre as lâminas de um estudo, o mesmo método, anteriormente apresentado, que é aplicado para a normalização de escala pode ser utilizado para realizar a normalização entre as várias lâminas (Yang et al., 2001). Durbin & Rocke (2004) propõem outros três métodos para estabilizar a variabilidade entre as lâminas, que são alternativas oriundas dos métodos utilizados em experimentos com oligonucleotídeos.

A normalização entre lâminas é bastante utilizada em experimentos com oligonucleotídeos, uma vez que cada tratamento de interesse é avaliado em uma diferente lâmina. Diferentes métodos são propostos na literatura para estabilizar a variância para este tipo de experimento (ver, por exemplo, Durbin et al., 2002; Li & Wong, 2001). Bolstad et al. (2003) apresentam uma técnica de normalização quantílica interessante para experimentos com oligonucleotídeos, baseada no gráfico Q-Q J-dimensional, onde J é o número total de lâminas a serem padronizadas.

3.2. Métodos de Normalização Paramétricos

Além dos métodos não-paramétricos de normalização apresentados neste capítulo, há, também, métodos paramétricos que serão vistos com maior detalhe no capítulo que segue. Dentre estes métodos, o primeiro foi proposto por Kerr & Churchill (2000) que adotaram um modelo de análise de variância clássico (ANOVA) que permite uma análise mais sistemática das fontes de variação que

podem influenciar as medidas de intensidade de expressão em um experimento de *microarrays*. É modelada a intensidade da medida de expressão levando em consideração possíveis fontes de variação como a lâmina e o canal de cor, bem como o tratamento e os diferentes genes avaliados. Com a inclusão destas variáveis no modelo, a normalização se faz no próprio modelo em que os resíduos representam os valores “normalizados” da medida de intensidade. Um segundo modelo, sugerido por Wolfinger (2001), é um modelo de efeitos mistos que é bastante flexível para modelar possíveis dependências entre as respostas de intensidade de expressão dentro da lâmina.

Capítulo 4

Métodos Estatísticos para Identificação de Genes Diferencialmente Expressos

Neste capítulo serão apresentadas alternativas de análise estatística para a identificação de genes diferencialmente expressos, tanto às relacionadas aos testes que podem ser utilizados, como às que se referem as alternativas de correções para os problemas de múltiplos testes, existentes em experimentos com *microarrays*.

Vários métodos de análise estatística podem ser utilizados dependendo do número e da estrutura dos tratamentos avaliados, das amostras disponíveis e do tipo de delineamento experimental. Nas análises mais simples em experimentos com *microarrays*, o interesse é comparar dois tipos de tratamentos e identificar genes diferencialmente expressos para estas duas condições. Nestas situações a comparação pode ser feita através da análise da estatística $M = \log_2(G/R)$ que representa a expressão diferencial entre dois tratamentos na escala logarítmica. Quando há réplicas biológicas no estudo a identificação de genes pode ser feita de forma paramétrica. Se o interesse é a comparação entre dois tratamentos utiliza-se a estatística *t*-Student como uma opção de análise. Quando há dois tratamentos de interesse com outros fatores a serem considerados na análise, ou quando há mais de dois tratamentos, pode-se utilizar modelos de Análise de Variância. Estes pontos serão melhor descritos nas seções a seguir.

4.1. Comparações entre Duas Condições de Interesse

4.1.1. Método Empírico Através do *Fold Change*

Uma vez que os dados estão normalizados com um dos métodos não paramétricos apresentados no capítulo anterior, a maneira mais simples utilizada para a identificação de genes diferencialmente expressos se faz através do módulo das médias das razões $\overline{M}$ ¹⁸ dada em (4.1). Considerando $M_{jg} = \log_2(G_{jg}/R_{jg}) = \log_2 G_{jg} - \log_2 R_{jg}$, a média para um determinado gene se dá por:

$$\overline{M}_g = \frac{\sum_{j=1}^J M_{jg}}{J} \quad (4.1)$$

onde,

$\overline{M}_g$: é a média da diferença entre o $\log_2$ das intensidades dos tratamentos avaliados para o g -ésimo gene ($g=1, \dots, G$);

J ($j=1, \dots, J$) é o número de lâminas ou réplicas do experimento.

Os genes diferencialmente expressos (DE) são definidos por meio de um ponto de corte empírico e são considerados DE quando ultrapassam este valor denominado *Fold-Change* (Cui & Churchill, 2003). Este método foi bastante utilizado nos primeiros estudos com *microarrays* cujo objetivo era principalmente exploratório. Adotando-se M_c como um ponto de corte, um determinado gene será considerado DE por meio do seguinte critério:

$$|\overline{M}_g| \geq M_c$$

Uma das grandes desvantagens deste método é não considerar um fator de escala, isto é, a variabilidade dos genes nas várias réplicas, ou seja, assume, para todos os genes, uma mesma variabilidade (Speed, 2003). Sendo assim,

¹⁸ M – neste capítulo será utilizado M e não M* para se tratar do log da razão entre as intensidades normalizadas G e R, por considerar que os dados para a comparação e identificação de genes DE já estejam normalizados.

valores aberrantes de intensidade de expressão poderão ser considerados DE de forma errada.

Este método foi utilizado inicialmente para a comparação direta de dois tratamentos, mas pode ser também aplicado para comparação de tratamentos de forma indireta (ver item 2.3.2) onde, por exemplo, dois tratamentos de interesse (A e B) são alocados separadamente em duas lâminas e cada um destes tratamentos é hibridizado com uma mesma amostra de referência (TR). Sendo assim, para o g -ésimo gene avaliado nas lâminas j e j' temos que para cada tratamento A e B:

- $M_{A j g} = \log_2(A/ TR)_{j g}$
- $M_{B j' g} = \log_2(B/ TR)_{j' g}$

Logo, os genes DE podem ser identificados por meio do seguinte critério:

$$| \overline{M}_{A_g} - \overline{M}_{B_g} | \geq M_c$$

onde, $(\overline{M}_{A_g} - \overline{M}_{B_g})_g$ é a diferença entre as intensidades médias dos tratamentos A e B avaliados em diferentes lâminas, para g -ésimo gene ($g=1, \dots, G$). As médias das intensidades de M_A e de M_B , respectivamente para os tratamentos A e B, são dadas por:

$$\overline{M}_{A_g} = \frac{\sum_{j=1}^{J_A} M_{A j g}}{J_A} \quad \text{e} \quad \overline{M}_{B_g} = \frac{\sum_{j=1}^{J_B} M_{B j' g}}{J_B} \quad (4.2)$$

onde, J_A e J_B são os números totais de lâminas em que os tratamentos A e B são avaliados relativamente ao tratamento de referência TR, respectivamente.

4.1.2. Métodos Paramétricos - Estatística t

Na presença de réplicas biológicas, para a comparação das médias de intensidade de expressão gênica entre dois tratamentos, o uso da estatística t -Student, ou simplesmente estatística t , é uma alternativa apropriada para a identificação de genes diferencialmente expressos, pois leva em consideração a variabilidade de cada gene.

Uma escolha de análise é utilizar a estatística t considerando a variabilidade de cada gene em questão e, devido à aplicação de ambos os tratamentos em cada *spot*, considerar a estatística para amostras pareadas (Bussab e Morettin, 2002). Sendo o interesse verificar se a intensidade de expressão gênica é igual entre os dois tratamentos com fluorescência R e G, respectivamente, e que foram hibridizados em uma mesma lâmina, as hipóteses a serem testadas serão:

H_0 : a expressão dos genes nos dois tratamentos são iguais

H_a : a expressão dos genes nos dois tratamentos não são iguais

Considerando $M_{jg} = \log_2 G_{jg} - \log_2 R_{jg}$, e supondo $M_{jg} \sim N(\mu_g, \sigma_g^2)$, $j=1,2,\dots,J$, onde μ_g e σ_g^2 é a média e variância populacional de M_{jg} , respectivamente. Em termos de μ_g , as hipóteses são traduzidas por:

H_0 : $\mu_g = 0$

H_a : $\mu_g \neq 0$

Considerando um nível de significância α , o teste de hipótese para cada gene é realizado por meio da estatística t , dada por:

$$t_g = \frac{\overline{M}_g}{\frac{s_g}{\sqrt{n}}} \quad (4.3)$$

onde,

s_g : desvio-padrão amostral das repostas de intensidade diferencial para um gene específico;

n : é o número de réplicas ou o número de lâminas.

Sob H_0 , a estatística t_g , segue uma distribuição t -Student com $n-1$ graus de liberdade.

Um grande problema com esta forma de análise, utilizando a estatística t clássica, é que, em um experimento com *microarrays*, milhares de estatísticas t serão calculadas para os diversos genes avaliados e comparados entre si. Assim, a possível heterogeneidade na variabilidade entre os genes não está sendo considerada (Cui & Churchill, 2003). Pode também ocorrer situações onde a variabilidade de um determinado gene é pequena levando a um elevado valor da estatística t , mesmo tendo uma medida M pequena. Isto se dá pois, quanto menor a variabilidade maior a chance da medida de expressão do gene ser significativamente diferente de zero, sugerindo que o gene em questão seja considerado diferencialmente expresso (Speedy, 2003). Nestes estudos em que o tamanho da amostra é pequeno, é difícil de estimar a variabilidade de forma eficiente. Devido a esses problemas de pequeno tamanho de amostra, heterocedasticidade entre os genes e da grande quantidade de comparações, podem ocorrer uma grande quantidade de conclusões falso-positivas (considerar um gene DE quando, na verdade, ele não é).

Na tentativa de minimizar estes problemas e encontrar uma métrica apropriada de selecionar genes DE, são apresentadas na literatura formas de análise que penalizam a estatística t (denominada estatística t penalizada). Lönnstedt & Speed (2002) mostram que as estatísticas t penalizadas são superiores ao uso da tradicional estatística t para identificar genes diferencialmente expressos.

Com relação às estatísticas t penalizadas, uma alternativa mais simples é utilizar como fator de escala no denominador da estatística t um desvio padrão global considerando todos os genes (ou um sub-conjunto) ao invés do desvio padrão de cada gene, dada por:

$$t_g = \frac{\overline{M}_g}{s/\sqrt{n}} \quad (4.4)$$

onde,

s : é o desvio-padrão amostral considerando todos os genes do experimento (ou um sub-conjunto deles);

n : é o número de réplicas dos tratamentos, em geral, o número de lâminas.

Outra abordagem é o acréscimo de uma constante (c) no denominador da estatística t clássica para um gene específico g . Assim a estatística penalizada é dada por:

$$t_g = \frac{\overline{M}_g}{(c + s_g)/\sqrt{n}} \quad (4.5)$$

onde,

s_g : é o desvio-padrão amostral para um gene específico;

n : é o número de réplicas.

A constante c pode ser calculada de diversas formas. Tusher et al. (2001) sugere que c seja o menor valor dos coeficientes de variação (CV) calculados para as diversas estatísticas t penalizadas simuladas, considerando como constantes para a simulação os p -ésimos percentis ($p=1,\dots,100$) dos desvios-padrão dos genes do experimento. Desta forma, para cada gene são calculadas 100 estatísticas t considerando os 100 percentis e, assim, serão obtidos 100 CV. Maiores detalhes do cálculo podem ser vistos em Chu et al. (www-stat.stanford.edu/~tibs/SAM/). Por sua vez, Smyth et al. (2003) calcula os 100 percentis dos desvios-padrão dos genes do experimento e considera como constante c o percentil 90% destes desvios-padrão.

Há também uma outra opção, denominada estatística t regularizada, que é uma combinação da informação de um gene específico e da variância global calculada para um conjunto de genes (ver Cui & Churchill, 2003; Baldi & Long, 2001). A estatística é dada por:

$$t_g = \bar{M}_g / \sqrt{\frac{\nu(s/\sqrt{n})^2 + (n-1)(s_g/\sqrt{n})^2}{(\nu + n - 2)}} \quad (4.6)$$

onde,

s : é desvio-padrão amostral global para o conjunto total de genes no experimento;

s_g : é o desvio-padrão amostral para o g -ésimo gene;

n : é o número total de réplicas;

ν : é um parâmetro que determina a relativa contribuição do gene específico na variância global. Maiores detalhes do cálculo de ν pode ser visto em Baldi et al. (2001).

Uma outra alternativa aos métodos apresentados anteriormente é baseada na estatística B (Bayes Empírico). Essencialmente, B é o logaritmo da razão de probabilidades dada por:

$$B_{gj} = \log \frac{P(I_g = 1/M_{gj})}{P(I_g = 0/M_{gj})} \quad (4.7)$$

onde, j é índice de lâmina $j=1, \dots, J$,

e $\begin{cases} I_g = 1, \text{ se } g\text{-ésimo gene é DE} \\ I_g = 0, \text{ caso contrário.} \end{cases}$

Para o uso desta abordagem todos os dados são utilizados na estimativa dos parâmetros de uma distribuição a priori. A estimação será feita a partir do logaritmo da razão das probabilidades, onde o numerador é a probabilidade de

um gene ser diferencialmente expresso e o denominador é a probabilidade de um gene não ser diferencialmente expresso. Ambas as probabilidades são estimadas baseadas nos dados totais e são chamadas de probabilidades a posteriori (Cui & Churchill, 2003). Para maiores detalhes da estimativa da estatística B ver Lönnstedt & Speed (2002) e Long et al. (2001). A estatística é dada por:

$$t_g = \frac{\overline{M}_g}{s_g^* / \sqrt{n}} \quad (4.8)$$

onde s_g^* é o desvio-padrão amostral estimado para o g -ésimo gene a partir da razão a posteriori B.

O valor de p , o nível descritivo do teste, para estas estatísticas penalizadas é encontrado através de análise de permutação (Draghici, 2003). Em Cui & Churchill (2003) são citados diversos pacotes estatísticos que calculam diretamente estas estatísticas t penalizadas e seus correspondentes níveis descritivos.

A estatística t pode, também, ser utilizada nas situações em que o interesse é comparar dois tratamentos em uma comparação indireta. Suponha que dois tratamentos A e B são comparados e cada um hibridizado com um mesmo tratamento de referência (TR), pode-se ter a seguinte hipótese a ser testada:

H_0 : a expressão relativa dos genes entre os tratamentos são iguais

H_a : a expressão relativa dos genes entre os tratamentos não são iguais

Considerando $M_{A_{jg}} \sim N(\mu_{A_g}, \sigma_{A_g})$ e $M_{B_{jg}} \sim N(\mu_{B_g}, \sigma_{B_g})$, em termos de μ_{A_g} e μ_{B_g} , as hipóteses são traduzidas por:

$H_0: \mu_{A_g} = \mu_{B_g}$

$H_a: \mu_{A_g} \neq \mu_{B_g}$

Considerando um nível de significância α , o teste de hipóteses para cada gene é realizado por meio da estatística t clássica, para amostras independentes sendo dada por:

$$t_g = \frac{\overline{M}_{A_g} - \overline{M}_{B_g}}{s \sqrt{1/n_{A_g} + 1/n_{B_g}}} \quad (4.9)$$

onde,

$\overline{M}_A$: é a média de M_A calculada como anteriormente na expressão (4.2);

$\overline{M}_B$: é a média de M_B calculada como anteriormente na expressão (4.2);

s_{Ag} : desvio-padrão amostral de um gene específico no tratamento A;

s_{Bg} : desvio-padrão amostral de um gene específico no tratamento B;

n_{Ag} : número de réplicas para o tratamento A;

n_{Bg} : número de réplicas para o tratamento B.

