COPPE/UFRJ

IDENTIFICAÇÃO DAS ALTERAÇÕES NO PERFIL DE EXPRESSÃO GÊNICA DE CÉLULAS HUMANAS TRATADAS COM TALIDOMIDA

Renata Torres de Paiva

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Biomédica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Biomédica.

Orientadores: Flávio Fonseca Nobre
Ulisses Gazos Lopes

Rio de Janeiro

Setembro de 2009

IDENTIFICAÇÃO DAS ALTERAÇÕES NO PERFIL DE EXPRESSÃO GÊNICA DE CÉLULAS HUMANAS TRATADAS COM TALIDOMIDA

Renata Torres de Paiva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA BIOMÉDICA.

Aprovada por:

_____

Prof. Flávio Fonseca Nobre, Ph.D.

_____

Prof. Ulisses Gazos Lopes, D.Sc.

_____

Prof. Jurandir Nadal, D.Sc.

_____

Prof. Milton Ozório de Moraes, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

SETEMBRO DE 2009

Paiva, Renata Torres de

Identificação das Alterações no Perfil de Expressão Gênica de Células Humanas Tratadas com Talidomida/ Renata Torres de Paiva. – Rio de Janeiro: UFRJ/COPPE, 2009.

IX, 85 p.: il.; 29,7 cm.

Orientadores: Flávio Fonseca Nobre

Ulisses Gazos Lopes

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Biomédica, 2009.

Referências Bibliográficas: p. 77-85.

1. Microarranjos. 2. Talidomida. 3. ANOVA. I. Nobre, Flávio Fonseca, *et al*. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Titulo.

Agradecimentos

Ao Professor Flavio Fonseca Nobre, pela orientação, dedicação, carinho e amizade.

Ao Professor Ulisses Gazos Lopes, pela importante colaboração neste trabalho.

À Professora Rosimary Terezinha de Almeida, pelo carinho e atenção.

Ao Marcelo Ribeiro Alves, pela atenção, dedicação e amizade inestimáveis.

À minha família, José Roberto e Ana Beatriz, pela compreensão e apoio.

Aos demais amigos da COPPE por tudo.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

IDENTIFICAÇÃO DAS ALTERAÇÕES NO PERFIL DE EXPRESSÃO GÊNICA DE CÉLULAS HUMANAS TRATADAS COM TALIDOMIDA

Renata Torres de Paiva
Setembro/2009

Orientadores: Flavio Fonseca Nobre
Ulisses Gazos Lopes

Programa: Engenharia Biomédica

A Talidomida é um medicamento que apresenta propriedade imunomoduladoras, antiinflamatórias e antiangiogênicas. O modo pelo qual este medicamento exerce seus efeitos ainda não foi elucidado. Este estudo utilizou dados de experimentos de microarranjos de oligonucleotídeos de células mononucleares de sangue periférico (PBMC) tratadas com Talidomida para estudar alterações na expressão gênica e assim formular hipótese sobre como esta droga exerce seus efeitos. Genes diferencialmente expressos foram identificados usando modelagem ANOVA e o método Produto de Ordenação. Os resultados apontaram genes relacionados com resposta inflamatória crônica a estímulos antigênicos, apoptose celular, regulação negativa da secreção de citocina na resposta imune. A intersecção entre os resultados dos dois métodos aplicados identificou os genes TBCC, PTCH1 e ADH7.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

IDENTIFICATION OF THE PROFILE OF DIFFERENTIAL EXPRESSION OF HUMAN CELLS TREATED WITH THALIDOMIDE

Renata Torres de Paiva

September/2009

Advisors: Flavio Fonseca Nobre

Ulisses Gazos Lopes

Department: Biomedical Engineering

Thalidomide is a drug that shows immunomodulatory, anti-inflammatory and antiangiogenic properties. The manner by which this drug exerts its effects has not yet been elucidated. This study used data obtained from DNA oligoarrays having peripheral blood mononuclear cells (PBMC) treated with Thalidomide to study changes in gene expression and thus formulate hypothesis about how this drug exerts its effects. Differentially expressed genes were identified using ANOVA modeling and Rank Product method. The results showed genes related to inflammatory response to chronic antigenic stimulation, cell apoptosis, and negative regulation of secretion of cytokines in immune response. The intersection between the results of two methods identified TBCC, PTCH1 and ADH7.

SUMÁRIO

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

1. INTRODUÇÃO

A talidomida foi introduzida no mercado farmacêutico como sedativo, em 1953, pela companhia alemã Chemie Grunenthal. Foi amplamente comercializada como sedativo não barbitúrico, sem efeito residual, além de anti-emético, efetivo no tratamento de enjôos matinais comuns na gestação. No início da década de 60, a droga foi retirada do mercado devido ao seu efeito teratogênico, que levou ao nascimento de milhares de crianças com malformações congênitas de membros superiores e inferiores, conhecida como focomielia, além de outras malformações sistêmicas [1]. Nos Estados Unidos o medicamento nunca foi liberado, pois a agência reguladora de medicamentos, *Food and Drugs Administration* (FDA), considerou a necessidade de mais estudos sobre a segurança do medicamento, uma vez que havia relatos de neurite periférica e alterações tireoidianas após seu uso [2]. O ressurgimento do fármaco não tardou a ocorrer, quando no final da década de 60, descobriu-se seu potencial na resolução de uma grave reação sistêmica inflamatória que ocorre na hanseníase, a reação tipo II ou eritema nodoso hanseniano [3]. Diversos trabalhos, após esta descoberta demonstraram suas propriedades imunomoduladoras, anti-inflamatórias e antiangiogênicas. Atualmente, a talidomida é o medicamento de escolha no tratamento da reação tipo II, ou eritema nodoso hanseniano (ENH). [4].

Há relatos de seu uso no tratamento da artrite reumatóide, tuberculose crônica, doença de Behcet e doença de Crohn [3]. Por suas propriedades antiangiogênicas [5] é utilizada em pesquisas no tratamento de tumores sólidos [6] e, também, da miocardiopatia pós-infarto [7]. A talidomida se mostra ainda um tratamento importante

no mieloma múltiplo e doença enxerto versus hospedeiro, doenças para as quais tem seu uso aprovado no Brasil, além do eritema nodoso hanseniano [8].

A autorização para dispensação e uso da talidomida no Brasil foi dada pela portaria Portaria SVS/MS Nº 354, de 15 de agosto de 1997, limitando seu uso aos seguintes programas oficiais:

i. Hanseníase – para tratamento da reação hanseniana tipo II ou eritema nodoso.

ii. Doenças Sexualmente Transmissíveis/Síndrome da Imunodeficiência Adquirida (DST/SIDA)– no tratamento de ulcerações graves recorrentes ocasionadas pela SIDA.

iii. Doenças crônico-degenerativas – incluindo lúpus eritematoso sistêmico e doença enxerto *vs* hospedeiro.

Entretanto, apesar de ter seu uso aprovado no tratamento destas doenças e ser utilizada em tratamentos experimentais de dezenas de outras, seus mecanismos moleculares de ação ainda não foram elucidados [4].