O desvio padrão amostral s_g comum a A e B, é obtido por:

$$s_g^2 = \frac{(n_{Ag} - 1)s_{Ag}^2 + (n_{Bg} - 1)s_{Bg}^2}{n_{Ag} + n_{Bg} - 2} \quad (4.10)$$

Sob H_0 , a estatística t_g segue uma distribuição t -Student com $(n_{Ag} + n_{Bg} - 2) \cdot \gamma$ graus de liberdade, onde γ é a correção feita nos graus de liberdade se a hipótese de homocedasticidade não estiver satisfeita (ver Magalhães & Lima, 2000).

As estatísticas t penalizadas podem ser generalizadas a outros experimentos como, por exemplo, para situações de comparações indiretas.

Considerando a estatística t penalizada tem-se:

$$t_g = \frac{\overline{M}_{A_g} + \overline{M}_{B_g}}{\sqrt{(c + s_g^2) * \sqrt{1/n_{A_g} + 1/n_{B_g}}}} \quad (4.11)$$

onde:

n_{Ag} : número de réplicas para o tratamento A;

n_{Bg} : número de réplicas para o tratamento B;

s_g : é o desvio-padrão amostral para o g -ésimo gene, comum aos tratamentos A e B, dado em (4.10).

A constante c é calculada como nas situações citadas anteriormente para a comparação direta entre dois tratamentos e o valor de p para esta estatística penalizada é, também, encontrado através da análise de permutação.

4.2. Análise para Comparações com mais de Dois Tratamentos de Interesse

Nas situações onde o interesse é comparar mais do que dois tratamentos, a análise não pode ser feita simplesmente calculando a razão entre as medidas de intensidade entre dois tratamentos avaliados em uma mesma lâmina. Para estas situações uma abordagem de análise utilizada é o modelo de Análise de Variância - ANOVA (Neter et al., 1996). Para análise dos dados de experimentos com *microarrays*, o modelo de ANOVA utiliza como variável resposta de interesse, na grande maioria das vezes, o logaritmo das intensidades de expressão e não o logaritmo da razão das intensidades, dado que o efeito de tratamento é modelado como variável preditora. Há alguns delineamentos de experimentos, como o apresentado na Figura 2.3d do capítulo 2, onde todos os tratamentos são comparados com um tratamento de referência (comparação

indireta), em que é possível utilizar como variável resposta no modelo da ANOVA o logaritmo das razões de intensidade relativamente ao tratamento TR, porém esta é uma alternativa menos eficiente, conforme discutido em Cui & Churchill (2003).

O modelo baseado na ANOVA no contexto de experimentos com *microarrays* foi inicialmente proposto por Kerr et al. (2000). Como alternativa aos métodos de normalização não paramétricos descritos na seção 3.1 do capítulo anterior. Esta proposta permite a normalização dos dados no próprio modelo em uma análise mais sistemática, pois pode levar em consideração, no referido modelo, o ajuste para todas as variáveis indesejáveis que influenciam a medida de intensidade de expressão como: a lâmina, o canal de cor, as agulhas, dentre outras. Desta forma, é realizada a calibração dos dados no próprio modelo de forma paramétrica. De maneira geral, nos modelos apresentados a seguir, é utilizada como variável resposta a medida de intensidade de expressão gênica. Esta medida é dada por y_{ijcg} , que é o logaritmo (na base 2) da intensidade da expressão do g -ésimo gene do i -ésimo tratamento hibridizado pelo c -ésimo canal de coloração na j -ésima lâmina, onde,

$g=1,\dots,G$, sendo G o número de genes avaliados no experimento;

$j=1,\dots,J$, sendo J o número de lâmina ou réplicas do experimento;

$i=1,\dots,I$, sendo I o número de tratamentos do experimento;

$c=1,2$ é o canal de coloração utilizado.

A seguir serão apresentadas duas alternativas de formulação de modelos de Análise de Variância adotando modelos com efeitos fixos e modelos mistos (com efeitos fixos e aleatórios). Para as duas abordagens serão apresentados modelos para o caso geral e para as situações onde é utilizado o planejamento com referência (TR), por ser este muito utilizado nos estudos com *microarrays*.

4.2.1. Modelo com Efeitos Fixos

Na abordagem de análise apresentada por Kerr et al., (2000), o modelo proposto considera as fontes de variação (como lâmina, canal de coloração, etc.) e os fatores os quais se tem interesse de comparação como efeitos fixos. Neste modelo o termo de interesse é a interação entre os fatores gene e tratamento, denotado por GT, que avalia a medida de intensidade do i -ésimo tratamento para o g -ésimo gene e, assim, verifica a expressão do gene em cada tratamento. Sendo assim, o interesse é comparar, para cada gene, se há diferença de expressão entre os tratamentos. O modelo de efeitos fixos considerado assume independência entre as respostas de expressão dos genes avaliados em uma mesma lâmina.

O modelo linear inicialmente considerado por Kerr et al., (2000) é dado por:

$$y_{ijcg} = \mu + A_j + T_i + D_c + G_g + (GT)_{gi} + \varepsilon_{ijcg}, \quad (4.12)$$

sendo que outros modelos foram propostos com a inclusão de novas interações (Kerr & Churchill, 2001a), dados por:

$$y_{ijcg} = \mu + A_j + T_i + D_c + G_g + (AD)_{jc} + (GT)_{gi} + (DG)_{cg} + (AG)_{ig} + \varepsilon_{ijcg} \quad (4.13)$$

onde,

y_{ijcg} : é $\log_2$ da medida de intensidade da expressão gênica do g -ésimo gene no i -ésimo tratamento, colorido pelo c -ésimo canal de coloração na j -ésima lâmina;

μ : média geral das intensidades de expressão gênica;

T_i : é o efeito do i -ésimo tratamento;

A_j : é o efeito do j -ésima lâmina;

D_c : é o efeito de c -ésimo canal de cor;

G_g : é o efeito do g -ésimo gene;

$(AD)_{jc}$: é o efeito da c -ésima coloração na j -ésima lâmina;

$(DG)_{cg}$: é o efeito da c -ésima coloração no g -ésimo gene;

$(GT)_{gi}$: é o efeito do g -ésimo gene no i -ésimo tratamento;

$(AG)_{ig}$: é o efeito do g -ésimo gene na j -ésima lâmina;

ε_{ijcg} : é o erro aleatório, assumidos como independentes e identicamente distribuídos seguindo um modelo Normal homocedástico, $N(0, \sigma^2)$.

As restrições do modelo são:

$$\begin{aligned} \sum_{i=1}^I A_i &= \sum_{c=1}^C D_c = \sum_{i=1}^I T_i = \sum_{j=1}^J AD_{jc} = \sum_{c=1}^C AD_{jc} = \sum_{g=1}^G DG_{cg} = \sum_{c=1}^C DG_{cg} = \sum_{g=1}^G G_g \\ &= \sum_{i=1}^I GT_{gi} = \sum_{g=1}^G GT_{gi} = \sum_{j=1}^J AG_{gj} = \sum_{g=1}^G AG_{gj} = 0. \end{aligned}$$

Logo, as medidas em y_{ijcg} são independentes e normalmente distribuídas com esperança ($E[y_{ijcg}]$) e variância ($Var[y_{ijcg}]$) dadas, respectivamente, por:

$$\bullet \quad E[y_{ijcg}] = \mu + A_j + T_i + D_c + G_g + (AT)_{ji} + (GT)_{gi} + (DG)_{cg} + (AG)_{ig} = \mu_{ijcg}; \quad (4.14)$$

$$\bullet \quad Var[y_{ijcg}] = Var[\varepsilon_{ijcg}] = \sigma^2. \quad (4.15)$$

Tendo y_{ijcg} uma distribuição Normal, em termos dos parâmetros μ_{g1} e μ_{g2} e...e μ_{gI} , que levam em conta o efeito de interação gene-tratamento, a hipótese de interesse a ser testada para cada gene neste modelo é dada por:

$$H_0: \mu_{g1} = \mu_{g2} = \dots = \mu_{gI}; \quad g=1, 2, \dots, G \quad (4.16)$$

H_a : existe pelo menos uma diferença entre os tratamentos.

Considerando um nível de significância α , o teste de hipóteses para cada gene pode ser realizado por meio da estatística F , seguindo com testes de comparações múltiplas.

Fatores adicionais podem ser incluídos ou excluídos do modelo, dependendo da estrutura dos fatores, do interesse do estudo e do delineamento

experimental adotado. Kerr & Churchill (2001a) e Shu & Wang (2005) mostram diferentes alternativas de construção do modelo de ANOVA, de acordo com o delineamento do experimento. Por exemplo, os efeitos de canal de cor e suas interações são estimáveis nos experimentos com os canais de cores invertidos, os chamados experimentos *Dye-Swap* (ver Figura 2.4a) ou em experimentos denominados *looping* como o apresentado na Figura 2.3c. No caso mais comum, em que cada tratamento é avaliado em um único canal de fluorescência, tais efeitos estão confundidos. Outras considerações, por exemplo, é que TG e AG são parcialmente confundidos quando há mais de dois tratamentos; as interações DT e AT, não aparecem no modelo pois, em geral, estão confundidos com A e T respectivamente (Kerr & Churchill, 2001a).

Modelo para os Dados com Tratamento de Referência

No experimento realizado com os 24 ratos, que serve de motivação para este estudo, cada um dos quatro tratamentos de interesse foi comparado com uma amostra de referência. Neste tipo de delineamento, as variações que podem ocorrer devido à coloração não são necessárias, pois há confundimento com tratamento, uma vez que em cada lâmina o tratamento de interesse sempre é rotulado com a mesma coloração, não podendo ser estimado qualquer efeito de coloração que interfira nos tratamentos. A interação entre lâmina e gene (AG) também não deve entrar neste modelo, pois esta interação se confunde com a interação GT (Kerr et al., 2001). O modelo é dado da seguinte forma:

$$y_{ijg} = \mu + A_j + T_i + G_g + (GT)_{gi} + \varepsilon_{ijg} \quad (4.17)$$

onde,

y_{ijg} : é $\log_2$ da medida de intensidade da expressão gênica do g -ésimo gene no i -ésimo tratamento na j -ésima lâmina;

μ : média global;

T_i : é o efeito relativo do i -ésimo tratamento;

A_j : é o efeito da j -ésima lâmina;

G_g : é o efeito do g -ésimo gene;

$(GT)_{gi}$: é o efeito do g -ésimo gene no i -ésimo tratamento;

$\varepsilon_{ijg} \sim N(0, \sigma^2)$, independente e identicamente distribuídos.

As restrições do modelo são:

$$\sum_{i=1}^I A_i = \sum_{i=1}^I T_i = \sum_{g=1}^G G_g = \sum_{i=1}^I GT_{gi} = \sum_{g=1}^G GT_{gi} = 0.$$

A esperança e a variância de y_{ijcg} são dadas, respectivamente, por:

$$\bullet \quad E[y_{ijg}] = \mu + A_j + T_i + G_g + (GT)_{gi} = \mu_{ijg} \quad (4.18)$$

$$\bullet \quad \text{Var}[y_{ijg}] = \text{Var}[\varepsilon_{ijg}] = \sigma^2 \quad (4.19)$$

Tendo y_{ijg} uma distribuição Normal, em termos dos parâmetros μ_{g1} e μ_{g2} e...e μ_{gI} , que levam em conta o efeito de interação gene-tratamento, a hipótese de interesse a ser testada é a mesma dada em (4.16).

Como anteriormente, para um dado α , o teste de hipóteses para cada gene pode ser realizado por meio da estatística F seguindo com os testes de comparações múltiplas.

O modelo utilizando efeitos fixos é bastante flexível sobre os fatores que podem ser incluídos ou excluídos do modelo. A seleção de variáveis vai depender da estrutura dos fatores envolvidos e do tipo de delineamento experimental adotado.

No caso de experimentos com comparações de referência, o modelo de ANOVA, ao invés de ter y_{ijg} como variável resposta, pode considerar como variável resposta M_{ijg} que é o logaritmo (na base 2) da razão entre as

intensidades da expressão gênica do g -ésimo gene avaliado para o i -ésimo tratamento, relativamente ao tratamento de referência na j -ésima lâmina. Além de ajustar as medidas de expressão dos tratamentos de interesse para o tratamento de referência (TR) por meio da variável M, outra alternativa mais indicada é a inclusão das respostas de intensidade de TR como covariáveis no modelo, isto é,

$$M_{ijg} = \mu + A_j + T_i + G_g + TR_{ijg} + (GT)_{gi} + \varepsilon_{ijg} \quad (4.20)$$

onde, T_i é o efeito dos tratamentos de interesse e TR_{ijg} é o valor das repostas de intensidade de expressão do tratamento de referência hibridizado junto com o tratamento i no g -ésimo gene da j -ésima lâmina.

Há vários métodos de inferência apropriados para a estimação dos parâmetros destes modelos, a citar: o método de Mínimos Quadrados (MMQ), o método de Máxima Verossimilhança (MMV) e o de Mínimos Quadrados Ponderados (MMQP), este último sendo indicado quando há heterogeneidade das variâncias da resposta, por exemplo, entre os tratamentos ou entre grupos de genes (ver, por exemplo, Neter et al., 1996).

4.2.2. Modelo com Efeitos Fixos e Aleatórios (Modelos Mistos)

No modelo com efeitos mistos (ver Searle, 1997; Brown & Prescott, 1999; Neter et al., 1996) formalizado no contexto de experimentos com *microarrays*, alguns dos fatores são tratados como aleatórios e outros como fixos, e será avaliado um termo de erro (componente da variância) separadamente para cada efeito aleatório. Também neste modelo o termo de interesse é a interação entre gene e tratamento (GT), considerado um fator fixo, pois quer se verificar se, para um determinado gene, há diferença entre os níveis de expressão gênica nos tratamentos.

O modelo utilizando efeitos aleatórios para experimentos com *microarrays* foi inicialmente proposto por Wolfinger et al. (2001) por apresentar vantagens computacionais na estimação dos parâmetros. O modelo sugerido considera o ajuste das respostas de intensidade de expressão em duas fases. A primeira se baseia em um modelo para a normalização dos dados e a segunda fase em um modelo genético. No modelo de normalização são ajustadas as possíveis variações presentes neste tipo de experimento que podem afetar os dados (como efeito de lâmina, de tratamento e interação entre lâmina e tratamento quando estes são estimáveis) e os resíduos deste modelo representam os valores de intensidade de expressão “normalizados” que serão utilizados no modelo genético. O objetivo deste último modelo é identificar os genes diferencialmente expressos entre os tratamentos.

Em experimentos com *microarrays* a utilização de um modelo misto produz uma abordagem mais vantajosa, pois leva em conta a possível correlação existente entre as observações em diferentes níveis das variáveis envolvidas, permitindo uma melhor e mais completa utilização das informações existentes nos dados gerados destes experimentos (Cui & Churchill, 2003). Desta forma, são levadas em consideração, por exemplo, a possível correlação existente entre as respostas de expressão dos genes em uma mesma lâmina.

Modelo Geral

Modelo Misto para Normalização:

Considere a formalização do seguinte modelo misto:

$$y_{ijcg} = \mu + A_j + T_i + D_c + (AT)_{ij} + r_{ijcg} \quad (4.21)$$

onde,

y_{ijcg} : é $\log_2$ da medida de intensidade de expressão gênica do g -ésimo gene no i -ésimo tratamento, colorido pelo c -ésimo canal de coloração na j -ésima lâmina;

μ : é a média global;

T_i : efeito do i -ésimo tratamento;

D_c : efeito do c -ésimo canal de coloração;

Com os seguintes efeitos aleatórios:

$A_j \sim N(0, \sigma_A^2)$ efeito aleatório da j -ésima lâmina;

$(AT)_{ij} \sim N(0, \sigma_{AT}^2)$ - efeito aleatório do i -ésimo tratamento na j -ésima lâmina;

$r_{ijcg} \sim N(0, \sigma_e^2)$ - efeito aleatório residual.

As restrições do modelo são:

$\sum_{i=1}^I T_i = \sum_{c=1}^C D_c = 0$ e que os termos aleatórios são não correlacionados.

Nos modelos mistos para normalização dos dados de experimentos com *microarrays*, o efeito de canal de coloração será considerado como efeito fixo pois, em geral, há apenas duas possíveis colorações: R e G. O efeito de lâmina é considerado aleatório e, deste modo, todas as respostas de intensidade de expressão mensuradas dentro de uma mesma lâmina são consideradas correlacionadas. Neste caso, todos os termos de interação com a variável lâmina são também considerados aleatórios. Além disso, note que o ajuste do modelo (4.21) é realizado considerando todas as observações (dentro e entre lâminas) para todos os genes.

A esperança, variância e covariância das respostas de expressão y_{ijcg} são dadas, respectivamente, por:

$$\bullet E[y_{ijcg}] = \mu + T_i + D_c = \mu_{ic} \quad (4.22)$$

$$\bullet \text{Var}[y_{ijcg}] = \text{Var}[A_j + (AT)_{ij} + r_{ijcg}] = \sigma_A^2 + \sigma_{AT}^2 + \sigma_e^2 \quad (4.23)$$

$$\bullet \quad \text{Cov}(y_{ijcg}, y_{i'j'c'g'}) = \begin{cases} \sigma_A^2 & \text{para } j = j', i \neq i', c \neq c', \forall g, g' \\ \sigma_A^2 + \sigma_{AT}^2 & \text{para } j = j', i = i', c = c', \forall g, g' \\ 0 & \text{para } j \neq j', \forall i, i', g, g', c, c' \end{cases}$$

Modelo Misto Genético

O modelo misto genético, cujo objetivo é identificar genes DE, tem como variável resposta o resíduo do modelo ajustado para as normalizações apresentado em (4.20) e é dado por:

$$r_{ijcg} = G_g + (DG)_{cg} + (GT)_{gi} + (AG)_{jg} + \gamma_{ijcg} \quad (4.24)$$

onde,

r_{ijcg} : é a estimativa do resíduo do modelo normalizado para o g -ésimo gene no i -ésimo tratamento, colorido pelo c -ésimo canal de coloração na j -ésima lâmina

G_g : é o efeito do g -ésimo gene;

$(GT)_{gi}$: é o efeito de interação do i -ésimo tratamento para o g -ésimo gene;

$(DG)_{cg}$: é o efeito de interação do c -ésimo canal de coloração no g -ésimo gene.

Com os seguintes efeitos aleatórios:

$(AG)_{jg} \sim N(0, \sigma_{AG}^2)$ - efeito aleatório do g -ésimo gene na j -ésima lâmina;

$\gamma_{ijcg} \sim N(0, \sigma_\gamma^2)$ - efeito aleatório residual.

As restrições do modelo são:

$\sum_{g=1}^G G_g = \sum_{i=1}^I GT_{gi} = \sum_{g=1}^G GT_{gi} = \sum_{g=1}^G DG_{cg} = \sum_{c=1}^C DG_{cg} = 0$ e que os componentes aleatórios são não correlacionados.

A esperança, a variância e a covariância de r_{ijcg} são dadas por:

- $E[r_{ijcg}] = G_g + (DG)_{cg} + (GT)_{gi} = \mu_{icg}$ (4.25)

- $Var[r_{ijcg}] = Var[(AG)_{ig} + \gamma_{icg}] = \sigma_{AG}^2 + \sigma_r^2$ (4.26)

- $Cov(r_{ijcg}, r_{i'j'c'g'}) = \begin{cases} \sigma_{GA}^2 & \text{para } j = j', g = g', i \neq i', c \neq c' \\ 0 & \text{para } j \neq j' \vee g \neq g', \forall c, c', i, i' \end{cases}$

Tendo r_{ijcg} uma distribuição Normal, em termos dos parâmetros $\mu_{g1}, \mu_{g2}, \dots, \mu_{gI}$, que levam em conta o efeito de interação gene-tratamento, a hipótese de interesse a ser testada para cada gene neste modelo é dada por:

$$H_0: \mu_{g1} = \mu_{g2} = \dots = \mu_{gI}; g=1, 2, \dots, G$$

H_a : existe pelo menos uma diferença entre os tratamentos

Note que o ajuste do modelo (4.24) é realizado com as observações (dentro e entre lâminas) de cada gene, diferente do ajuste do modelo de normalização (4.21) que utilizou as observações de todos os genes.

Considerando um nível de significância α , o teste de hipótese para cada gene é realizado por meio da estatística F apropriada para comparação de efeitos fixos em modelos mistos (Neter et al., 1996; Mendenhall, 1968; Searle, 1997), seguindo de testes de comparações múltiplas entre tratamentos.

Modelo Misto Completo

Uma alternativa ao ajuste das respostas de expressão em dois estágios é formular os dois modelos conjuntamente tal como em (4.13). Considere a seguinte formulação do modelo misto considerando os dois estágios simultaneamente (normalização e genético):

$$y_{ijcg} = \mu + A_j + T_i + D_c + G_g + (AT)_{ji} + (GT)_{gi} + (DG)_{cg} + (AG)_{ig} + v_{ijcg} \quad (4.27)$$

onde,

y_{ijcg} : é $\log_2$ da medida de intensidade de expressão gênica do g -ésimo gene no i -ésimo tratamento, colorido pelo c -ésimo canal de coloração na j -ésima lâmina;

μ : média global;

T_i : efeito do i -ésimo tratamento;

G_g : é o efeito do g -ésimo gene;

D_c : efeito do c -ésimo canal de coloração;

$(GT)_{gi}$ – é o efeito de interação do i -ésimo tratamento com o g -ésimo gene;

$(DG)_{cg}$ – é o efeito de interação do c -ésimo canal de coloração no g -ésimo gene.

Com os seguintes efeitos aleatórios:

$A_j \sim N(0, \sigma_A^2)$ efeito aleatório da j -ésima lâmina;

$(AT)_{ji} \sim N(0, \sigma_{AT}^2)$ - efeito aleatório do i -ésimo tratamento na j -ésima lâmina;

$(AG)_{ig} \sim N(0, \sigma_{AG}^2)$ - efeito aleatório da j -ésima lâmina no g -ésimo gene;

$v_{ijcg} \sim N(0, \sigma_v^2)$ - efeito residual.

As restrições do modelo são:

$$\sum_{i=1}^I T_i = \sum_{c=1}^C D_c = \sum_{g=1}^G G_g = \sum_{i=1}^I (GT)_{gi} = \sum_{g=1}^G (GT)_{gi} = \sum_{g=1}^G (DG)_{cg} = \sum_{c=1}^C (DG)_{cg} = 0, \text{ e que}$$

os componentes aleatórios são não correlacionados.

A esperança, variância e covariância de y_{ijcg} são dadas, respectivamente, por:

$$\bullet E[y_{ijcg}] = \mu + T_i + G_g + (GT)_{gi} + D_c + (DG)_{cg} = \mu_{icg} \quad (4.28)$$

$$\bullet \text{Var}[y_{ijcg}] = [A_j + (AT)_{ji} + (AG)_{ig} + v_{ijcg}] = \sigma_A^2 + \sigma_{AT}^2 + \sigma_{AG}^2 + \sigma_v^2 \quad (4.29)$$

$$\bullet \quad \text{Cov}(y_{ijcg}, y_{i'j'c'g'}) = \begin{cases} \sigma_A^2 & \text{para } j = j', i \neq i', c \neq c', \forall g, g' \\ \sigma_A^2 + \sigma_{AT}^2 & \text{para } j = j', i = i', c = c', g \neq g' \\ \sigma_A^2 + \sigma_{AG}^2 & \text{para } j = j', g = g', i \neq i', c \neq c' \\ 0 & \text{para } j \neq j', \forall i, i', g, g', c, c' \end{cases}$$

Como no modelo misto genético, a hipótese a ser testada é dada por (4.16) e, considerando um valor de α fixado, será testada por meio da estatística F apropriada (Neter et al., 1996; Mendenhall, 1968; Searle, 1997).

Modelo para os Dados com Tratamento de Referência

A seguir é apresentado o modelo misto em dois estágios, para o caso de experimentos com comparações indiretas, onde todos os tratamentos a serem comparados são avaliados relativamente à uma amostra de referência (TR). Vale lembrar que a resposta de interesse pode ser dada também pela razão das intensidades entre os tratamentos de interesse e o tratamento de referência.

Modelo para Normalização:

O modelo misto pode ser formulado da seguinte forma:

$$y_{ijg} = \mu + T_i + A_j + (AT)_{ij} + \varepsilon_{ijg} \quad (4.30)$$

onde,

y_{ijg} : é $\log_2$ da medida de intensidade de expressão gênica do g -ésimo gene no i -ésimo tratamento na j -ésima lâmina;

μ : média global;

T_i : efeito do i -ésimo tratamento;

Com os seguintes efeitos aleatórios:

$A_j \sim N(0, \sigma_A^2)$ efeito aleatório da j -ésima lâmina;

$(AT)_{ji} \sim N(0, \sigma_{AT}^2)$ - efeito aleatório do i -ésimo tratamento na j -ésima lâmina;

$\varepsilon_{ijg} \sim N(0, \sigma_e^2)$ - efeito aleatório residual.

As restrições do modelo são:

$\sum_{i=1}^I T_i = 0$ e que efeitos aleatórios são não correlacionados

A esperança, variância e covariância de y_{ijg} são dadas, respectivamente, por:

$$\bullet \quad E[y_{ijg}] = \mu + T_i = \mu_i \quad (4.31)$$

$$\bullet \quad Var[y_{ijg}] = Var[A_j + (AT)_{ji} + \varepsilon_{ijg}] = \sigma_A^2 + \sigma_{AT}^2 + \sigma_e^2 \quad (4.32)$$

$$\bullet \quad Cov(y_{ijg}, y_{i'j'g'}) = \begin{cases} \sigma_A^2 & \text{para } j = j', i \neq i', \forall g, g' \\ \sigma_A^2 + \sigma_{AT}^2 & \text{para } j = j, i \neq i', \forall g, g' \\ 0 & \text{para } i \neq i', \forall i, i', g, g', \end{cases}$$

Modelo Misto Genético

O modelo misto genético tem como variável resposta o resíduo do modelo ajustado para as normalizações apresentado anteriormente, dado por:

$$r_{ijg} = G_g + (GT)_{gi} + (AG)_{jg} + \gamma_{ijg} \quad (4.33)$$

onde,

r_{ijg} : é a estimativa do resíduo do modelo normalizado;

G_g : é o efeito do g -ésimo gene;

$(GT)_{gi}$: é o efeito de interação do i -ésimo tratamento para o g -ésimo gene.

Com os seguintes efeitos aleatórios:

$(AG)_{jg} \sim N(0, \sigma_{AG}^2)$ - efeito aleatório da j -ésima lâmina no g -ésimo gene

$\gamma_{ijg} \sim N(0, \sigma_{\gamma}^2)$ - efeito residual

As restrições do modelo são:

$\sum_{g=1}^G G_g = \sum_{i=1}^I GT_{gi} = \sum_{g=1}^G GT_{gi} = 0$ e que efeitos aleatórios são não correlacionados.

A esperança, a variância e a covariância de r_{ijg} são dadas por:

$$\bullet \quad E[r_{ijg}] = \mu + G_g + (GT)_{gi} = \mu_{ig} \quad (4.34)$$

$$\bullet \quad \text{Var}[r_{ijg}] = \text{Var}[(AG)_{jg} + \gamma_{ijg}] = \sigma_{AG}^2 + \sigma_{\gamma}^2 \quad (4.35)$$

$$\bullet \quad \text{Cov}(r_{ijg}, r_{i'j'g'}) = \begin{cases} \sigma_{AG}^2 & \text{para } g = g', j = j', i \neq i' \\ 0 & \text{para } j \neq j', \forall g, g', i, i' \end{cases}$$

Como no modelo misto genético, a hipótese a ser testada é dada por (4.16) e, considerando um valor de α fixado será testada por meio da estatística F apropriada.

Como ressaltado para o modelo de efeitos fixos, também para a proposta de modelos mistos considerando o delineamento de referência, pode-se adotar M_{ijg} como variável resposta ou mesmo incluir as intensidades de tratamento de referência como covariável no modelo de normalização (4.24).

Modelo Completo:

A formulação do modelo misto incluindo os dois estágios simultaneamente se dá por:

$$y_{ijg} = \mu + A_j + T_i + G_g + (AT)_{ji} + (GT)_{gi} + (AG)_{ig} + \nu_{ijg} \quad (4.36)$$

onde,

y_{ijg} : é $\log_2$ da medida de intensidade de expressão gênica do g -ésimo gene no i -ésimo tratamento, na j -ésima lâmina;

μ : média global;

T_i : efeito do i -ésimo tratamento (incluindo o tratamento TR);

G_g : é o efeito do g -ésimo gene;

$(GT)_{gi}$ – é o efeito de interação do i -ésimo tratamento para o g -ésimo gene.

Com os seguintes efeitos aleatórios:

$A_j \sim N(0, \sigma_A^2)$ efeito aleatório do j -ésima lâmina;

$(AT)_{ij} \sim N(0, \sigma_{AT}^2)$ - efeito aleatório relativo do i -ésimo tratamento na j -ésima lâmina;

$(AG)_{ig} \sim N(0, \sigma_{AG}^2)$ - efeito aleatório da j -ésima lâmina no g -ésimo gene;

$\nu_{ijg} \sim N(0, \sigma_\nu^2)$ - efeito residual.

As restrições do modelo são:

$\sum_{i=1}^I T_i = \sum_{g=1}^G G_g = \sum_{i=1}^I GT_{gi} = \sum_{g=1}^G GT_{gi} = 0$, e que os efeitos aleatórios são não correlacionados.

A esperança, variância e covariância de y_{ijg} são dadas, respectivamente, por:

$$\bullet \quad E[y_{ijg}] = \mu + T_i + G_g + (GT)_{gi} = \mu_{ig} \quad (4.37)$$

$$\bullet \quad \text{Var}[y_{ijg}] = \text{Var}[A_j + (AT)_{ji} + (AG)_{ig} + \nu_{ijg}] = \sigma_A^2 + \sigma_{AT}^2 + \sigma_{AG}^2 + \sigma_\nu^2 \quad (4.38)$$

$$\bullet \text{Cov}(v_{ijg}, v_{i'j'g'}) = \begin{cases} \sigma_A^2 & \text{para } j = j', i \neq i', \forall g, g' \\ \sigma_A^2 + \sigma_{AT}^2 & \text{para } j = j, i = i', \forall g, g' \\ \sigma_A^2 + \sigma_{GA}^2 & \text{para } j = j', g = g', i \neq i' \\ 0 & \text{para } j \neq j', \forall g, g', i, i' \end{cases}$$

Tendo y_{ijg} uma distribuição Normal, em termos dos parâmetros μ_{g1} e $\mu_{g2}, \dots, \mu_{gI}$, a hipótese de interesse a ser testada para este modelo é dada por:

$$H_0: \mu_{g1} = \mu_{g2} = \dots = \mu_{gI}; g=1, 2, \dots, G;$$

H_a : existe pelo menos uma diferença entre as médias

Considerando um nível de significância α o teste de hipótese para cada gene é realizado por meio da estatística F apropriada para comparação de efeitos fixos em modelos mistos (Neter et al., 1996 Mendenhall, 1968; Searle, 1997).

Rosa et al. (2005) mostram algumas alternativas de análise considerando o tipo de delineamento e a análise estatística via modelos mistos e modelos bayesianos em experimentos de cDNA.

Há vários métodos apropriados para a estimação dos parâmetros de um modelo com efeitos aleatórios: o método de Máxima Verossimilhança, o método de Máxima Verossimilhança Restrita para as componentes de variância, o de Mínimos Quadrados Generalizados Iterativo, bem como o método Bayesiano (Searle, 1997; Brown & Prescott, 1999).