Alguns trabalhos apontam que sua utilização leva a uma alteração dos níveis do fator de necrose tumoral (TNF), uma importante citocina que participa de diversos processos biológicos como morte celular programada (apoptose) e resposta imunológica [4]. Há, também, relatos de que a talidomida exerce efeitos inibitórios sobre a interleucina-12 (IL-12), uma citocina cuja principal função é aumentar a secreção de interferon-γ (IFN-γ) pelas células NK (*natural killers*) e pelos linfócitos T auxiliares (CD4). Outros trabalhos apontam uma ação direta da talidomida sobre os linfócitos T auxiliares (T *helper*). Os linfócitos T são uma importante fonte de citocinas, que apresentam em sua superfície celular receptores específicos capazes de reconhecer patógenos. Há dois principais subconjuntos de linfócitos T, que são

distinguidos pela presença de moléculas de superfície denominadas CD4 e CD8. Linfócitos T que expressam $CD4^+$ em sua superfície são conhecidos como T *helper* (Th) e, podem ser subdivididos em Th1 e Th2. A resposta Th1 é pró-inflamatória, responsável por ativar macrófagos e restringir o crescimento de patógenos intracelulares e perpetuar a resposta imune (memória imunológica). O interferon-γ (IFN-γ) é a principal citocina Th1, além da IL-12. Uma resposta Th1 exacerbada pode levar à injúria tecidual. As citocinas Th2 incluem as interleucinas (IL) 4, 5 e 13, que estão associadas à produção de imunoglobulina E (IgE) e à resposta eosinofílica em atopia. A resposta Th2 exerce controle sobre a resposta Th1 através da IL-4, enquanto a resposta Th1 inibe a Th2 através da IL-2 e do IFN-γ, reduzindo a formação de células Th2 [9].

Uma forma de tentar desvendar os mecanismos moleculares de ação da talidomida é estudar as alterações nos padrões de expressão gênica desencadeada por sua utilização [10]. Para isto, pode-se utilizar microarranjos de DNA, que são arranjos de grandes conjuntos de sequências de DNA imobilizados em um substrato sólido. Estes são uma importante ferramenta capaz de analisar milhares de genes simultaneamente, possibilitando a investigação de aspectos cardinais do crescimento e desenvolvimento celular, exploração de mecanismos de doenças, identificação de marcadores de câncer e, também, o desenvolvimento de novos medicamentos [11]. Na indústria farmacêutica uma das aplicações dos microarranjos é identificar o padrão de expressão gênica de células expostas a pequenas moléculas farmacológicas e toxinas, gerando informações sobre o mecanismo de toxicidade e potencial carcinogênico [12].

Vários trabalhos têm sido realizados utilizando microarranjos de DNA para traçar o perfil de alteração gênica induzido por fármacos, seja identificando os

mecanismos de ação de um agente terapêutico, seja na identificação de alvos moleculares para ação de uma droga [13]. KUDOH *et al.*[14] analisaram as alterações de expressão em células cancerosas (MCF-7) expostas ao agente quimioterapêutico anticancerígeno doxorrubicina ou ao promotor tumoral 12-Otetradecanoilforbol-13-acetato (TPA) usando microarranjos. Neste trabalho os autores verificaram que as alterações de expressão são compatíveis com a resposta fenotípica e bioquímica a estes agentes. Desta forma, demonstraram a possibilidade de se obter perfis moleculares de drogas anticancerígenas utilizando microarranjos de DNA e, assim, gerar informações sobre o mecanismo de resistência de drogas, além de sugerir métodos alternativos de tratamento.

A utilização de microarranjos envolve diversas etapas para que a informação resultante do experimento seja confiável. Estas etapas abrangem desde o planejamento e execução experimental até a interpretação adequada dos resultados. A análise dos dados é uma etapa crucial para o sucesso do experimento, havendo diversos métodos disponíveis, sem que nenhum possa ser considerado “padrão-ouro”. As abordagens propostas são bastante variadas e os métodos enfocam alguma característica considerada importante no conjunto de dados. A escolha do método deve considerar o número de réplicas, a variância dos dados, além de outras importantes características experimentais que influenciarão fundamentalmente a estimativa dos parâmetros amostrais [15].

Existem muitas estratégias para identificar a expressão diferencial de um determinado gene, dentre as quais a razão de expressão (*fold-change,* FC), que foi a primeira abordagem empregada. Nesta, os genes são considerados diferencialmente expressos quando a razão entre as intensidades de controle e tratamento ficam além (ou aquém) de um limiar fixo de corte, por exemplo, dois. Esta técnica é

estatisticamente ineficiente, principalmente porque não considera as inúmeras variações sistemáticas e biológicas que ocorrem neste tipo de experimento. Embora algumas variações sistemáticas possam ser removidas por técnicas de normalização, as variações biológicas entre amostras e as variações fisiológicas, por exemplo, são mais difíceis de manusear [16].

Abordagens mais sofisticadas foram sugeridas para experimentos que apresentavam lâminas replicadas, utilizando de testes estatísticos de hipótese, por exemplo, teste t e suas variações, ANOVA, métodos Bayesianos, levando em consideração o tamanho do efeito e a variabilidade. Métodos que empregam modelos ANOVA apresentam a vantagem adicional de levar em consideração as diversas fontes de variabilidade. Ajustar um valor de corte para a expressão diferencial não é simples, pois é preciso equilibrar os falsos positivos (erro tipo I) e os falsos negativos (erro tipo II). Além disto, realizar testes estatísticos de milhares de genes leva à questão dos testes múltiplos de hipóteses. Esta questão é frequentemente abordada utilizando métodos como, por exemplo, a correção de Bonferroni que controla o *family-wise error rate* (FWER), que é a probabilidade de se obter pelo menos um falso positivo dentre todas as hipóteses testadas [17]. Porém, métodos que controlam o FWER são muito conservadores e, nestes experimentos pode ser mais interessante permitir a presença de poucos genes falso positivos. Sendo assim, pode ser mais prático controlar a taxa de falsa descoberta *(false discovery rate,* FDR*),* que corresponde à proporção de falsos positivos entre todas as hipóteses rejeitadas; ou seja, os genes identificados como diferencialmente expressos quando na verdade não o são.

Alguns estudos de microarranjos ainda utilizam um pequeno número de réplicas, o que dificulta a estimativa dos parâmetros amostrais. A fim de contornar este

problema, alguns métodos de análise foram propostos utilizando o ordenamento dos valores de expressão associado a alguma estatística não-paramétrica na identificação de genes diferencialmente expressos [18], [19], [20], [21]. Entre estes podemos citar o Produto de Ordenação proposto por BREITLING *et al.* (2004), baseado no cálculo do produto da posição do gene em uma lista ordenada de acordo com os valores de razão de expressão.

1.1. OBJETIVO GERAL

Este trabalho tem como objetivo traçar um perfil de alteração gênica de células mononucleares de sangue periférico (PBMC) de indivíduos saudáveis tratadas com talidomida, utilizando microarranjos de DNA.

1.1.1. OBJETIVO ESPECÍFICO

Identificar genes diferencialmente expressos em microarranjos de DNA empregando os métodos estatísticos: modelo ANOVA e Produto de Ordenação.

Agrupar os genes diferencialmente expressos de acordo com a ontologia gênica, para identificar possíveis alvos moleculares da talidomida.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. FUNDAMENTOS DE BIOLOGIA MOLECULAR

O gene é a unidade constitutiva fundamental de qualquer sistema vivo e pode ser definido como um segmento de DNA (ou RNA, em alguns casos) que codifica a informação necessária para produzir um produto biológico funcional. O produto final é, geralmente, uma proteína, podendo também originar diferentes tipos de RNA. Neste caso, considera-se como gene todas as regiões do DNA que são transcritas e, também, as regiões regulatórias para o processo de transcrição [22].

Uma molécula de DNA consiste de duas longas cadeias polinucleotídicas composta de quatro tipos de subunidades nucleotídicas. Os nucleotídeos são compostos por uma base nitrogenada, um grupo fosfato e uma pentose (ribose ou desoxirribose) (Figura 1). No caso dos nucleotídeos no DNA o açúcar é a desoxirribose ligada a um único grupo fosfato e a base pode ser a adenina (A), citosina (C), guanina (G) ou timina (T). Os nucleotídeos são covalentemente unidos em cadeia através dos açúcares e fosfatos, formando uma estrutura alternante entre pentose e fosfato (Figura 2) [22], [23].

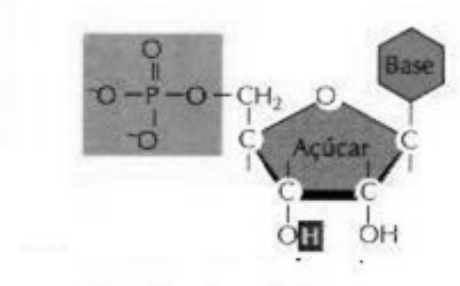

Figura 1 - Representação esquemática de um nucleotídeo, onde se pode obsservar a presença de um açúcar (pentose), um radical fosfato e uma base nitrogenada. Extraído de Lehninger *et al*.

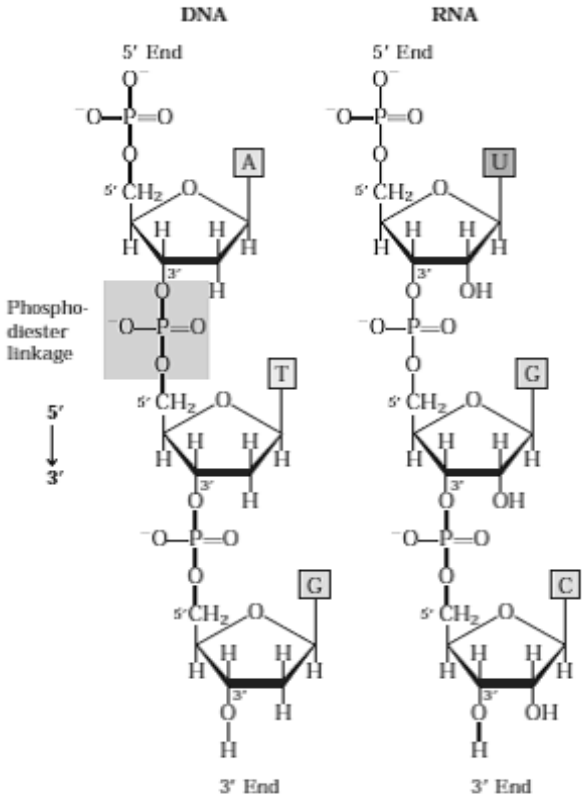

Figura 2: Ligações fosfodiéster ocorrendo o sentido 5'→3' na estrutura covalente do DNA e RNA. Extraído de Lehninger *et al*.

A formulação do modelo da estrutura de dupla hélice do DNA (Figura 3), por James Watson e Francis Crick in 1953, sugere um modelo de como a informação genética pode ser copiada de modo que possa ser transmitida de geração a geração. A compreensão de como a informação no DNA pode ser convertida em proteínas funcionais veio da descoberta do RNA mensageiro e do RNA transportador e com o desvendamento do código genético. Estas, entre outras descobertas, estabelecem o pilar da biologia molecular, englobando os três principais processos na utilização da informação genética pela célula. O primeiro processo é a replicação, onde o DNA é

copiado em duas moléculas filhas com sequências nucleotídicas idênticas. O segundo processo é a transcrição, onde parte da informação genética do DNA é precisamente copiada para o RNA. O terceiro processo é a tradução, na qual a mensagem genética codificada pelo RNA mensageiro é traduzida nos ribossomos em um polinucleotídeo com uma sequência particular de aminoácidos [23].

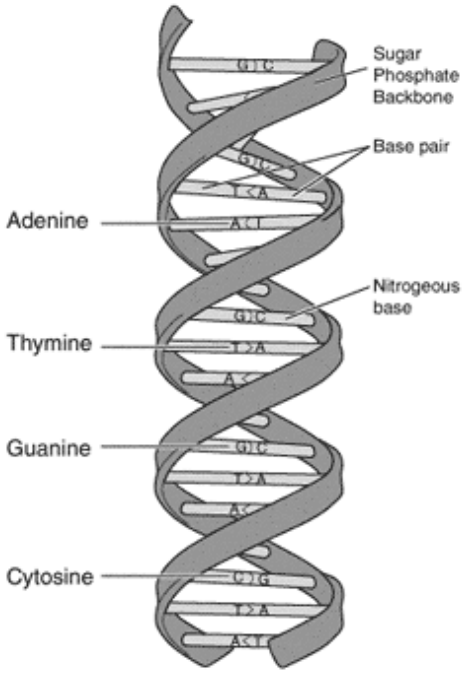

Figura 3: Modelo da estrutura de dupla hélice do DNA proposto por Watson e Crick . Extraído de Lehninger et al.

A molécula de RNA transcrita de um gene é complementar à fita codificadora do gene. Duas fitas únicas complementares de moléculas de ácido nucleico (DNA ou mRNA) tendem a hibridizar, formando uma única molécula de fita dupla. Duas fitas-únicas de moléculas de ácido nucleico que não são totalmente complementares podem também hibridizar, mas quanto maior a complementaridade, mais forte a ligação [23].

Apesar do mesmo DNA genômico estar presente, com algumas poucas exceções específicas, em todas as células de um organismo, nem todas as células são as mesmas. As diferenças entre elas devem-se a diferentes subconjuntos de genes que estão expressos em cada um dos diferentes tipos celulares e em respostas a estímulos, de forma que o padrão de expressão gênica reflete tanto o tipo celular quanto suas condições [22].

Para compreender a função de um gene é preciso compreender seu mecanismo de expressão e sob quais situações estes mecanismo se altera. Até o final da década de 80, a análise da expressão e regulação gênica era realizada utilizando técnicas de análise de genes e proteínas do tipo *dotting blots*, tais como *Northern blot*, *Southern blot*. Estas técnicas consistem de pequenos arranjos de material genético (DNA, RNA ou oligonucleotídeos, respectivamente), que permitem avaliar suas quantidades relativas. No *Southern Blotting*, pequenos fragmentos desnaturados de DNA, digeridos por enzimas de restrição e separados por tamanho por eletroforese em gel, são transferidos para uma membrana, marcados radioativamente e hibridizados. Já no *Northern Blotting*, um oligonucleotídeo marcado por radiação é utilizado na hibridização com RNA mensageiro (RNAm), também previamente separado por gel de eletroforese.

Outra técnica empregada na análise da expressão gênica é a de macroarranjos, utilizados na análise da expressão de dezenas ou centenas de genes. Os macroarranjos são normalmente impressos em membranas de nylon de aproximadamente 8 x 12 cm de tamanho com até 5000 poços, de aproximadamente 300 mícrons de diâmetro. Porém, a restrição da difusão das moléculas ocasionadas pelas membranas de nylon limita o aumento do número de poços impressos e dificulta a hibridização.