Esta alternativa de análise em dois estágios pode, também, ser aplicada a modelos com efeitos fixos, com resultados idênticos ao modelo em um único estágio como apresentado em (4.13) (Cui & Churchill, 2003). No modelo com efeitos mistos em dois estágios, não necessariamente, os resultados são

idênticos ao modelo em um único estágio. Mesmos resultados seriam obtidos se não houvesse a presença de um fator comum aos dois modelos, como por exemplo, a lâmina presente no modelo de normalização e na interação AG do modelo genético. Apesar de poder ocorrer, em algumas situações, resultados com pequenas divergências, opta-se pelo modelo em dois estágios. Isto se dá devido a grande quantidade de estimações no modelo em um estágio, acarretando em problemas de convergência dos algoritmos numéricos e impossibilitando a estimação dos parâmetros. O modelo em dois estágios traz vantagens computacionais, permitindo a estimação dos parâmetros de interesse (Yang, 2003) e Wolfinger et al. (2001) cita que os resultados são similares, principalmente em situações onde não há perda de dados. Yang (2003) apresenta uma abordagem para se obter, no modelo em dois estágios, resultados idênticos ao modelo em um estágio.

A análise paramétrica realizada utilizando um modelo de efeitos fixos teve uma contribuição importante aos estudos com experimentos com *microarray* e, além disso, é bastante superior à análise realizada com normalizações não paramétricas, pois leva explicitamente em consideração na análise as diversas fontes de variação sistemáticas que estão presentes em experimentos com *microarrays* (Rosa et al., 2005). Entretanto, diversas vantagens já foram apontadas em se utilizar modelos com efeitos aleatórios. Uma delas, nos estudos com experimentos de *microarrays*, é o fato de que não há o interesse em se comparar os efeitos específicos das fontes de variação como lâmina, canal de coloração, entre outros) e sim verificar o efeito que a inclusão de um determinado fator aleatório tem nas inferências sobre o efeito fixo de interesse que, em geral, produz um ganho em precisão. A outra vantagem é que as conclusões obtidas para os efeitos fixos de interesse em um modelo misto podem ser generalizadas para uma população envolvendo os efeitos aleatórios incorporados no modelo.

4.2.3. Estatística *F* Penalizada

A estatística utilizada para a identificação de diferenças entre os tratamentos, no caso de modelos clássicos de ANOVA é a estatística *F* seguida por testes de comparações múltiplas. Assim, como apresentado para o teste *t*, além do método clássico de cálculo de *F*, há várias correções que podem ser feitas no teste *F* (Cui & Churchill, 2003).

Considerando o modelo de ANOVA clássica dada na expressão (4.38) e um nível de significância α , o teste de hipóteses para a comparação entre os diversos tratamentos para cada gene (*GT*) é realizado por meio da estatística *F* clássica. A seguir são, também, apresentadas estatísticas alternativas, na classe das estatísticas *F* penalizadas, que seguem uma abordagem semelhante àquelas apresentadas para a estatística *t*. As estatísticas *F* penalizadas apresentadas a seguir podem, também, ser utilizadas para comparações por meio de modelos mistos.

Para as estatísticas *F* penalizadas descritas a seguir o valor do nível descritivo *p* só poderá ser encontrado através de análise de permutação (Draghici, 2003). Cui & Churchill (2003) apontam que a estatística *F*, considerando a variância global, para todos os genes e todos os tratamentos, é a mais poderosa, porém está sujeita ao mesmo problema encontrado na análise de *fold-change*, onde a presença de valores aberrantes (*outliers*) pode afetar os resultados. As estatísticas apresentadas a seguir são calculadas para um gene específico e, desta forma, o índice para gene não foi indicado nas fórmulas.

Teste *F* clássico:

$$F = \frac{(SQR_o - SQR_1)/(gl_o - gl_1)}{SQR_1 / gl_1} \quad (4.39)$$

onde,

SQR_0 – soma de quadrados dos resíduos do modelo para a hipótese nula (não há diferença entre os tratamentos);

gl_0 – graus de liberdade sob a hipótese nula;

SQR_1 – soma de quadrados dos resíduos sob a hipótese alternativa;

gl_1 – graus de liberdade sob a hipótese alternativas.

Teste F considerando a variância comum de todos os genes:

$$F = \frac{(SQR_0 - SQR_1)/(gl_0 - gl_1)}{s^2} \quad (4.40)$$

onde,

s^2 – variância amostral da resposta de expressão comum a todos os genes.

Teste F para apenas um gene específico:

A idéia é análoga à abordagem apresentada anteriormente para o teste t regularizado, onde é utilizado a variabilidade específica de cada gene e a variabilidade total, considerando todos os genes. Tem-se:

$$F = \frac{(SQR_0 - SQR_1)/(gl_0 - gl_1)}{[(SQR_1 / gl_1) + s^2]/2} \quad (4.41)$$

onde s^2 é dado como em 4.40.

4.3. Gráfico Vulcão

O gráfico Vulcão (do inglês *Volcano*) é uma maneira exploratória de identificar genes diferencialmente expressos dispondo, simultaneamente, em um mesmo gráfico, os resultados das duas análises apresentadas anteriormente: a empírica (*fold-change*) e a paramétrica (via testes).

O gráfico Vulcão é bastante esclarecedor e de visualização prática. É um gráfico de dispersão que apresenta os resultados encontrados através da análise do *fold-change* e um teste paramétrico apropriado de acordo com o delineamento do experimento, sendo este teste um dos apresentados anteriormente. No eixo das abscissas (X) são colocados os valores do *fold-change* ($\bar{M}$) para cada gene e no eixo das ordenadas é colocado o valor de z ($z = -\log_{10} p_g$), onde p_g é o valor do nível descritivo (ver, por exemplo, Neter et al., 1996) da estatística do teste para o g -ésimo gene (ver Speed, 2003; Wolfinger et al., 2001; Draghici, 2003).

Para a definição dos pontos diferencialmente expressos deverá ser levado em consideração os resultados das duas análises, definindo-se empiricamente um ponto de corte para cada uma das formas de análise. Suponha o exemplo considerado na Figura 4.1. Foi definido como ponto de corte para o *fold-change* $|M| > 2$ e um p -valor $< 0,00005$, conforme as linhas tracejadas. Serão considerados genes diferencialmente expressos (DE) aqueles que satisfizerem estas duas condições apontadas nas áreas hachuradas na Figura 4.1a. Nesta figura também pode-se observar os valores falso-positivos, definidos como aqueles que aparecem como DE utilizando-se o método de *fold-change*, porém não aparecem no método paramétrico, e os valores falso-negativos, aqui sendo considerados como aqueles que não aparecem como DE pelo método de *fold-change*, mas aparecem pelo método paramétrico. O gráfico da Figura 10b é um exemplo de como os dados ficam dispostos em um gráfico Vulcão.

Os gráficos denominados Vulcão podem ser, também, apresentados quando há interesse em se comparar diversos tratamentos, sendo cada ponto no gráfico o resultado da comparação de cada dois tratamentos, ou seja, quando há mais do que dois tratamentos no experimento, nos gráficos são colocados os resultados de cada comparação entre dois tratamentos para cada gene. Supondo que foi avaliado três tratamentos (A, B e C), um dado ponto no gráfico representa a comparação de, por exemplo, A com B para um determinado gene; um outro ponto será a comparação entre A e C, para o mesmo gene e assim por diante. Desta forma, nos gráficos, para cada gene, aparecerão três pontos representando as comparações entre os tratamentos dois a dois (A e B; A e C; B e C).

a)

b)

Figura 4.1 – Representação do Gráfico Vulcão (Wolfinger et al., 2001)

4.4. O Problema dos Múltiplos Testes

Supondo que seja adotado um nível de significância α , para a comparação realizada para um determinado gene, temos que há uma probabilidade α de ocorrer um resultado falso positivo, isto é, de se rejeitar a hipótese nula quando ela é verdadeira. Para estudos com experimentos de *microarrays*, há uma probabilidade α de que o nível de expressão de um determinado gene seja DE, quando na verdade ele não é DE (Erro tipo I).

Considerando uma abordagem formal, onde quer se testar a hipótese nula (H_0) de que a média da intensidade de expressão gênica é igual entre dois tratamentos A e B, o resultado do teste é avaliado por meio do valor do nível descritivo (p) baseado em algum modelo e pressuposto.

Seja a hipótese a ser testada para um específico gene definida como:

$$H_{0g}: \mu_{Ag} - \mu_{Bg} = 0$$

$$H_{ag}: \mu_{Ag} - \mu_{Bg} \neq 0$$

Como citado, o nível de significância para este teste é α dado por:

$$\alpha = P(\bar{M}_g \neq 0 / H_{0g} \text{ é verdadeiro})$$

Sendo assim, a probabilidade do resultado estar correto, ou seja, do gene ser considerado diferencialmente expresso, e ele realmente o ser, é dada por:

$$P(\bar{M}_g \neq 0 / H_{0g} \text{ é falso}) = 1 - \alpha$$

Considerando todo o experimento, isto é, para $g=1, \dots, G$ temos que a probabilidade de não ocorrer falso-positivos em todo o experimento é dada por:

$$\prod_{g=1}^G P_g(\bar{M}_g \neq 0 / H_{0g} \text{ é falso}) = (1-\alpha)(1-\alpha)(1-\alpha)\dots(1-\alpha) = (1-\alpha)^G$$

Sendo assim o nível de significância do teste para todo o experimento é dado por:

$$\alpha^* = P(\bar{M}_g \neq 0 / H_0 \text{ é verdadeiro}) = 1 - (1 - \alpha)^G$$

onde,

α^* - é o nível de significância adotado para todo o experimento, considerando as comparações para os G genes.

Como em experimentos com *microarrays* o número de genes avaliados simultaneamente é muito grande, a comparação será realizada milhares de vezes, levando ao problema de múltiplos testes, onde se pode acumular um número excessivamente grande de resultados falso-positivos (genes considerados DE, quando na verdade não o são).

Há vários métodos propostos na literatura para corrigir o problema de múltiplos testes e alguns detalhes serão apresentados a seguir. Com o apoio destes procedimentos, após um valor de α ter sido corrigido para uma única hipótese H_{0g} a ser testada, este será definido como o nível de entrada do teste no qual a hipótese H_{0g} poderá ser rejeitada, e um novo valor crítico dos testes será considerado. Há várias abordagens para se calcular o p-valor ajustado dependendo da rigidez da correção (Dudoit et al., 2002). Desta forma, a hipótese H_{0g} será rejeitada para o g -ésimo gene no controle FWER (*Family-Wise Error Rate*) se:

$$p \leq \alpha^*$$

4.4.1. Controle FWER (*Family-Wise Error Rate*)

Nesta metodologia é controlada a probabilidade de acumular um ou mais resultados falso-positivos em situações de múltiplos testes estatísticos.

Dentre as alternativas, a correção de Bonferroni é o método mais simples e conhecido para problemas de múltiplos testes, onde o novo nível de significância utilizado (α^*) será o nível de significância inicialmente adotado (α) dividido pelo número de comparações (m), isto é,

$$\alpha^* = \alpha / m$$

onde,

α : nível de significância adotado para a família de comparações;

α^* : nível de significância individual corrigido

Este ajuste obtido pelo critério de Bonferroni é considerado um método conservador, pois se um gene é DE após a correção ajustada, então o gene é realmente DE entre os grupos comparados. Porém se um dado gene não é considerado DE de acordo com esta correção, ele pode, ainda assim, ser DE sob outros critérios (Draghici, 2003).

4.4.2. Controle FDR (*False Discovery Rate*)

Uma outra alternativa é o controle FDR (do inglês *False Discovery Rate*). Trata-se de uma abordagem utilizada para a correção dos múltiplos testes que considera a proporção de falsos positivos entre todos os genes considerados diferencialmente expressos, ou seja, aqueles onde a hipótese a ser testada foi rejeitada.

Os métodos de ajuste FWER para múltiplos testes consideram que as comparações são independentes. O método de FDR leva em consideração a correlação entre as comparações.

	número de não rejeitados	número de rejeitados	
Número de H_0 verdadeiros	U	V	m_0
Número de H_0 não verdadeiros	T	S	m_1
	W	R	m

Tabela 4.1 – Tabela resumo para o problema de múltiplos testes (Ge et al., 2003)

Considere a Tabela 4.1, onde,

R - é o número de hipóteses rejeitadas;

W - é o número de hipóteses não rejeitadas;

V - é o número de falso-positivos;

T - é o número de falso-negativos;

m_0 - é o número de hipóteses nulas verdadeiras;

m_1 - é o número de hipóteses nulas não verdadeiras.

A seguir será apresentado como se obtém o critério FDR. Considere Q uma variável aleatória, tal que:

$$Q = \begin{cases} \frac{V}{R} & \text{se } R > 0 \\ 0 & \text{se } R = 0 \end{cases}$$

onde,

Q - é a proporção de falso-positivos entre aqueles que a hipótese nula foi rejeitada

Sendo assim,

$$FDR = E(Q)$$

isto é, FRD é o valor esperado de Q .

A seguir, estão apresentados os passos para esta abordagem (Draghici, 2003):

Para calcular m testes a serem realizados,

- calcular o valor do nível descritivo p_r obtido a partir de algum modelo estatístico adotado;

- estes valores são ordenados de forma crescente sendo $r=1, \dots, R$, onde,

p_r - é o r -ésimo p_r valor após a ordenação;

- comparar os valores dos $p_{(r)}$ através do seguinte critério:

$$p_1 < \frac{1 * \alpha}{r}, p_2 < \frac{2 * \alpha}{r}, p_3 < \frac{3 * \alpha}{r}, \dots, p_r < \frac{r * \alpha}{r}$$

- se $p_r < \frac{1 * \alpha}{r}$, para o r -ésimo valor de p_r ordenado, o gene correspondente a este resultado será considerado DE.

Há também uma variação do FDR , denominada $pFDR$ (do inglês *positive false discovery rate*) que foi proposto por (Storey, 2002). A estatística FDR , neste caso, é multiplicada por um fator π_0 que é uma estimativa da proporção de genes que não são DE entre todos os genes. Como π_0 é um valor entre 0 e 1, os valores de $pFDR$ são menores do que os valores de FDR , ou seja, este critério é mais rigoroso para a identificação de genes DE (Cui & Churchill, 2003).

Os métodos FWER são mais conservadores e devem ser utilizados quando uma menor quantidade de comparações é realizada. O método FDR é mais apropriado, uma vez que leva em consideração a possível dependência entre as comparações e é computacionalmente eficiente.

Foram apresentados apenas alguns métodos de correção para múltiplos testes bastante utilizados para experimentos com *microarrays*, porém outras

alternativas são consideradas por diversos autores (ver, por exemplo, Ge et al., 2003; Speed, 2003; Draghici, 2003). Os vários métodos de correção para o problema de múltiplos testes estão implementados em softwares e pacotes estatísticos. Cui & Churchill (2003) apresentam diversas possibilidades.

Capítulo 5

Aplicação

Nos Capítulos 3 e 4 foram apresentadas as possíveis formas de análise estatística de dados em experimentos com *microarrays*. No presente capítulo são apresentados os resultados obtidos da análise estatística realizada com os dados dos ratos submetidos aos quatro tratamentos apresentados no primeiro capítulo. Inicialmente serão realizadas análises visando identificar genes DE entre os tratamentos, através da ANOVA com um fator, com os dados já normalizados por um método não paramétrico. Na seqüência, será utilizada uma análise paramétrica através de um Modelo Misto seguindo o modelo de Wolfinger et al. (2001) e, por último, um modelo de ANOVA com vários fatores com efeitos fixos. Serão verificados os genes diferencialmente expressos através destes modelos. Os resultados encontrados nas análises paramétricas serão comparados empiricamente com o método de *Fold-Change* através do gráfico Vulcão. Para as análises considerando os modelos de efeitos fixos o método de estimação utilizado foi o de máxima verossimilhança. Nas análises considerando modelos de efeitos mistos o método de estimação utilizado foi o de máxima verossimilhança restrita. Nas análises foram utilizados os recursos computacionais do programa SAS 8.02 e para a análise do modelo com efeitos mistos foi utilizada a macro adaptada da versão apresentada por Wolfinger et al. (2001). Os códigos da programação utilizada no ajuste dos modelos estão apresentados no Apêndice 2.