A partir da década de 90, foi desenvolvido o microarranjo de DNA, uma técnica de análise de alta varredura, capaz de analisar milhares de genes de uma só vez. Estes arranjos consistem em uma matriz ordenada de milhares de diferentes sequências de DNA que podem ser utilizados para medir a variação de expressão gênica [24]. A principal distinção entre as técnicas de *dotting blots* e os microarranjos de DNA está na utilização de um substrato impermeável, rígido, tal como o vidro, que tem uma série de vantagens práticas sobre membranas porosas e gel, entre elas a facilidade dos ácidos nucleicos alvo em encontrar o acesso imediato às sondas, sem difusão em poros, pois o líquido não penetra a superfície de apoio.

Os níveis de expressão gênica, modulados em resposta a uma variedade de sinais intra e extracelulares, sugerem um "instantâneo" das proteínas que estão sendo expressas. Assim, se o objeto de estudo é, por exemplo, o tratamento de células humanas com uma droga em particular, pode-se observar através da análise de microarranjos, alterações na expressão de genes necessários para a resposta àquela droga [25]. Além disto, a abordagem global, realizada com microarranjos, fornece uma visão geral tanto dos efeitos transcricionais imediatos de uma droga, quanto da resposta secundária às perturbações do metabolismo.

A análise de expressão gênica pode ser realizada usando um esquema de marcação tanto de uma quanto de duas cores. A análise de microarranjos com um marcador é realizada para arranjos preparados por fotolitografia patenteado sob o nome comercial GeneChip pela empresa *Affymetrix* (http://www.affymetrix.com), onde há a síntese de oligonucleotídeo *in situ* em um suporte sólido empregando a fotolitografia para construir cada elemento do arranjo, nucleotídeo a nucleotídeo, acima de 20 pares de base [26]. Neste método, perfis de expressão para cada

amostra são gerados em uma lâmina usando um único marcador fluorescente, tal como ficoeritina, e as diferentes imagens são comparadas. Um protocolo de duas cores foi desenvolvido posteriormente, onde duas amostras de cDNA são marcadas

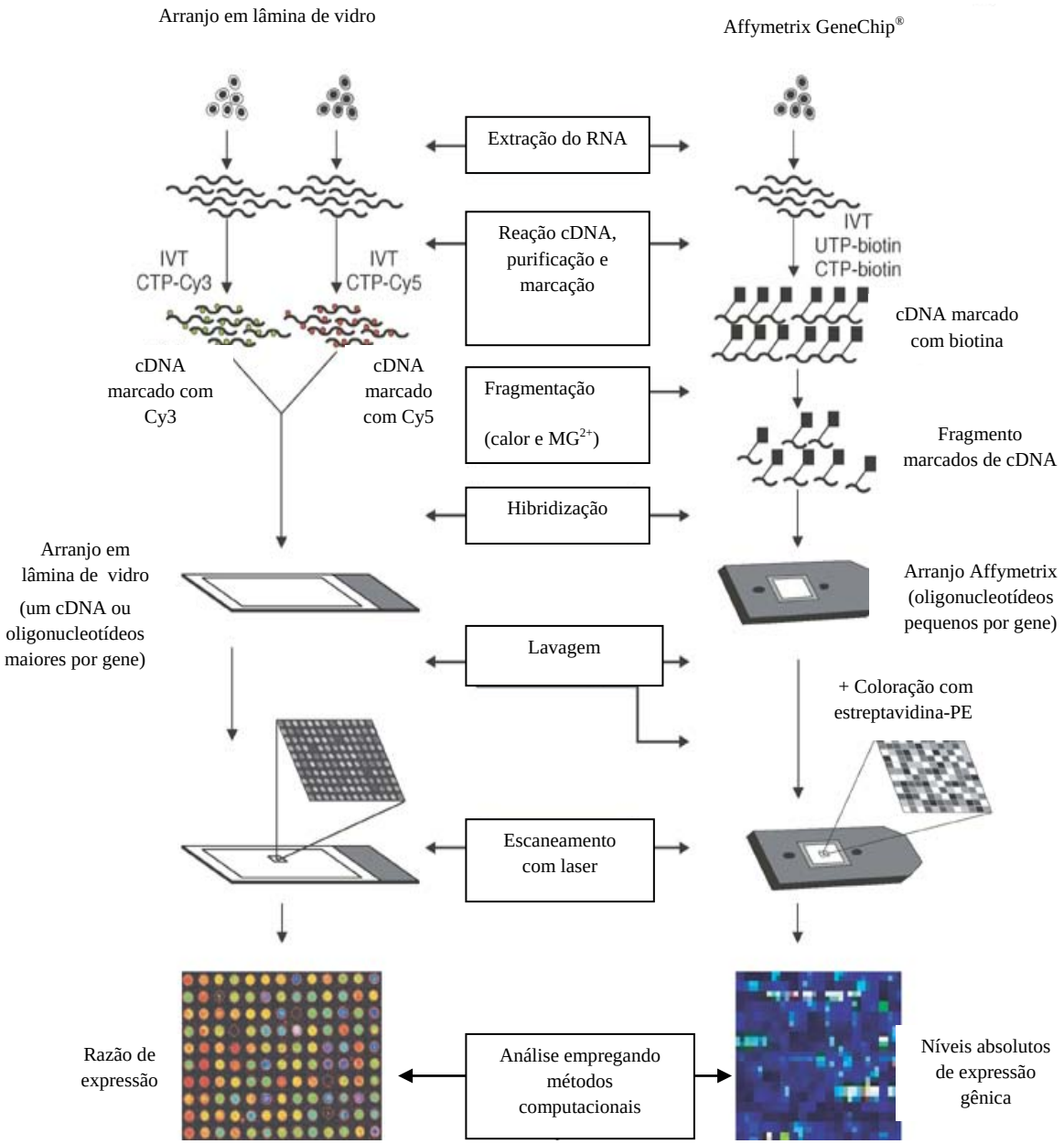

Figura 4: Esquema comparativo da fabricação de microarranjos de DNA e do microarranjo tipo Affymetrix. Modificado de Causton *et al.*

separadamente com diferentes corantes fluorescentes, por exemplo, cianina três (Cy3) e cianina cinco (Cy5). Estas amostras marcadas são hibridizadas em um arranjo impresso de DNA; os microarranjos são escaneados, os sinais fluorescentes podem ser sobrepostos para visualizar os genes que são ativados ou reprimidos [26]. Uma representação das etapas experimentais de cada uma destas tecnologias pode ser vista na Figura 4. Os microarranjos de DNA podem ser fabricados usando oligonucleotídeos curtos (15-25 nucleotídeos), oligonucleotídeos de cadeia longa (50-120 nucleotídeos) ou cDNA amplificado por PCR (100-3000 pares de base).

No cenário da descoberta de drogas e arena terapêutica, os microarranjos vêm sendo utilizados para análise de células e tecidos em diferentes condições, polimorfismos de único nucleotídeo (SNPs), toxigenômica e farmacogenômica [26].

2.2. ETAPAS DO EXPERIMENTO DE MICROARRANJOS DE DNA

O planejamento experimental deve considerar qualidade, custo, disponibilidade de recursos técnicos e precisão requerida nos resultados. Não há com determinar o

desenho correto para um estudo, mas deve haver um equilíbrio entre a questão em estudo e a disponibilidade de recursos.

O desenho de referência, que está entre as diversas possibilidades de desenho de estudo, é o mais simples e mais comumente usado em experimentos de microarranjos [27].

A despeito de ser geralmente menos poderoso que alternativas como o desenho de *loop* [28, 29], [30], [31], o desenho de referência apresenta algumas vantagens como a simplicidade na análise estatística e facilidade de comparação entre diferentes resultados no contexto de uma meta-análise [28].