Em todas as análises, optou-se pelo método de Bonferroni para correção do problema dos múltiplos testes. Este método, apesar de excessivamente conservador, principalmente no contexto de *microarrays*, foi escolhido neste trabalho como um critério de correção para os múltiplos testes, visando comparar os genes DE identificados segundo diferentes abordagens de análise

estatística. Ao ser adotado o método de Bonferroni, pode-se ter um aumento do número de resultados falso-negativos em prol da proteção aos erros do tipo I (resultados falso-positivos). Estes aspectos podem ser investigados por meio do gráfico Vulcão.

Os dados do estudo estão dispostos da seguinte forma: foram analisados 24 animais divididos em quatro tratamentos conforme descrito no primeiro capítulo. A análise de cada animal foi realizada em uma diferente lâmina e em cada lâmina a amostra de mRNA de cada animal foi comparada com uma amostra de referência, que foi a mesma para todos os animais. Desta forma, é considerado um total de 5 tratamentos para as análises (4 tratamentos de interesse e TR). Em cada lâmina do estudo há 12.768 *spots*, ou seja, 12.768 genes podem ser avaliados porém, na leitura da imagem pelo scanner alguns *spots* podem apresentar problemas e serem eliminados da análise. O software que avalia a medida de intensidade das colorações, indicando a medida da expressão gênica, mostra um código indicando o grau de eficiência da leitura de cada gene avaliado. Este indicador é denominado *Flags* podendo ter as pontuações 0, -50, -75 e menor do que -75. Valor 0 para a variável Flag aponta *spots* com boas especificações para análise, *flags* -50 podem indicar alguns dados não confiáveis e os anotados com *flags* -75 ou menos são observações consideradas de má qualidade devido a algum tipo de problema (problemas na hibridização, sujeira, etc.). A distribuição dos dados com relação aos *flags* nas 24 lâminas está mostrada na Tabela 5.1, onde se pode observar que 124.595 (40,6%) dados foram excluídos da análise, dado que optou-se pelo critério de eliminar genes com medidas de expressão de qualidade igual ou inferior a *flag* -50.

Tabela 5.1 – Distribuição dos dados por Flags

Flag	N	%
Flag = -75	960	0,3
Flag = -50	123635	40,3
Flag = 0	181837	59,3
Total	306432	100,0

No Apêndice 1 está apresentado um exemplo dos dados para um dos animais do primeiro tratamento. No banco de dados há a informação da intensidade para o tratamento de interesse (G), bem como a respectiva intensidade da amostra de referência para cada gene (R), ainda não corrigidos pelo *background*.

5.1. Modelo I – Análise dos Dados Utilizando Normalização Não Paramétrica

Neste primeiro modelo, inicialmente foram construídos gráficos de dispersão de $M \times A$ (MA-Plot), para verificar a necessidade se de fazer normalizações devido a algumas tendências. Os gráficos estão apresentados na Figura 5.1, onde podemos observar que para todos os tratamentos há necessidade de normalização por apresentarem uma tendência. Até mesmo para o tratamento 1 que teve pouco desvio com relação à linha central ($M=0$). Verificada a necessidade de correções, foram feitas normalizações utilizando o método de Lowess apresentado no Capítulo 3. Após a normalização realizada em cada uma das lâminas, foram novamente construídos gráficos MA-Plot com os dados já corrigidos (M^*A -Plot). Os resultados dos gráficos estão apresentados na Figura 5.2, onde podemos observar que os dados estão ajustados e não apresentam nenhuma tendência considerada marcante.

Figura 5.1 – Gráficos MA-PLOT para os diferentes tratamentos

Figura 5.2 – Gráficos MA-PLOT para os diferentes tratamentos com dados normalizados

Com os dados já normalizados prosseguiu-se a análise, onde foi utilizado o método de Análise de variância com um fator (a variável tratamento). A comparação entre os tratamentos foi feita para cada gene. Foram realizados ajustes dos múltiplos testes através do método de Bonferroni (α dividido pelo número de comparações de interesse, que para este caso foi 60.341).

A variável resposta para este modelo foi a diferença entre os logaritmos de intensidades G e R ($M_{jg} = \log_2 G_{jg} - \log_2 R_{jg}$).

Segue o modelo :

$$M_{ij} = \mu_g + T_{ig} + \varepsilon_{ij}$$

onde

M_{ij} é $\log_2$ da razão das medidas das intensidade G e R normalizadas para o j -ésimo animal (lâmina) no i -ésimo tratamento do g -ésimo gene

$g=1, \dots, 12.768$ é o número de genes avaliados no experimento

$j=1, \dots, 24$ é o número de animais do experimento

$i=1, \dots, 4$ é o número de tratamentos de interesse do experimento

μ_g : média global considerando o g -ésimo gene

T_{ig} é o efeito relativo do i -ésimo tratamento (comparado com o tratamento de referência TR)

$\varepsilon_{ij} \sim N(0, \sigma^2)$ e independentes

Tendo M_{ij} uma distribuição Normal, com média $\mu_{ig} = \mu_g + T_{ig}$. Em termos de $\mu_{g1}, \mu_{g2}, \mu_{g3}$ e μ_{g4} , a hipótese de interesse a ser testada para este modelo é dada por:

$$H_0: \mu_{g1} = \mu_{g2} = \mu_{g3} = \mu_{g4} \Leftrightarrow H_0: T_{ig} = 0 \quad \forall i = 1, 2, 3, 4$$

H_a : existe pelo menos uma diferença entre as médias

O ajuste do Modelo I permitiu considerar DE apenas 17 genes. Na Tabela 5.2 segue a distribuição da quantidade de genes DE para cada comparação entre tratamentos, considerando o resultado do ajuste do Modelo I, indicando que alguns genes foram DE em mais de dois tratamentos. Para os 17 genes houve 25 comparações, estatisticamente significantes. A Tabela 5.3 mostra as estimativas das diferenças de expressão entre os tratamentos para todos os genes DE entre pelo menos dois tratamentos, ou seja, que obtiveram nível descritivo menor do que $8,29 \cdot 10^{-7}$ (valor obtido após a correção de Bonferroni onde foi dividido o valor do nível de significância de 5% para cada teste individual pelo número de comparações estimadas). Observando as estimativas

das diferenças pode-se notar, por exemplo, que há diferença significativa na expressão esperada dos genes G1444, G5047 e G6652 entre o tratamento 1 e os demais tratamentos, sendo que os G1444 e G6652 apresentam menor expressão gênica no tratamento 1, isto é, em animais diabéticos e não exercitados do que nos demais. Já o gene G5047 se apresenta mais expressos em ratos diabéticos não exercitados (tratamento 1) do que nos outros três tratamentos. Os genes G328, G1793, G1924, G6602, G8234, G9794, G9897 e G11570 só apresentam diferença entre os tratamentos 2 e 4, evidenciando que são mais expressos em ratos que receberam o tratamento 2, relativamente ao tratamento 4. O gene G8129 é DE quando comparado o tratamento 4 com os tratamentos 2 e 3, sendo que este gene se expressa mais em ratos que passaram pelo tratamento 2 e 3 do que em animais não diabéticos e não exercitados (tratamento 4).

Tabela 5.2 – Genes considerados DE para cada comparação entre os tratamentos – Modelo I

	<i>n</i>	%
Tratamento1 x Tratamento2	4	16
Tratamento1 x Tratamento3	5	20
Tratamento1 x Tratamento4	6	24
Tratamento2 x Tratamento3	0	0
Tratamento2 x Tratamento4	9	36
Tratamento3 x Tratamento4	1	4
Total	25	100

Depois de verificado os resultados do modelo, foi construído o gráfico Vulcão entre os resultados dos níveis descritivos do Modelo I versus o *fold change*. Observando a Figura 5.3, e considerando as duas formas de análise contidas neste gráfico, nota-se que das 25 comparações com genes DE, apenas 13 são consideradas DE pelos dois critérios. Pode-se observar também os genes que aparecem com alto valor de *fold change*, sugerindo serem DE, mas na verdade não são. Note que, para esta abordagem, o critério de validação de genes está fixado em valores empíricos de níveis descritivos individuais menores ou iguais a 10^{-5} ($-\log_p=5$)

5.2. Modelo II – Anova com Modelos Mistos (Modelo de Normalização e Genético)

Considerando o modelo proposto por Wolfinger et al. (2001), esta análise é realizada em dois estágios: primeiro é feita a normalização (paramétrica) dos dados, seguindo para a análise genética, ambas utilizando um modelo misto, com efeitos fixos e aleatórios. A seguir estão apresentados os modelos, considerado nas expressões 4.29 e 4.32, onde foi utilizada como variável resposta o logaritmo (na base 2) das intensidades das medidas de expressão do j -ésimo animal no i -ésimo tratamento para o g -ésimo gene (y_{ijg}), sendo:

Modelo para normalização:

$$y_{ijg} = \mu + T_i + A_j + (AT)_{ij} + \varepsilon_{ijg}$$

Modelo Misto Genético:

$$r_{ijg} = G_g + (GT)_{gi} + (AG)_{ig} + \gamma_{ijg}$$

onde,

$g=1,\dots,12.768$ é o número de genes avaliados no experimento;

$j=1,\dots,24$ é o número de lâmina ou réplicas do experimento;

$i=1,\dots,4$ é o número de tratamentos do experimento;

No modelo genético, tendo r_{ijg} uma distribuição Normal, em termos de $\mu_{g1}, \mu_{g2}, \mu_{g3}, \mu_{g4}$ e $\mu_{g\text{ref}}$, que são as médias de cada tratamento para um determinado gene, a hipótese de interesse a ser testada via este modelo é dada por:

$$H_0: \mu_{g1} = \mu_{g2} = \mu_{g3} = \mu_{g4} = \mu_{g\text{ref}} \Leftrightarrow H_0: GT_{gi} = 0 \quad \forall i = 1, 2, 3, 4$$

H_a : existe pelo menos uma diferença entre as médias

Na Tabela 5.4 segue a distribuição da quantidade de genes DE para cada comparação entre tratamentos, considerando o resultado do Modelo II. A Tabela 5.5 mostra as estimativas das diferenças das médias de expressão entre tratamentos apontando quais os genes DE que obtiveram nível descritivo menor do que $8,23 \times 10^{-7}$, valor encontrado após a correção de Bonferroni. Podemos notar que animais não diabéticos e não exercitados (tratamento 4) apresentaram uma maior quantidade de genes DE com relação aos demais. As comparações com o tratamento de interesse não foram apresentadas por não ser interesse do estudo.

Tabela 5.4 – Distribuição do número de genes DE consideradas por comparações – Modelo II

	<i>n</i>	%
Tratamento1 x Tratamento2	5	7,1
Tratamento1 x Tratamento3	10	14,3
Tratamento1 x Tratamento4	17	24,3
Tratamento2 x Tratamento3	1	1,4
Tratamento2 x Tratamento4	20	28,6
Tratamento3 x Tratamento4	17	24,3
Total	70	100,0

Observando as estimativas das diferenças, pode-se notar, como exemplo, que há diferença dos genes G1444, G10255, G11160, e G11258 entre o tratamento 1 e os demais tratamentos, sendo que os genes G1444, G10255, e G11160 apresentam menor expressão gênica no tratamento 1 do que nos demais. Já o gene G11258 se apresenta mais expresso no tratamento 1 do que os demais. Alguns genes só se diferenciam nos tratamentos 1 e 3: G10690, G11152 e G11171. Há uma grande quantidade de genes que se diferenciam apenas entre os tratamentos 2 e 4, ou entre os tratamentos 3 e 4.

Como para a análise anterior, depois de verificado os resultados do ajuste do modelo, foram validados os resultados dos níveis descritivos do Modelo II versus o *fold change* através do gráfico Vulcão. Observando a Figura 5.4, e considerando as duas formas de análise contida neste gráfico, nota-se que a grande maioria dos genes considerados DE pelo método paramétrico,

continua sendo DE no fold change, mas há alguns genes como, por exemplo, os genes 5817, 2505, 5819, 11497 e 11960 que são falso-negativos.

Tabela 5.5 – Estimativa da média da diferença de intensidade de expressão (na escala $\log_2$) entre os tratamentos para os genes considerados DE – Modelo II

	Tratamento1 x Tratamento2		Tratamento1 x Tratamento3		Tratamento1 x Tratamento4		Tratamento2 x Tratamento3		Tratamento2 x Tratamento4		Tratamento3 x Tratamento4	
	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão
5	,	,	,	,	,	,	8,65	0,47	8,30	0,45	,	,
246	,	,	,	,	-3,11	0,49	,	,	-2,79	0,47	,	,
533	,	,	,	,	-1,65	0,26	,	,	,	,	,	,
580	,	,	,	,	,	,	,	,	-1,72	0,23	,	,
588	,	,	,	,	,	,	,	,	-1,52	0,25	,	,
1444	-6,47	0,43	-6,57	0,42	-6,96	0,43	,	,	,	,	,	,
2174	,	,	,	,	-2,16	0,38	,	,	-2,18	0,37	,	,
2505	,	,	,	,	-1,64	0,25	,	,	-2,04	0,26	-1,44	0,24
2506	,	,	,	,	,	,	,	,	-1,76	0,25	-1,38	0,24
3353	,	,	,	,	-1,96	0,33	,	,	,	,	,	,
3823	,	,	,	,	-1,51	0,26	,	,	-1,72	0,26	,	,
3826	,	,	,	,	-2,06	0,33	,	,	,	,	,	,
4054	,	,	,	,	-2,71	0,40	,	,	,	,	,	,
4101	,	,	,	,	,	,	,	,	-1,55	0,24	,	,
4137	,	,	,	,	,	,	,	,	-1,78	0,26	,	,
4165	,	,	,	,	,	,	,	,	-1,86	0,31	,	,
4166	,	,	,	,	,	,	,	,	-1,22	0,20	,	,
4236	,	,	-3,23	0,46	-3,72	0,48	,	,	,	,	,	,
4349	,	,	,	,	,	,	,	,	,	,	-1,10	0,18
4350	,	,	,	,	,	,	,	,	,	,	-1,37	0,21
4380	,	,	,	,	,	,	,	,	,	,	-1,17	0,20
4651	,	,	,	,	,	,	,	,	-1,64	0,27	,	,
5280	,	,	,	,	-2,19	0,36	,	,	-2,18	0,36	,	,
5376	,	,	,	,	-1,92	0,32	,	,	-1,91	0,31	,	,
5430	,	,	,	,	,	,	,	,	,	,	-1,21	0,21
5483	,	,	,	,	-2,73	0,45	,	,	,	,	,	,
5715	,	,	,	,	,	,	,	,	,	,	-1,35	0,24
5801	,	,	,	,	,	,	,	,	,	,	-1,40	0,19
5811	,	,	,	,	,	,	,	,	,	,	-1,37	0,23
5817	,	,	,	,	,	,	,	,	-1,64	0,18	,	,
5819	,	,	,	,	,	,	,	,	,	,	-1,38	0,21
5891	,	,	-6,73	0,32	-7,84	0,34	,	,	,	,	,	,
5927	,	,	,	,	,	,	,	,	-1,32	0,18	,	,
5932	,	,	,	,	,	,	,	,	,	,	-1,13	0,16
10255	-7,89	0,61	-6,74	0,63	-7,32	0,63	,	,	,	,	,	,
10274	-4,12	0,54	-4,25	0,52	,	,	,	,	,	,	,	,
10690	,	,	-2,11	0,35	,	,	,	,	,	,	,	,
11152	,	,	-3,30	0,46	,	,	,	,	,	,	,	,
11160	-4,51	0,65	-4,85	0,63	-4,58	0,66	,	,	,	,	,	,
11171	,	,	-3,40	0,48	,	,	,	,	,	,	,	,

Tabela 5.5 (continuação) – Estimativa da média da diferença de intensidade de expressão (na escala $\log_2$) entre os tratamentos para os genes considerados DE – Modelo II

	Tratamento1 x Tratamento2		Tratamento1 x Tratamento3		Tratamento1 x Tratamento4		Tratamento2 x Tratamento3		Tratamento2 x Tratamento4		Tratamento3 x Tratamento4	
	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão
11258	7,18	0,71	7,24	0,71	8,60	0,75	,	,	,	,	,	,
11371	,	,	,	,	,	,	,	,	,	,	1,24	0,19
11493	,	,	,	,	,	,	,	,	,	,	1,41	0,20
11497	,	,	,	,	,	,	,	,	1,76	0,25	1,54	0,23
11523	,	,	,	,	,	,	,	,	1,87	0,25	1,83	0,20
11960	,	,	,	,	,	,	,	,	1,47	0,20	1,36	0,18
12344	,	,	,	,	,	,	,	,	,	,	1,77	0,28

Gráfico Vulcão – Modelo II

Figura 5.4 – Gráfico Vulcão para resultados do Modelo II e Fold Change

5.3. Modelo III – Anova com Efeitos Mistos em um Único Modelo

O modelo apresentado a seguir é também um modelo misto como o anterior, porém a análise foi feita em apenas um estágio conforme a expressão (4.36). Neste modelo, a variável lâmina entrou no modelo como efeitos aleatórios por poder apresentar variação sistemática em experimentos com *microarrays*.