O desenho de referência consiste na hibridização de cada amostra de teste com um tipo comum de amostra. O termo desenho de referência, porém, é usado para diferentes *layouts* experimentos: KERR e CHURCHILL [29] e GLONEK e SOLOMON [32] descrevem a utilização de um desenho de referência com replicação, onde a amostra de referência incluía a replicação da referência no mesmo nível das amostras de tratamento, STEIBEL e ROSA [33] denominaram este desenho como de referência clássico; DOBBIN e SIMON [28] descrevem a utilização de um desenho de referência onde uma amostra de referência comum é co-hibridizada com cada amostra de teste, sendo este desenho denominado desenho de referência comum (CRD). Em geral, no desenho de referência clássico e no desenho de referência comum o grupo de referência não é de interesse *per se*. Para minimizar este problema, o grupo de referência pode ser representado por um dos tratamentos de interesse, tais como tempo inicial em um experimento serie temporal [31] ou uma cepa *wild type* [34]. STEIBEL e ROSA denominam este desenho como de desenho de referência replicada. A Figura 5 esquematiza os três tipos de desenho: (a) Desenho de referência comum, com quatro replicações (subindices), a referência (R) é a mesma em todos os arranjos; (b) Desenho de referência clássico com quatro replicações (subindices) em 12 arranjos, a amostra de referência é replicada; (c) desenho de referência replicado, no qual seis replicações dos tratamentos B e C são hibridizados juntamente com as replicações independentes do tratamento controle (A). As setas indicam o sentido de marcação, indo do Cy3 (verde) pata o Cy5 (vermelho).

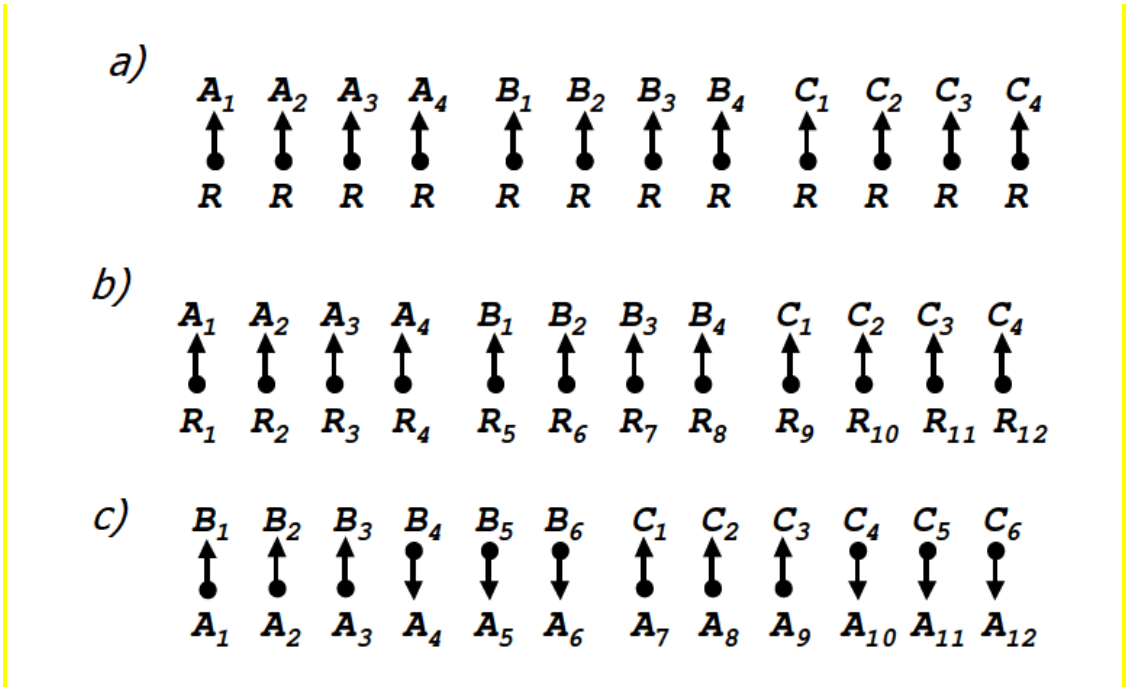

Figura 5: Tipos de desenho de referência para experimentos com três tratamentos (A, B e C). (a) Desenho de referência comum, com apenas uma referência para todos os tratamentos; (b) Desenho de referência, com referencias replicadas no mesmo nível dos tratamentos; (c) desenho de referência replicado, onde o tratamento A é empregado como referência, sendo replicado no mesmo nível dos outros tratamentos. Extraído de STEIBEL e ROSA [33].

Após a determinação da pergunta experimental e do desenho de análise seguem-se as etapas experimentais propriamente ditas, que podem ser divididas em duas partes principais: (i) processamento do material e coleta de dados; e (ii) processamento da informação. Em um experimento de microarranjos, o processamento do material e a coleta de dados podem ainda ser subdivididos em cinco etapas: (1) construção do arranjo; (2) preparação das amostras do material biológico a ser estudado; (3) extração e marcação do RNA (ou cDNA) das amostras, e (4) hibridização das amostras marcadas no arranjo; (5) leitura das lâminas hibridizadas. A imagem escaneada é o ponto inicial para o processamento da informação [35].

O processamento da informação pode ser subdividido em: (1) quantificação da imagem; (2) normalização dos dados e integração; (3) análise das alterações na expressão gênica e (4) a partir das análises geração de novas hipóteses acerca de processos biológicos.

A seguir serão destacados pontos principais de cada uma destas etapas.

2.2.1. PROCESSAMENTO DO MATERIAL E COLETA DE DADOS

Os arranjos são obtidos pela impressão do material genético, como DNA ou oligonucleotídeos pré-sintetizados, em locais específicos em uma lâmina de vidro ([36]). Neste método, pequenas quantidades de material genético são impressos em poços (spots) em uma superfície sólida (lâmina). Esta impressão pode ser realizada de duas formas: (1) utilizando unidades passivas, que aplicam a solução de DNA com uma ponteira (ou ponteiras múltiplas) que toca a superfície sólida (Figura 6) ou (2) unidades ativas, que utilizam dispositivos baseados na tecnologia de jato de tinta, sem contato direto com a superfície. A impressão deve produzir poços com morfologia consistente. Para isto, além da escolha e manuseio mais adequados do equipamento para impressão, esta deve ser realizada em ambiente com temperatura e umidade controladas [37].

2.2.1.1. Preparo da amostra, marcação e hibridização

A preparação do experimento se inicia com a extração do RNA (ou mRNA) das amostras de tecido ou células, o DNA contaminante é digerido por tratamento com DNAses e o material resultante testado quanto à sua pureza. O RNA é transcrito em cDNA, num processo conhecido como transcrição reversa, e marcado, direta ou

indiretamente, com fluorescência. A marcação direta do RNA é obtida pela produção de cDNA a partir do RNA usando a enzima transcriptase reversa e, em seguida, incorporando os marcadores fluorescentes, mais comumente Cy3 e Cy5 [37].

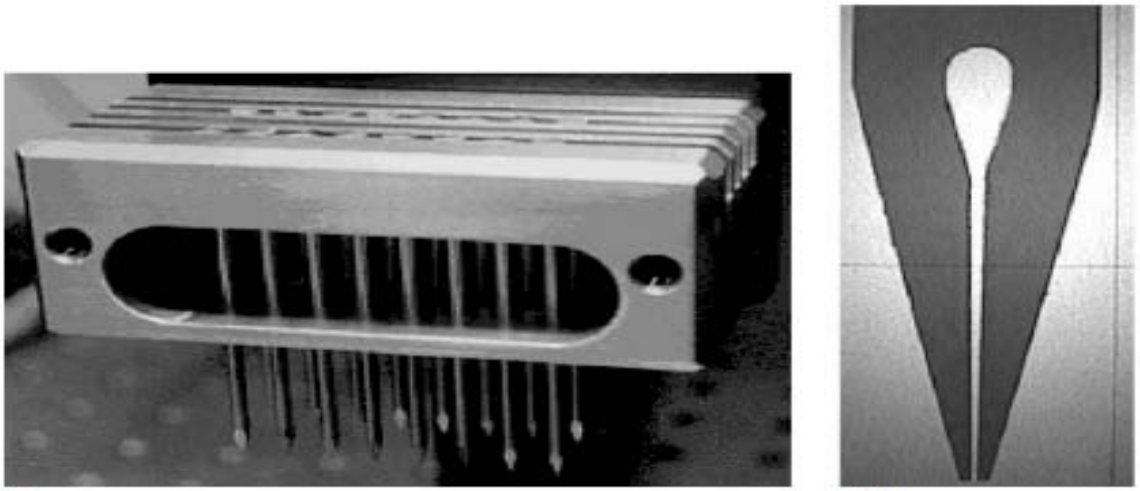

Figura 6: Conjunto de ponteiras de utilizado no robô para impressão dos microarranjos. À direita, representação da ponteira com reservatório para solução de impressão. Extraído de Causton et al [35].

Outros fluoróforos estão disponíveis (por exemplo, TAMRO, Texas vermelho), mas são menos utilizados. No processo de marcação indireta, um grupo reativo, normalmente uma amina primária, é incorporada primeiro ao cDNA, e o marcador é então acoplado em uma reação separada. Este método mostra-se mais eficiente devido à incorporação de uma pequena molécula durante a etapa de transcrição reversa.

Depois de realizada a marcação das sondas, os nucleotídeos não ligados são eliminados. Isto é tipicamente feito em coluna cromatográfica ou por precipitação da amostra em etanol [37]. Em seguida este material marcado é hibridizado na lâmina.

2.2.2. PROCESSAMENTO DA INFORMAÇÃO

2.2.2.1. Aquisição da imagem

A aquisição da imagem é o primeiro passo na análise de dados. Uma má qualidade de imagem exige muitas manipulações, o que influenciará no resultado da análise. Há duas condições para se obter uma imagem de elevada qualidade. A primeira condição é a de que todas as etapas na construção do arranjo, extração do RNA, marcação e hibridação sejam adequadamente padronizadas, para minimizar a variabilidade experimental e manter alta a razão sinal-ruído. A segunda condição depende da escolha adequada dos parâmetros de varredura. O ajuste de ganho do tubo fotomultiplicador (PMT) realizado durante o processo de digitalização deve garantir o equilíbrio global das intensidades entre os dois canais (Cy3 e Cy5). Este equilíbrio pode ser avaliado pela: (i) inspeção visual da imagem de varredura, onde os genes não-diferencialmente expressos aparecem amarelos (ou seja, proporção igual a um) em uma imagem equilibrada; (ii) análise de sobreposição dos histogramas de *pixels* para ambos os canais. Em uma imagem de boa qualidade, os histogramas devm sese sobrepõem tanto quanto possível. E, (iii) cálculo do fator de normalização global para todos os pontos contidos nos dois canais. Pode-se, por exemplo, utilizar como fator a soma dos sinais em um dos canais dividida pela soma de sinais no outro. Uma imagem bem balanceada deve ter um fator próximo de um. A escolha da resolução adequada para varredura depende da especificação do arranjo. A regra geral é que a resolução deve ser de pelo menos 10% do diâmetro do poço. Ao mesmo tempo, o número de poços com *pixels* saturados deve ser mínimo (por exemplo, entre 3 a 5 poços saturados em um arranjo com aproximadamente 6000 elementos).

Também, deve-se evitar a excessiva digitalização da lâmina para impedir o foto-branqueamento [16].

Após o escaneamento da imagem, o reconhecimento dos poços é realizado automaticamente na maioria dos *softwares* de analise de imagem, havendo, em geral, necessidade de pequenos ajustes manuais. Segue-se, então, a etapa de segmentação da imagem escaneada. A segmentação tem por objetivo fragmentar uma imagem, em unidades homogêneas, considerando algumas de suas características intrínsecas como, por exemplo, o nível de cinza dos *pixels*, textura e contraste. Em microarranjos, a segmentação é usada para diferenciar os *pixels* de *foreground* (intensidade de sinal verde e vermelho em cada poço) dos *pixels* de *background* (fluorescência adjacente ao poço de hibridização).

Existem muitos algoritmos para segmentação da imagem incluindo segmentação por círculo fixo, por círculo adaptativo, por forma adaptativa e histograma de segmentação. Existem também vários algoritmos para estimativa do *background*, por exemplo, *background* constante, *background* local e abertura morfológica (*morphological opening*). Esses algoritmos são executados em diferentes programas de análise de imagem. A escolha adequada dos algoritmos depende da qualidade das imagens originais. Por exemplo, a segmentação de círculo adaptativa, que estima o diâmetro separadamente para cada poço, funciona melhor quando todos os poços são circulares [31].

Após o processo de segmentação, as intensidades dos *pixels* dentro da primeira e segunda máscara (ou seja, áreas definidas pelo software como a imagem do *foreground* e *background*, respectivamente), são calculadas separadamente para gerar os valores de intensidade para *foreground* e *background.* A mediana é utilizada

quando há valores extremos nos poços enviesando a distribuição de intensidade dos *pixels*.

Após a determinação dos valores de *background* e *foreground*, a subtração destes valores fornece a intensidade de expressão em cada canal, para cálculo da razão de expressão e outras análises [16]. Como o *background* pode surgir de um número indefinido de fontes, incluindo lavagem incompleta da lâmina após a hibridização, características da lâmina que ligam RNA ou marcador e imprecisão na localização dos poços (segmentação) durante a fase de aquisição de imagem, a subtração de *background* é uma solução para redução do viés [38].

Na etapa de análise, para a utilização de alguns algoritmos podem ser necessárias suposições sobre a estrutura dos dados, como variância homogênea, por exemplo, o que pode produzir variações e incertezas nos resultados. Além disto, o ruído experimental também contribui para a grande variabilidade e aumento da incerteza [39]. Para contornar estes problemas algumas etapas de pré-processamento, como avaliação da qualidade da lâmina através da identificação de poços com sinal de baixa intensidade, por exemplo, podem ser realizadas antes da etapa de inferência estatística.