A seguir está apresentado o modelo:

Modelo Completo:

$$y_{ijg} = \mu + A_j + T_i + G_g + (AT)_{ji} + (GT)_{gi} + (AG)_{ig} + \nu_{ijg}$$

onde,

$g=1,\dots,12768$ é o número de genes avaliados no experimento

$j=1,\dots,24$ é o número de lâmina ou réplicas do experimento

$i=1,\dots,5$ é o número de tratamentos do experimento

Supondo que Y_{ig} tem uma distribuição Normal, em termos de μ_{g1} , μ_{g2} , μ_{g3} , μ_{g4} e $\mu_{g\text{ref}}$, que são as médias de cada tratamento para um determinado gene, a hipótese de interesse a ser testada para este modelo é dada por:

$$H_0: \mu_{g1} = \mu_{g2} = \mu_{g3} = \mu_{g4} = \mu_{g\text{ref}} \Leftrightarrow H_0: GT_{gi} = 0 \quad \forall i = 1, 2, 3, 4.$$

H_a : existe pelo menos uma diferença entre as médias

Pode-se observar na Tabela 5.6 que houve uma menor quantidade de genes DE entre os tratamentos e que a maior quantidade de genes DE, quando comparado entre os tratamentos, se deu entre os tratamentos 1 e 4. A Tabela 5.7 mostra os resultados dos testes, apontando quais os genes DE que, após a

correção de Bonferroni, obtiveram comparações entre tratamentos com nível descritivo menor do que $6,4 \times 10^{-6}$. Observando as estimativas das diferenças, pode-se notar nos genes G1021 e G10092 que houve diferença entre os tratamentos 3 e 4. Para os demais genes houve diferença entre o tratamento 1 e o tratamento 4.

Observando os resultados da Figura 5.5, pode-se notar que se cruzando os dois critérios, o modelo ajustado e *fold change*, nenhum gene se apresenta DE. Há apenas alguns genes que poderiam ser considerados DE pelo critério do *fold change*, mas não pelo método paramétrico.

O que se espera é que esse modelo (em um único estágio) apresente resultados semelhantes ao modelo anterior (em dois estágios). A pequena quantidade de genes DE encontrada através deste modelo se dá devido a problemas computacionais, dificultando a possibilidade de estimação das diferenças entre os tratamentos. Desta forma, como não é possível chegar a resultados conclusivos via este modelo devido a limitações computacionais (não convergência dos algoritmos numéricos), este modelo não foi considerado na comparação entre os modelos, apresentada na seção 5.5.

Tabela 5.6 – Distribuição do número de comparações consideradas DE –Modelo III

	<i>n</i>	%
Tratamento1 x Tratamento2	1	11,1
Tratamento1 x Tratamento3	1	11,1
Tratamento1 x Tratamento4	5	55,6
Tratamento3 x Tratamento4	2	22,2
Total	9	100,0

Tabela 5.7 - Estimativa da média da diferença de intensidade de expressão (na escala $\log_2$) entre os tratamentos para os genes considerados DE – Modelo III

	Tratamento1 x Tratamento2		Tratamento1 x Tratamento3		Tratamento1 x Tratamento4		Tratamento3 x Tratamento4	
	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão
1021	,	,	,	,	,	,	-1,62	0,28
10086	,	,	,	,	-2,23	0,44	,	,
10092	,	,	,	,	,	,	-1,83	0,39
10094	,	,	,	,	-2,66	0,55	,	,
10255	-7,43	0,87	,	,	-7,40	0,90	,	,
10527	,	,	,	,	-3,17	0,56	,	,
10690	,	,	-2,47	0,45	-2,48	0,49	,	,

Gráfico Vulcão – Modelo III

Figura 5.5 Gráfico Vulcão para resultados do Modelo III e Fold Change

5.4. Modelo IV – Anova com Modelos de Efeitos Fixos

A seguir será aplicado o modelo sugerido por Kerr et al. (2000), e apresentado no capítulo anterior. O modelo é dado por:

$$y_{ijg} = \mu + A_j + T_i + (AT)_{ji} + G_g + (GT)_{gi} + (GA)_{gi} + \varepsilon_{ijg}$$

onde,

$g=1,\dots,12768$ genes avaliados no experimento

$j=1,\dots,24$ é o número de lâmina ou réplicas do experimento

$i=1,\dots,5$ é o número de tratamentos do experimento

Tendo Y_{ig} uma distribuição Normal, em termos de $\mu_{g1}, \mu_{g2}, \mu_{g3}, \mu_{g4}$ e $\mu_{g\text{ ref}}$, que são as médias de cada tratamento para um determinado gene, a hipótese de interesse a ser testada para este modelo é dada por:

$$H_0: \mu_{g1} = \mu_{g2} = \mu_{g3} = \mu_{g4} = \mu_{g\text{ ref}}$$

H_a : pelo menos a média de um dos tratamentos não é igual

Na Tabela 5.8 segue a distribuição da quantidade de genes DE para cada comparação entre tratamentos, considerando o ajuste do modelo IV. A Tabela 5.9 mostra os resultados dos testes apontando quais os genes DE que obtiveram nível descritivo menor do que $4,8 \cdot 10^{-7}$, valor encontrado após a correção de Bonferroni. Nos resultados a seguir só foram apresentados as comparações entre os quatro tratamentos. As comparações realizadas com o tratamento de referência (TR) não foram apresentadas, por não ser este o interesse do estudo.

Tabela 5.8 – Distribuição do número de comparações consideradas DE

	<i>n</i>	%
Tratamento1 x Tratamento2	5	11,9
Tratamento1 x Tratamento3	7	16,7
Tratamento1 x Tratamento4	6	14,3
Tratamento2 x Tratamento3	1	2,4
Tratamento2 x Tratamento4	15	35,7
Tratamento3 x Tratamento4	8	19,0
Total	42	100,0

Tabela 5.9 – Estimativa da média da diferença de intensidade de expressão (na escala log₂) entre os tratamentos para os genes considerados DE – Modelo IV

	Tratamento1 x Tratamento2		Tratamento1 x Tratamento3		Tratamento1 x Tratamento4		Tratamento2 x Tratamento3		Tratamento2 x Tratamento4		Tratamento3 x Tratamento4	
	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão	Média	Erro- padrão
5	,	,	,	,	,	,	8,73	0,47	8,39	0,45	,	,
867	,	,	,	,	,	,	,	,	-1,60	0,24	,	,
1444	-6,81	0,44	-6,78	0,43	-7,29	0,44	,	,	,	,	,	,
2505	,	,	,	,	,	,	,	,	-1,95	0,27	,	,
2506	,	,	,	,	,	,	,	,	-1,73	0,25	,	,
3823	,	,	,	,	,	,	,	,	-1,66	0,28	,	,
4137	,	,	,	,	,	,	,	,	-1,85	0,27	,	,
4349	,	,	,	,	,	,	,	,	,	,	-1,17	0,18
5430	,	,	,	,	,	,	,	,	,	,	-1,30	0,22
5817	,	,	,	,	,	,	,	,	-1,60	0,19	,	,
5891	,	,	-6,63	0,40	-7,23	0,42	,	,	,	,	,	,
5982	,	,	,	,	,	,	,	,	-0,92	0,14	,	,
6652	,	,	-7,22	0,88	-7,58	0,92	,	,	,	,	,	,
7103	,	,	,	,	,	,	,	,	-1,84	0,29	,	,
7322	,	,	,	,	,	,	,	,	,	,	-1,02	0,14
7345	3,50	0,43	3,37	0,40	,	,	,	,	,	,	,	,
7360	,	,	,	,	,	,	,	,	-1,96	0,31	,	,
7364	,	,	,	,	,	,	,	,	-1,69	0,25	,	,
7489	-5,65	0,41	-5,20	0,39	-4,96	0,40	,	,	,	,	,	,
7631	,	,	,	,	,	,	,	,	-1,15	0,17	-1,31	0,16
8667	-4,01	0,54	,	,	-3,85	0,54	,	,	,	,	,	,
11160	,	,	-5,09	0,75	,	,	,	,	,	,	,	,
11258	7,91	0,76	8,12	0,76	9,66	0,80	,	,	,	,	,	,
11371	,	,	,	,	,	,	,	,	,	,	1,23	0,19
11497	,	,	,	,	,	,	,	,	1,84	0,25	1,55	0,23
11523	,	,	,	,	,	,	,	,	1,94	0,26	1,88	0,21
11960	,	,	,	,	,	,	,	,	1,44	0,20	1,32	0,19

A seguir foram validados os resultados dos níveis descritivos do Modelo IV versus o *fold change* através do gráfico Vulcão. Observando a Figura 5.4, e considerando as duas formas de análise contida neste gráfico, nota-se que a maioria dos genes considerados DE pelo método paramétrico via modelo com efeitos fixos, continuam sendo DE. Há alguns genes como, por exemplo, 7631, 5817, 7631, 11258, 11497 e 11523 que apontaram ser DE pelo método paramétrico, mas não são pelo método empírico.

Figura 5.6 – Gráfico Vulcão para resultados do Modelo IV e Fold Change

5.5. Comparação entre os modelos

Nesta parte da análise serão verificadas as semelhanças e diferenças apresentadas ao se comparar os resultados dos três modelos I, II, e IV. Serão apresentados, separadamente, os resultados para cada uma das 6 comparações de interesse que foram realizadas no estudo: tratamento1 x tratamento2; tratamento1 x tratamento3; tratamento1 x tratamento4; tratamento2 x tratamento3; tratamento2 x tratamento4; tratamento3 x tratamento4. Nas tabelas a seguir foram apresentadas as mesmas estimativas para as diferenças apresentadas anteriormente, porém classificadas entre os diferentes modelos.

Para a comparação entre o tratamento 1 e o tratamento 2, pode-se observar na Tabela 5.10 que houve similaridade entre os modelos I, II e IV para apenas 1 gene (G1444). Além deste, um gene foi DE nos modelos I e IV (G7489) e um gene nos modelos II e IV (G11258). Os demais genes apareceram DE apenas em um dos modelos.

Para a comparação entre o Tratamento 1 e o Tratamento 3, também pode-se verificar (Tabela 5.11) que apenas houve similaridade entre os modelos I, II e IV para dois genes (G1444 e G5891). Houve similaridade entre dois modelos para o gene G10690 (Modelos I e II), para o gene G6652 (Modelo I e Modelo IV) e para os genes G11160 e G11258 (Modelos II e IV). Os outros genes apareceram DE apenas em um dos modelos.

Para a comparação entre o tratamento 1 e o tratamento 4 (Tabela 5.12), houve similaridade entre os modelos I, II e IV para três genes (G1444, G5891 e G11258). Similaridade entre dois modelos apareceu para o gene G6652 nos modelos I e IV, e para o gene G10255 nos modelos II e IV. Os outros genes apareceram DE apenas em um dos modelos.

Tabela 5.10 – Genes DE para a comparação Tratamento 1 x Tratamento 2

Genes	Modelo I	Modelo II	Modelo IV
G1444	-7,18	-6,47	-6,81
G5047	4,01	,	,
G6652	-7,49	,	,
G7345	,	,	3,50
G7489	-6,29	,	-5,65
G8667	,	,	-4,01
G10255	,	-7,89	,
G10274	,	-4,12	,
G11160	,	-4,51	,
G11258	,	7,18	7,91

Tabela 5.11 – Genes DE para a comparação Tratamento 1 x Tratamento 3

Genes	Modelo I	Modelo II	Modelo IV
G1444	-7,24	-6,57	-6,78
G4236	,	-3,23	,
G5047	4,70	,	,
G5891	-7,22	-6,73	-6,63
G6652	-7,85	,	-7,22
G7345	,	,	3,37
G7489	,	,	-5,20
G10255	,	-6,74	,
G10274	,	-4,25	,
G10690	-2,61	-2,11	,
G11152	,	-3,30	,
G11160	,	-4,85	-5,09
G11171	,	-3,40	,
G11258	,	7,24	8,12

5.12 – Genes DE para a comparação Tratamento 1 x Tratamento 4

<i>Genes</i>	<i>Modelo I</i>	<i>Modelo II</i>	<i>Modelo IV</i>
G246	,	-3,11	,
G533	,	-1,65	,
G1444	-7,63	-6,96	-7,29
G2174	,	-2,16	,
G2505	,	-1,64	,
G3353	,	-1,96	,
G3823	,	-1,51	,
G3826	,	-2,06	,
G4054	,	-2,71	,
G4236	,	-3,72	,
G5047	4,51	,	,
G5280	,	-2,19	,
G5376	,	-1,92	,
G5483	,	-2,73	,
G5891	-7,92	-7,84	-7,23
G6652	-8,50	,	-7,58
G7030	-2,51	,	,
G7489	,	,	-4,96
G8667	,	,	-3,85
G10255	,	-7,32	,
G11160	,	-4,58	,
G11258	9,13	8,60	9,66

Na comparação entre o tratamento 2 e o tratamento 3 apresentada na Tabela 5.13, o G5 apareceu como DE, mesmo assim, apenas nos modelos II e IV.

Tabela 5.13 – Genes DE para a comparação Tratamento 2 x Tratamento 3

<i>Genes</i>	<i>Modelo I</i>	<i>Modelo II</i>	<i>Modelo IV</i>
G5	,	8,65	8,73

Na comparação entre o Tratamento 2 e o Tratamento 4 (ver Tabela 5.14), nenhum gene foi identificado como DE para mais de dois tratamentos. As similaridades entre dois modelos foram identificadas apenas para os modelos II e IV para os genes G5, G2505, G2506, G3823, G4137, G5817, G11497, G11523 e G11960. Os demais genes apareceram DE apenas em um dos modelos.