2.2.2.2. Avaliação da qualidade dos dados

REIMERS *et al.* [40] propuseram método de avaliação da qualidade dos dados através da visualização de vieses regionais, que correspondem às extremidades das lâminas onde a leitura pode apresentar menor intensidade, e outros artefatos espaciais, como arranhões ou poeira depositada, por exemplo. Adicionalmente, um índice estatístico é empregado para quantificar e investigar a distribuição destes vieses regionais. WANG *et al.* [39], considerando a necessidade de utilizar tanto

características de intensidade quanto de informação espacial, propuseram um conjunto de critérios para avaliação da qualidade. Neste trabalho, os autores utilizaram cinco critérios de pontuação para cada poço no arranjo baseados em seu tamanho, relação sinal-ruído (RSR), uniformidade de *background* e condição de saturação, demonstrando que a variabilidade inerente em dados de expressão deve-se em grande parte à qualidade dos poços e que a remoção daqueles poços de baixa qualidade aumenta a confiabilidade das medidas. Um dos grandes problemas destas abordagens é a grande quantidade de poços eliminados por não se mostrarem satisfatórios em relação ao critério de qualidade. Baseado nisto, TRAN *et al.* [41] sugeriram um método que utiliza uma correlação entre média e mediana das intensidades de *pixels* como uma medida de qualidade dos poços a fim de eliminar aqueles de má qualidade, sem necessidade de inspeção visual. Os autores sugerem que uma correlação entre média e mediana acima de 85% seria capaz de não só de manter mais dados, mas também que os dados mantidos apresentariam maior precisão.

Outro critério de avaliação, a razão sinal ruído, quantifica a qualidade do sinal em relação ao ruído sistemático. Desta forma, quanto menor a razão sinal ruído maior a variabilidade das medições e menor a sua precisão [39]. COOMBES *et al.* [42] consideram como poços de boa qualidade aqueles onde a razão sinal ruído é maior que dois.

Entretanto, o rigor dos processos de filtragem pode influenciar (de maneira positiva ou negativa) o resultado da análise. Uma filtragem muito estringente pode eliminar genes cuja alteração na expressão é baixa, porém apresentam relação importante com a questão em estudo. Por outro lado, processos de filtragem pouco

exigentes mantêm um elevado número de genes não informativos, o que aumenta a quantidade de testes de hipótese e influenciam o resultado [43].

2.2.2.3. Transformação logarítmica

Os dados brutos de microarranjos são geralmente transformados tomando-se o logaritmo antes da análise estatística. Transformação é um procedimento matemático, onde uma nova variável é construída ou derivada do original pela aplicação de uma determinada fórmula ou função. A transformação dos dados pode ser usada para obtenção de propriedades estatísticas favoráveis, por exemplo, estabilização da variância e aditividade. A transformação logarítmica (geralmente base dois) pode ser aplicada, também, para facilitar o cálculo da razão de expressão dos sinais de fluorescência originais. A transformação não somente converte as razões em diferenças entre os dois canais para cada poço, mas também estabiliza a variância dos poços de alta intensidade. Para fins estatísticos, a transformação logarítmica converte o erro multiplicativo em erro aditivo. Se os erros são proporcionais a intensidade de sinal na escala original, eles se tornarão constantes através da variação da intensidade de sinal, em escala logarítmica [44].

2.2.2.4. Normalização

Outra etapa do pré-processamento é a normalização. Seu objetivo é minimizar as variações sistemáticas nas medidas dos níveis de expressão gênica de duas amostras co-hibridizadas. Desta forma, as diferenças biológicas podem ser mais bem distinguidas, além de facilitar a comparação dos níveis de expressão entre as lâminas [45]. A normalização visa minimizar ou eliminar a variabilidade introduzida pelas

diferenças de incorporação na etapa de marcação e pela instabilidade dos marcadores fluorescentes, Cy3 e Cy5, além da diferenças nas condições experimentais, que podem influenciar as intensidades de hibridização.

Muitos métodos de normalização têm sido descritos na literatura. Estes métodos podem ser divididos em dois grupos: (1) métodos lineares, que envolvem a estimativa de uma ou mais constantes globais para uma lâmina ou o ajuste de uma curva de regressão local ao logaritmo da intensidade no canal vermelho (R) *versus* logaritmo da intensidade no canal verde (G), e (2) métodos não-lineares, que envolvem a transformação dos dados ao longo dos eixos $\frac{log_2(RG)}{2}$ *versus* $\log_2 R/G$ ajustando uma ou mais curvas de regressão. O método lowess (*Local Weighted Polynomial Regression*), um método não linear, proposto por Cleveland ([46]), para alisamento de diagramas de dispersão robusto e resistente a *outliers*. Ele se baseia em regressões polinomiais locais ponderadas, de forma que pesos maiores são dados a pontos concordantes com o conjunto de dados e pesos menores são dados a *outliers*. O ajuste polinomial é realizado pelo método de mínimos quadrados ponderados, de acordo com os pesos definidos, obtendo valores estimados para a razão logarítmica R/G para cada valor de $\frac{log_2(RG)}{2}$. Os resíduos da subtração dos valores observados pelos estimados são empregados no re-escalonamento dos dados. Quando uma curva é ajustada a todos os pontos na lâmina, o método é denominado normalização por lâmina, quando uma curva é ajustada a cada grupo de impressão determinado pelas ponteiras de impressão, o método é denominado normalização “por ponteiras” [44]. Neste último, ocasionalmente um efeito de extremidades pode ser visto, pois um pequeno conjunto de pontos é tomado por vez.

A etapa seguinte à normalização é a de identificação de genes diferencialmente expressos. A escolha do método de análise não é trivial e vai depender do desenho de estudo, da disponibilidade de replicações da pergunta do estudo. Há uma grande quantidade de métodos de análise disponíveis, cada um enfatizando uma ou mais características do conjunto de dados. A identificação de genes diferencialmente expressos ainda deve, por vezes, considerar a questão da multiplicidade de hipóteses testadas.

3. REVISÃO DA LITERATURA

3.1. TALIDOMIDA

Talidomida é um derivado sintético não-polar do ácido glutâmico. Quimicamente, é uma N-ftalimidoglutarimida consistindo de um único átomo de carbono central assimétrico de ftalidimida à esquerda e um anel de glutarimida à direita. O anel ftalidimida parece estar associado aos efeitos teratogênicos, enquanto o anel de glutarimida, que é estruturalmente semelhante a outros sedativos, parece estar associado à sedação. Ela existe como duas formas oticamente ativas, os enantiômeros R (+) e S (-), que interconvertem rapidamente *in vivo* e também como uma mistura racêmica oticamente inativa. Ela contém quatro ligações amida, sensíveis à clivagem hidrolítica. É pouco solúvel em etanol e água, tornando impossível o emprego de formulações intravenosas, é solúvel em lipídeos, e atravessa facilmente a barreira placentária [47].

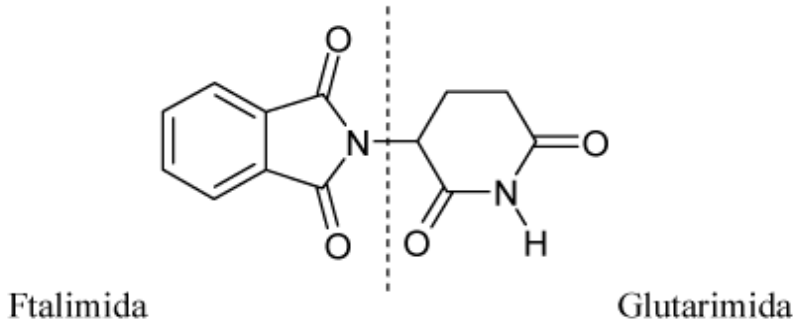

Figura 7: Estrutura da talidomida. Extraída de Borges e Froelich (2001).

Sua variedade de efeitos biológicos está associada à sua capacidade de interferir com o sistema imunológico de acordo com o tipo celular ou via de ativação. SAMPAIO *et al.* [48] demonstraram que a talidomida inibe a produção TNF por monócitos do sangue humano, o que pode estar relacionado a muitos dos efeito imunológicos da droga. O TNF é uma citocina produzida principalmente por macrófagos e monócitos, que desempenha um papel fundamental em muitas doenças autoimunes e inflamatórias crônicas, tais como eritema nodoso hanseniano, doença inflamatória intestinal, doença de Crohn. A redução da atividade desta citocina se tornou um objetivo no tratamento de doenças mediadas pelo TNF [4].