Tabela 5.14 – Genes DE para a comparação Tratamento 2 x Tratamento 4

<i>Genes</i>	<i>Modelo I</i>	<i>Modelo II</i>	<i>Modelo IV</i>
G5	,	8,30	8,39
G246	,	-2,79	,
G328	1,54	,	,
G580	,	-1,72	,
G588	,	-1,52	,
G867	,	,	-1,60
G1793	1,59	,	,
G1924	1,55	,	,
G2174	,	-2,18	,
G2505	,	-2,04	-1,95
G2506	,	-1,76	-1,73
G3823	,	-1,72	-1,66
G4101	,	-1,55	,
G4137	,	-1,78	-1,85
G4165	,	-1,86	,
G4166	,	-1,22	,
G4651	,	-1,64	,
G5280	,	-2,18	,
G5376	,	-1,91	,
G5817	,	-1,64	-1,60
G5927	,	-1,32	,
G5982	,	,	-0,92
G6602	1,81	,	,
G7103	,	,	-1,84
G7360	,	,	-1,96
G7364	,	,	-1,69
G7631	,	,	-1,15
G8129	1,68	,	,
G8234	1,74	,	,
G9794	2,07	,	,
G9897	1,96	,	,
G11497	,	1,76	1,84
G11523	,	1,87	1,94
G11570	3,07	,	,
G11960	,	1,47	1,44

Como no caso anterior, para a comparação entre o Tratamento 3 e o Tratamento 4 (Tabela 5.15), nenhum gene foi identificado como DE para mais do que dois tratamentos. As similaridades entre dois modelos foram identificadas apenas para os modelos II e IV para os genes G4349, G5430, G11371, G11497, G11523 e G11960. Os demais genes apareceram DE em apenas um dos modelos.

Tabela 5.15 – Genes DE para a comparação Tratamento 3 x Tratamento 4

Genes	Modelo I	Modelo II	Modelo IV
G2505	,	-1,44	,
G2506	,	-1,38	,
G4349	,	-1,10	-1,17
G4350	,	-1,37	,
G4380	,	-1,17	,
G5430	,	-1,21	-1,30
G5715	,	-1,35	,
G5801	,	-1,40	,
G5811	,	-1,37	,
G5819	,	-1,38	,
G5932	,	-1,13	,
G7322	,	,	-1,02
G7631	,	,	-1,31
G8129	1,55	,	,
G11371	,	1,24	1,23
G11493	,	1,41	,
G11497	,	1,54	1,55
G11523	,	1,83	1,88
G11960	,	1,36	1,32
G12344	,	1,77	,

Comparação empírica de Eficiência

Outra abordagem realizada para comparar os modelos foi avaliar uma medida resumo dos erros-padrão de cada um dos modelos, através da média dos erros-padrão obtidos em todas as comparações (Tabela 5.16). Além da avaliação do erro-padrão, foi feita uma avaliação da eficiência relativa. Vinciotti et al. (2005), apresentam uma medida empírica para avaliar a eficiência relativa entre modelos, baseada na variância estimada para cada comparação. Esta eficiência relativa ER é dada por:

$$ER = \sqrt{\frac{\sum_{g=1}^G \sum_{z_1=1}^{Z_1} VC_{gz_1}}{\sum_{g=1}^G \sum_{z_2=1}^{Z_2} VR_{gz_2}}} \quad (5.1)$$

onde,

VC_{gz_1} : é a variância estimada para z_1 -ésima comparação do modelo de comparação no g -ésimo gene;

VR_{gz_1} : é a variância estimada para z_1 -ésima comparação do modelo de referência no g -ésimo gene;

$g=1, \dots, G$ é o número de genes.

$z_1=1, \dots, Z_1$ é o número de comparações realizadas no modelo de comparação;

$z_2=1, \dots, Z_2$ é o número de comparações realizadas no modelo de referência.

Os modelos neste trabalho não possuíam a mesma quantidade de comparações, pois no Modelo I o número de níveis de tratamentos avaliados é 4, uma vez que os tratamentos de interesse são relativos ao tratamento de referência ($M_{jg} = \log_2 G_{jg} - \log_2 R_{jg}$). Nos demais, há 5 níveis de tratamentos avaliados, uma vez que o tratamento de referência é considerado como mais um nível. Desta forma optou-se por considerar as médias das variabilidades ao invés da soma, como apresentado em (5.2). Sendo assim, a fórmula ajustada é dada por:

$$ER = \sqrt{\frac{(\sum_{g=1}^G \sum_{z_1=1}^{Z_1} VC_{gz_1}) / G * Z_1}{(\sum_{g=1}^G \sum_{z_2=1}^{Z_2} VR_{gz_2}) / G * Z_2}} \quad (5.2)$$

Nestas medidas de eficiência quanto menor for o valor, mais eficiente é o modelo de comparação (numerador) relativamente ao modelo de referência (denominador).

Observando os valores da Tabela 5.16, nota-se que, ao comparar os valores das médias dos erros padrão, o Modelo II é o que apresenta menor variabilidade, sugerindo ser o melhor modelo entre os demais, seguido do

Modelo IV e Modelo I, respectivamente. Pelas comparações relativas avaliadas na Tabela 5.15, também, se observa os mesmos resultados.

Tabela 5.16 – Média de Erro padrão para cada modelo considerando todas as comparações realizadas

<i>Média de Erro padrão para todos os contratos</i>	
Modelo 1	0,7
Modelo 2	0,4
Modelo 4	0,5

Tabela 5.17 – Eficiência relativa entre os modelos

<i>modelo de comparação</i>	<i>modelo de referência</i>		
	<i>Modelo I</i>	<i>Modelo II</i>	<i>Modelo IV</i>
Modelo I	1		
Modelo II	0,79	1	
Modelo IV	0,85	1,07	1

As comparações realizadas entre os resultados dos ajustes dos vários modelos propostos, apesar de serem de natureza descritiva, mostram que as maiores diferenças de resultados ocorreram entre os procedimentos de normalização não paramétrica (Modelo I) e os demais modelos. Houve maior concordância entre os modelos de normalização paramétrica via modelos de efeitos fixos (Modelo IV) e mistos em dois estágios (Modelo II). O modelo com efeitos mistos em dois estágios permitiu que um maior número de genes DE fossem identificados. Além disto, ele sugere, também, ser o melhor modelo por apresentar a menor estimativa da medida de variabilidade entre as comparações.

O modelo com efeitos mistos pode ser considerado um melhor modelo, pois, além de mais preciso, o modelo com efeitos mistos é mais flexível, permitindo a modelagem de covariâncias entre as observações e a inclusão de

novos fatores no modelo para a normalização dos dados, além ter a possibilidade de que os erros sejam modelados com outras distribuições além do normal, como é o caso da distribuição t , obtendo resultados mais robustos em situações com distribuições de caudas pesadas (Lange et al., 1989).

É importante enfatizar que para a finalidade específica de comparação entre os modelos, estratégias mais sofisticadas precisam ser consideradas, como estudos de simulação e estatísticas de ajustes apropriadas ao problema. O critério de Akaike (Searle, 1997), não foi utilizado neste caso, dado que para a construção os modelos de normalização paramétrica, seria necessário um número maior de parâmetros do que nas alternativas não paramétricas, para as quais fica difícil a decisão sobre o número de parâmetros usados na obtenção da conclusão final sobre os genes DE.

Além disso, um critério decisivo na comparação entre os modelos é a análise de diagnóstico de cada um dos ajustes envolvidos. Neste trabalho, considerando o número excessivamente grande de ajustes sob cada modelo adotado (cerca de 12.768 modelos para os vários genes) a análise de diagnóstico foi feita (resultados não apresentados) somente para os genes identificados como DE em algumas das comparações entre tratamentos e considerando os modelos de efeito fixos. Por limitações deste tipo, entre outros motivos, a literatura tem recomendado que os resultados das análises estatísticas de dados de experimentos com *microarrays* sejam usados com cuidado num contexto mais exploratório e indicativo de padrão de expressão do que conclusivo.

Capítulo 6

Considerações Finais

Os experimentos com *microarrays* são bastante poderosos e de grande importância em diversas áreas como a agricultura e área médica. Por ser um experimento complexo e com a possibilidade de ocorrer em diversas fontes de variação, muitas técnicas de análise estatística são estudadas visando melhores resultados com uma conclusão mais eficiente para a identificação de genes DE.

Para um melhor resultado, no início do estudo deve-se investir no bom planejamento do experimento. Algumas abordagens foram apresentadas neste trabalho como os delineamentos em blocos incompletos com referência ou *looping*. Devido à grande importância desta etapa da análise, diversos estudos estão sendo feitos comparando diferentes tipos de delineamentos (ver, por exemplo, Yang & Speed, 2002; Rosa et al., 2005), mas sem conclusões definitivas.

Após definir um bom delineamento, a próxima etapa, que é um dos maiores esforços nestes experimentos, é minimizar a influência que fontes de variações sistemáticas podem ter nos resultados. Com este intuito há diversas alternativas de normalizações dos dados, tanto as não paramétricas, como as paramétricas.

Diversas formas de análise foram avaliadas neste trabalho: os resultados considerando o *fold change* são bastante simples, porém ter como base apenas este método para identificar genes DE pode levar a um grande número de falso-positivos, uma vez que ele não leva em consideração a variabilidade dos dados e seu ponto de corte é escolhido de forma empírica. Desta forma, se faz

necessário utilizar outras abordagens de análise como as apresentadas neste trabalho.

No capítulo de aplicação dos dados foram avaliados três diferentes modelos: modelos com normalizações não paramétricas (Modelo I), modelos mistos em dois estágios (Modelo II) e o modelo com efeitos fixos (Modelo IV). Observando os três modelos avaliados, notou-se que eles apresentam resultados diferentes. Os dois modelos que mais se aproximaram em relação aos resultados para genes DE foram os modelos II e IV, o que utilizou a ANOVA da forma sugerida por Wolfinger et al. (2001) com modelos mistos e o modelo de efeitos fixos sugerido por Kerr et al. (2000), respectivamente. Nestes dois modelos a quantidade de genes DE foi maior do que no Modelo I, além de haver maior similaridade entre os resultados dado que genes comuns foram identificados nestes dois modelos. Considerando os resultados obtidos pelo gráfico Vulcão, observou-se que, em geral, houve sempre algum gene considerado DE pelo método paramétrico, porém com fatores de *fold change* menores que dois em módulo. Sendo assim, alguns genes considerados DE pelo método não paramétrico podem ser considerados falso-negativos.

Pelos resultados empíricos deste trabalho o modelo que conduziu a estimativas mais precisas foi o Modelo II, cujos resultados se apresentam bastante semelhantes ao modelo IV. O modelo com efeitos mistos tem uma série de qualidades como, por exemplo, utilizar de forma mais apropriada as informações presentes em um experimento com *microarrays*. Ele considera como aleatórias as possíveis fontes de variação existentes nestes experimentos e assim leva em consideração as variações presentes nestas fontes (lâmina, spot, etc.). É, também, um modelo, como o modelo de efeitos fixos, que permite ajuste por outras distribuições além do normal, como é o caso da distribuição *t*, na presença de distribuições assimétricas, obtendo resultados mais robustos. Uma vez que não há um interesse específico em comparar os níveis dos fatores considerados como aleatórios este modelo é vantajoso (Cui & Churchill, 2003).

Como exemplo, pode-se citar as lâminas, onde dificilmente haverá interesse em comparar os resultados entre diferentes lâminas.

Nos modelos apresentados, foi considerado que a variável de interesse apresentava uma distribuição Normal. Análises subsequentes a este estudo podem ser feitas considerando outros modelos mais gerais, com outras distribuições como, por exemplo, distribuições assimétricas. Na classe de distribuições mais gerais para as medidas de expressão pode não ser necessário usar transformações logarítmicas dos dados.

Neste trabalho foram apresentadas algumas possibilidades de correções para múltiplos testes. O ajuste FDR apresentado no capítulo 4 leva em consideração as dependências entre os resultados dos testes, porém novas alternativas, que levem em consideração padrões de correlação mais estruturados, como o padrão espacial de distribuição dos genes na lâmina, devem ser pesquisadas.

Outras abordagens de análise, visando melhores resultados nestes experimento, como o controle de falso-positivos e falso-negativos estão ainda na infância e devem ser mais pesquisados. Pensando no problema de escala (a métrica) a ser utilizada para a comparação das expressões em experimentos com *microarrays*, tem surgido a possibilidade de análise através da Teoria da Resposta ao Item. Ainda, para se avaliar o comportamento de conjuntos específicos de genes com expressão diferencial, tem se utilizado a técnica de Modelos de Equações Estruturais como uma alternativa de análise supervisionada.

Devido à grande importância e possibilidade de aplicação, um crescente interesse vem surgindo na literatura sobre os experimentos com *microarrays*. Devido às peculiaridades e dificuldades destes experimentos, as técnicas de análise estatística devem ainda ser bastante exploradas e avaliadas.

Referências Bibliográficas

- Baldi, P. & Long A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17:509-519.
- Bolstad, B.M.; Irizarry, R.A.; Astrand, M. & Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2):185-193.
- Brown, H.K. & Prescott, R.J. (1999). *Applied Mixed Models in Medicine*. John Wiley and Son: Chichester, NY.
- Bueno, M. R. P. O (1997). Projeto genoma humano. *Bioética* 5(2):143-155.
- Bussab, W.O. & Morettin, P.A. (2002). *Estatística Básica*. 5ª ed., Editora Saraiva: São Paulo, SP.
- Chen C.C.; Lai D.; Kao C.Y. & Chen C.N. (2004), Microarray Image Pattern Analysis, *The 17th Chinese Image Processing and Pattern Recognition*, B3-6: Hualien, Taiwan.
- Cleveland, W. & Loader, C. (1996). Rejoinder to Discussion of *Smoothing by Local Regression: Principles and Methods*. In Hardle, W. & Schimek, M. G., editors, *Statistical Theory and Computational Aspects of Smoothing*. 113-120: Springer, New York.
- Chu, G.; Narasimhan, B.; Tibshirani, R. & Tusher, (2003). V.G. SAM: Significance Analysis of Microarrays V1.12: User's guide and technical document, Stanford University.
- Coe, M. (2003). You want Ketchup with your DNA Chips? An Overview of Expression Microarrays. *BioTeach J.* 1:89-93.

Churchill, G.A. (2002) Fundamental of experimental design for cDNA microarrays. *Nature Genetics* 32(2):490-495.

Cristo E. B. (2003). *Métodos estatísticos na análise de experimentos de microarray*. IME/ USP: São Paulo, SP.

Cui, X., & Churchill, G.A. (2003). Statistical Tests for Differential Expression in cDNA Microarray Experiments, *Genome Biology* 4:201.

Dantas, O. D. (2004). *Uma técnica automática baseada em morfologia matemática para medida de sinal em imagens cDNA*. IME/ USP: São Paulo, SP.

Draghici. S., (2003). *Data Analysis Tools for DNA Microarrays*. 2ª ed. Chapman & Hall /CRC Press: Boca Raton, Florida.

Dudoit, S.; Yang, Y.H.; Callow, M.J. & Speed, T.P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sinica* 2:111-139.

Durbin, B.P. & Rocke, D.M. (2004), Variance-stabilizing transformations for two-color microarrays, *Bioinformatics* 20(5):660-667.

Durbin, B.P.; Hardin, J.S.; Hawkins, D.M. & Rocke, D.M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 18(1):S105-S110.

Farah, B. S. (1997). *DNA Segredos e Mistérios*. Ed. Sarvier: São Paulo, SP.

Filho, C.A.M. & Verjovski-Almeida, S. (2000). Genoma Clínico: O impacto da genética genômica na prática médica. *Biotecnologia Ciência e Desenvolvimento* 16:162-167.

Ge, Y., Dudoit, S., & Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *TEST* 12(1):1-44.

Hirata, R.D.C. (2005). *Aplicação da biologia molecular no imunodiagnóstico*, FCF/USP: São Paulo, SP. Disciplina Imonologia Clínica. 2005. Disponível em: <http://www.fcf.usp.br/Ensino/Graduacao/Disciplinas/LinkAula/BMAD%20Imuno%202.pdf>. Acessado em 25/11/2005.

Hartemink, A.J.; Gifford, D.K.; Jaakkola T.S. & Young R.A. (2001). Maximum Likelihood Estimation of Optimal Scaling Factors for Expression Array Normalization. *SPIE BIOS*: San Jose, California.

Holder, D.; Raubertas, R.F.; Pikounis, V.B.; Svetnik, V. & Soper, K. (2001). *Statistical analysis of high density oligonucleotide arrays: a Safer approach*. Proceedings of ASA Annual Meeting, Atlanta, GA.

Huber W. A.; Von Heydebreck & M. Vingron. (2003). *Handbook of Statistical Genetics*, 2nd ed. John Wiley and Son: Chichester, NY.

Irizarry, R. A. (2002). *Choosing Smoothness Parameters for Smoothing Splines by Minimizing an Estimate of Risk*. Johns Hopkins, School of Hygiene & Public Health Disponível em: <http://www.biostat.jhsph.edu/~ririzarr/papers/react-splines.pdf> . Acessado em 22/09/2005.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431(7011):931-945.

Irizarry, R.A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y.D.; Antonellis, K.J.; Scherf, U. & Speed, T.P. (2003). Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* 4(2):249-264.

Kendzierski, C.; Irizarry, R.A.; Chen, K.; Haag, J.D. & Gould, M.N. (2004). *To Pool or Not to Pool: A Question of Microarray Experimental Design*. Collection of Bioeststatistics Research Archive. Disponível em <http://www.bepress.com/jhubiostat/paper46/>. Acessado em 13/10/2005.

Kerr, M.; Martin, M. & Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* 7:819-837.

Kerr, M. & Churchill, G.A. (2001a), Experimental design for gene expression microarrays. *Biostatistics* 2:183-201.

Kerr, M. & Churchill, G.A. (2001b). Statistical Design and the Analysis of Gene Expression Microarray Data. *Genetical Research* 77:123-128.

Kerr, M.; Afshari, C. A.; Bennett L.; Bushel P.; Martinez, J.; Walker, N. & Churchill, G. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* 12:203-217.

Landgrebe, J.; Bretz, F., & Brunner, E. (2004). Efficient two-sample designs for microarray experiments with biological replications. *In Silico Biology* 4(4):461-470.

Lange K. L.; Little R. J. A.; & Taylor J. M. G (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896.

Lee, P.S. & Lee, K. H. (2000). *Genomic Analysis. Current Opinion in Biotechnology* 11:171-175.

Leon, J.M. (2002). *Mapeamento genético em populações humanas via modelos de Variância*. Dissertação de Mestrado. São Paulo: Instituto de Matemática e Estatística, Universidade de São Paulo.

Li C. & Wong W.H.(2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* 2(8):1-11. Disponível em <http://www.biomedcentral.com/content/pdf/gb-2001-2-8-research0032.pdf>. Acessado em 28/09/2005.

Long, A.D.; Mangalam, H.J.; Chan, B.Y.; Toller, L.; Hatfield, G.W. & Baldi, P. (2001). Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. *J Biol Chem.* 8, 276(23):19937-44.

Lönnstedt, I. & Speed, T. (2002). Replicated Microarray Data. *Statistica Sinica* 12: 31-46.

Magalhães, M.N. e Lima, A.C.P. (2000) .*Noções de Probabilidade e Estatística*, 2ª ed. IME/USP: São Paulo,SP.

Mendenhall, W. (1968) Introduction to linear models and the design and analysis of experiments. *Duxbury Press*. Wadsworth, Belmont, USA, 1968. 1 ed. 465 p.

Neter, J.; Kutner, M. H.; Nachtsheim, C. J. & Wasserman, W. (1996). *Applied Linear Statistical Models*. 4ª ed. U.S.A.: *Times Mirror Higher Education Group*. 1408 p.

Paula, G. A. (2004). *Modelos de regressão com apoio computacional*. IME-USP: São Paulo, SP.

Park, T.; Yi, S.; Kang, S.; Lee, S; Lee, Y. & Simon. (2003). Evaluation of normalization methods for microarray data. *Bioinformatics* 4:33.

Ramakrishnan, R.; Dorris, D.; Lublinsky, A.; Nguyen, A.; Domanus, M.; Prokhorova, A.; Gieser, L.; Touma, E.; Lockner, R.; Tata, M.; Zhu, X.; Patterson, M.; Shippy, R.; Sendera, T. J.; Mazumder, A. (2002). An assessment of Motorola CodeLink™ microarray performance for gene expression profiling applications. *Nucleic Acid Research* **30**, n^{os} 7 e 30.

Rosa, F.H.P. (2004), *Métodos Adaptivos em regressão não paramétrica e normalização de microarrays de cDNA*. IME/USP: São Paulo, SP.

Rosa, G. J. M., Steibel, J. P. & Tempelman, R. J. (2005). Reassessing design and analysis of two-color microarray experiments using mixed effects models. *Comparative and Functional Genomics* **6**:123-131.

Rosa, G.J.M. & Tempelman, R.J. (2003). *Delineamento e Análise de Experimentos de Expressão Gênica*. 10^o Simpósio de Estatística Aplicada à Experimentação Agronômica – Universidade Federal de Lavras: Minas Gerais.

Santos, J. M. ; Silva, R. M. ; Domingues, P.; Amado, F. & Santos, M. A. S. (2001). Genómica funcional em Aveiro. *Boletim Biotecnológico* **68** :13-18. Universidade de Aveiro, Portugal.

Satagopan, J.M. & Panageas, K.S. (2003). A statistic perspective on gene expression data analysis. *Statistics in Medicine* **22**:481-499.

Schuchhardt, J.; Beule, D.; Malik, A.; Wolski, E.; Eickhoff, H.; Lehrach, H. & Herzelt, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Research* **28**:10-47.

Searle, S.R (1997). *Linear Models*. Wiley-Interscience: 532p

Shu, W. & Wang, Y. (2005). *Optimal design for gene expression microarrays*, Technique report, Institute of Statistics, National Tsing Hua University.

Simas, H.S. (2003). *Aspectos Metodológicos em análise de series temporais epidemiológicas do efeito da poluição atmosférica na saúde pública: uma revisão bibliográfica e um estudo comparativo via simulação*. Dissertação de Mestrado. Instituto de Medicina Social: Universidade do Rio de Janeiro, RJ.

Simpson, A.J.G. e Caballero, O. L. (2000). Projeto Genoma Humano e suas implicações para a saúde humana: visão geral e contribuição brasileira para o projeto. *Revista de Bioética e Ética* 8(1):89-96.

Smyth, G. K.; Yang, Y.H. & Speed, T. P. (2003). Statistical issues in cDNA microarray data analysis. In: *Functional Genomics: Methods and Protocols*, M. J. Brownstein and A. B. Khodursky (eds.), *Methods in Molecular Biology* 224: 111-136. Humana Press: Totowa, NJ.

Soler, J.M.P.; Rosa, F.F.; Chiavegatto, S.; Aneas, I. & Krieger, J.E. (2004). Use of splines for normalization of microarray gene expression data. *Bioscience Journal, Especial*: 101-116.

Speed, T. (2003). *Statistic analysis of gene expression microarray data*. Chapman & Hall/ CRC.

Speed, T. (2000). Hints & Prejudices - Always Log Spot Intensities and Ratios, *Technical Report*, University of California, Bekerly. Disponível em www.stat.berkeley.edu/~terry/zarray/Html/log.html. Acessado em 12/07/2005.

Storey, J.D. (2002). A direct approach to the false discovery rate. *Journal of the Royal Statistical Society, Series B*, 64:479-498.

Tusher, V.G.; Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98:5116-5121.

Vinciotti V. ; Khanin R.; D'Alimonte D.; Liu X. ; Cattini N.; Hotchkiss G.; Bucca G.; Jesus O. de; Rasaiyaah J.; Smith C. P.; Kellam P. & Wit E. An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics* 21(4):492-501.

Wolfinger, R.D.; Gibson, G.; Wolfinger, E.D.; Bennett, L.; Hamadeh, H.; Bushel, P.; Afshari, C. & Paules, R.S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Comput Biol.*8(6):625-637.

Woo, Y.; Affourtit, J.; Daigle, S.; Viale, A.; Johnson, K.; Naggert, J.; Churchill, G. (2004). A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms, *J. Biomol Tech.* 15(4), 276-284.

Yang Y.H.; Dudoit, S.; Luu P. & Speed T. (2001). *Normalization for cDNA microarray data*. SPIE BIOS vol. 4266: San Jose, CA.

Yang, Y.H. & Speed, T.P. (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics* 3:579-588.

Yang Y.W.; Dudoit, S.; Luu, P.; Lin, D.M.; Peng, V.; Ngai, J.; & Speed T. P. (2002b). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* 30 (4) e15.

Yang H., Buckley M. J., Dudoit S. & Speed T.P. (2002c). Comparison of methods for image analysis on cDNA microarray data. *Computacional Graphical Statistics*, 11 (1): 108-136.

Yang X. (2003). *Optimal Design of Single Factor cDNA Microarray Experiments. and Mixed Models for Gene Expression Data*. Dissertation for degree of Doctor in Statistics to the Faculty of the Varginia Polytechnic Institute and State University, Virginia.

Zatz, M., Clonagem e Células-Tronco. 2004. Disponível em: http://genoma.ib.usp.br/artigos_clonagem.php. Acessado em 22/09/2005.

Wu, S. & Yan, H. (2003). Microarray Image Processing Based on Clustering and Morphological Analysis. Proceedings of the First Asia-Pacific bioinformatics conference on *Bioinformatics* 111-118 (Australian Computer Society Inc.).

www.affimetrix.com

www.axon.com/gn_Genomics.html

www.fapesp.br

www.genome.gov/10001477 (Understanding Our Genetic Inheritance: The United States Human Genome Project - The First Five Years: Fiscal Years 1991-1995. Springfield, VA: National Technical Information Service, April 1990. (<http://www.genome.gov/10001477>- acessado em 27/07/2005)

www.ghente.org/ciencia/genoma

Apêndice

Apêndice 1 – Exemplo de banco de dados para os 40 primeiros genes (spots) avaliados em um dos 24 animais do estudo

<i>gene</i>	<i>animal</i>	<i>Intencidade G (Cy3)</i>	<i>Correção BackGround para a intensidade G</i>	<i>Intencidade R (Cy5)</i>	<i>Correção BackGround para a intensidade R</i>	<i>Flags</i>
g1	1	527	392	875	660	-50
g2	1	485	423	782	766	-50
g3	1	533	411	830	797	-50
g4	1	376	407	756	618	-50
g5	1	524	429	546	507	-50
g6	1	314	470	794	560	-50
g7	1	324	435	361	592	-50
g8	1	407	396	542	647	-50
g9	1	541	339	643	741	-50
g10	1	388	349	1074	774	-50
g11	1	243	452	900	750	-50
g12	1	365	442	684	556	-50
g13	1	336	349	493	518	-50
g14	1	450	436	931	589	-50
g15	1	635	421	515	677	-50
g16	1	462	404	700	771	-50
g17	1	436	430	381	629	-50
g18	1	431	417	589	568	-50
g19	1	395	385	517	615	-50
g20	1	443	447	585	647	-50
g21	1	563	381	639	529	-50
g22	1	827	374	1308	733	-50
g23	1	684	404	1402	918	-50
g24	1	1501	475	2203	989	0
g25	1	3589	491	5110	642	0
g26	1	723	371	1364	497	-50
g27	1	610	318	744	550	-50
g28	1	701	319	1339	542	0
g29	1	1300	657	1128	842	0
g30	1	634	429	1074	715	-50
g31	1	1722	674	1734	889	0
g32	1	713	391	1334	667	-50
g33	1	1070	481	1904	498	0
g34	1	1326	445	1836	634	0
g35	1	1424	383	1950	556	0
g36	1	599	375	593	612	-50
g37	1	568	399	694	560	-50
g38	1	578	411	791	565	-50
g39	1	478	393	621	492	-50
g40	1	501	364	635	478	-50

Apêndice 2 – Códigos SAS para os programas utilizados no estudo

Modelos I – Modelo com normalização não paramétrica

```
/** Ajuste por Loess */
proc sort data=dados;
  by lamina ;
run;

proc loess data=dados;
  by lamina;
  model m = a / smooth=0.4;
  id ident;
  ods output OutputStatistics=resultados;
run;

/** Gráficos MAPLOT */

proc sort data=resultados;
  by tratamento a ;
run;

options reset=all;
symbol1 v=dot i=none c=black;
symbol2 v=none i=join c=blue;
symbol3 v=none i=none c=red;
proc gplot data= resultados;
  by tratamento;
  title1 'MA-PLOT';
  format DepVar 5.0 a 5.0;
  plot DepVar*a=1 pred*a=2 DepVar*a=3/ haxis=0 to 20 by 5 vaxis=-10 to
  10 vref=0 overlay;
run;

proc sort data=resultados;
  by lâmina tratamento gene ;
run;

ods listing close;
run;

/** Comparação entre os tratamentos */

proc mixed data= resultados;
  by gene;
  class tratamento;
  model norm = tratamento / outp=resteste;
  lsmeans tratamento / diff cl;
  ods output covparms= resteste_c tests3= resteste_t diffs=resteste_d;
run;
```

Modelos II – Modelo com efeitos mistos

```

/**** normalização ****/

proc mixed data=s2 covtest cl lognote;
  class lamina tratamento;
  model logi = tratamento /
    outp=resteste(keep=lamina gene spot tratamento resid);
  random lamina lamina*tratamento;
  lsmeans tratamento / diff cl;
run;

data resteste;
  set tese_f2.resteste;
  where resid not in (.);
run;

proc sort data=resteste;
  by gene lamina;
run;

/**** modelo genético****/

ods listing close;
run;

proc mixed data=resteste;
  by gene;
  class lamina tratamento;
  model resid = tratamento / outp=resteste_r;
  random lamina;
  lsmeans tratamento / diff;
  ods output covparms=resteste_c tests3= resteste_t diffs= resteste_d;
run;

ods listing;
run;

proc sort data= resteste_d;
  by gene;
run;

/*---Bonferroni adjustment---*/
data resteste_d;
  set resteste_d;
  if (probt = .) then delete;
  logp = log10(probt);
  neglogp = -logp - log10(69247);
  if (neglogp <= 0) then delete;
run;

```


Modelos III - Modelo com efeitos fixos

```

/**** normalização ****/

proc mixed data=s2 covtest cl lognote;
  class lamina tratamento;
  model logi = tratamento lamina lamina*tratamento/
    outp= resteste(keep=lamina gene spot tratamento resid);
  lsmeans tratamento / diff cl;
run;

Data resteste;
  set resteste;
  where resid not in (.);
run;

proc sort data= resteste;
  by gene lamina spot;
run;

/**** modelo genético****/

ods listing close;
run;

proc mixed data= resteste;
  by gene;
  class lamina tratamento;
  model resid = tratamento lamina/ outp=tese_f4.resteste_r;
  lsmeans tratamento / diff;
  ods output covparms= resteste_c tests3= resteste_t diffs=resteste_d;
run;

ods listing;
run;

data resteste_d;
  set resteste_d;
  if (probt = .) then delete;
run;

/*---Bonferroni adjustment---*/
data resteste_d1;
  set resteste_d;
  if (probt = .) then delete;
  logp = log10(probt);
  neglogp = -logp - log10(126992);
  if (neglogp <= 0) then delete;
run;

```