Outra ação da talidomida está relacionada à inibição da angiogênese. D'AMATO *et al.* [5] demonstraram que a talidomida administrada por via oral é um inibidor da angiogênese induzida pelo fator básico de crescimento de fibroblastos (bFGF). O mecanismo pelo qual a talidomida inibe a angiogênese é desconhecido, pois a droga não apresentou efeito sobre a proliferação de células endoteliais induzida por bFGF. A ação antiangiogênica é espécie-dependente e pode ser verificada *in vivo*, mas não *in vitro* o que aponta para uma ação indireta, realizada através de metabólitos ativos produzidos principalmente no fígado [49], [4].

As inúmeras atividades biológicas associadas à talidomida podem ser explicadas, em grande parte, por seus efeitos sobre a atividade do Fator Nuclear kappa B (NfkB) [50]. O NFkB é um fator transcricional envolvido em muitos processos fisiológicos, incluindo a regulação de genes relacionados à inflamação, tais como as citocinas TNF, IL-6, e IL-12, proteínas envolvidas na apoptose, como inibidor celular de apoptose proteína 2 (c-IAP2) e membros da família Bcl-2, além de fatores angiogênicos, por exemplo, o fator endotelial de crescimento vascular (VEGF) [51].

MOREIRA *et al.* [52] examinaram o mecanismo de inibição, pela talidomida, da produção de TNF induzida por lipopolissacarídeo (LPS) e descobriram que a droga aumenta a degradação do mRNA do TNF. Os autores verificaram que a inibição da produção de TNF era seletiva, sem alteração de outras citocinas induzidas por LPS. Os autores ainda indicam que a ação inibitória da talidomida ocorre por mecanismos distintos dos da pentoxifilina e dexametasona, sugerindo que estes agentes atuem como inibidores da síntese de TNF em diferentes pontos da via biossintética. Estas observações fornecem uma explicação para os efeitos sinergéticos destes medicamentos. A inibição seletiva da produção de TNF permite que a talidomida seja usada para o tratamento de condições inflamatórias onde é requerida uma redução do TNF, porém a imunidade deve ser mantida.

STEPHENS *et al.* [53] sugerem um mecanismo estrutural pelo qual a talidomida exerce seus efeitos. Seqüências consenso NFκB são regiões ricas em GC (CG-*box*) e, geralmente, aparecem repetidas em regiões promotoras. A estrutura da talidomida é semelhante a das bases purínicas do DNA (adenina (A) e guanina (G)), o que possibilitaria o silenciamento dos genes ao se intercalar em suas regiões promotoras, bloqueando o acesso de proteínas ao promotor.

HASLETT *et al.* [54] utilizando células mononucleares de sangue periférico (PBMC) de voluntários saudáveis, demonstraram que a talidomida é um potente co-estimulator de células T primárias *in vitro*, cooperando com a estimulação através do complexo receptor de células T para aumentar a proliferação de células T interleucina 2-mediada e a produção de interferon gama. O efeito co-estimulatório é maior sobre as células T CD8 +, do que sobre as células T CD4 +. A droga aumenta também a

resposta citotóxica primária das T CD8+, induzida por células dendríticas alogênicas na ausência de células T CD4 +.

HUANG *et al.* [55] utilizaram microarranjos de DNA para investigar o perfil de expressão gênica de pacientes de espondilite anquilosante em tratamento com talidomida. Neste estudo os autores utilizaram um arranjo de 588 genes para rastrear as alterações na expressão gênica após tratamento com talidomida (dose de 200mg/dia). O experimento contou com um *pool* de células mononucleares de sangue periférico (PBMC) de cinco pacientes antes do tratamento e três meses após a utilização da medicação, para o qual foram utilizados dois arranjos, um pré-tratamento e outro pós-tratamento. Os autores identificaram sete genes reprimidos pela talidomida: TNF, interleucina-1 beta (IL-1b), M1P-1, M1P-2, c-jun, ligante OX40 (ou CD134), um membro da superfamília de receptores TNFR que não é constitutivamente expresso em células T naïve e proteína MAL, associada à maturação de linfócitos T. Entretanto, a inibição destes genes não foi observada durante o tratamento com infliximab, um bloqueador do TNF, indicando a possibilidade de que o foco da terapia com talidomida não seja diretamente sobre o TNF.

VACCA *et al.* [56] utilizaram microarranjos de oligonucleotídeos (Affymetrix) para investigar a ação da s-talidomida (forma opticamente ativa da talidomida) no mieloma múltiplo. Os autores utilizaram arranjos de oligonucleotídeos de genoma humano, contendo 22.283 poços para cerca de 20.000 genes. A *s*-talidomida tem ação comprovada no mieloma múltiplo, embora tanto os efeitos antiangiogênicos quanto pro-apoptóticos, modos primários de ação terapêutica, não estejam esclarecidos. Neste estudo, os autores encontraram alterações na expressão de genes relacionados a apoptose (morte celular programada) e angiogênese, sendo as alterações mais

expressivas observadas em genes apoptóticos. Dentre os genes relacionados a apoptose alterados neste estudo, estão o Bcl-2 e o BAX. O gene Bcl-2 codifica uma proteína inibidora de apoptose. Por outro lado, o gene BAX codifica uma proteína que se dimeriza com a proteína Bcl-2, inibindo sua ação protetora da apoptose. Neste trabalho verificou-se um aumento da razão BAX:Bcl-2, devido à redução na expressão do Bcl-2, sugerindo um possível aumento do efeito citotóxico.

3.2. MÉTODOS DE DETERMINAÇÃO DA EXPRESSÃO DIFERENCIAL

A razão de expressão (*Fold Change*) foi o primeiro método usado para avaliar expressão diferencial e é uma medida razoável do tamanho do efeito. O cálculo da razão de expressão é realizado dividindo-se o valor de expressão das condições de interesse (tratamento) pelo valor da condição de controle. Este tipo de medida é de fácil interpretação e, quando associada a testes estatísticos, é capaz de fornecer informação sobre ordenamento dos resultados [57]. Um dos pontos críticos deste método é a determinação do limiar de corte mais adequado, acima (ou abaixo) dos quais os genes são considerados diferencialmente expressos. Também há o problema de, por não ser um método estatístico, não permitir a incorporação da variância e nem estimativas de confiança.

A fim de contornar a questão da inferência estatística, foi proposto o método de razão incomum (*unusual* ratio), baseado na seleção dos genes que estão a certa distância (mais ou menos dois desvios padrões, por exemplo) da média do experimento, chamada razão controle. Isto é realizado por uma transformação para a

variável padronizada z, que consiste na subtração dos valores log-transformados de intensidade do valor médio (μ) e divisão pelo desvio padrão (δ) (Equação 1).

$$Z = \frac{log\frac{R}{G} - \mu_{R/G}}{\sigma_{log}{}_{R/G}} \qquad \text{Eq. 1}$$

A seleção de genes diferencialmente expressos pode ser ainda realizada com base em testes estatísticos univariados, utilizando, por exemplo, a estatística t definida na Equação 2.

$$t = \frac{M_t - M_c}{s_p\sqrt{\frac{1}{n_t} + \frac{1}{n_c}}} \qquad \text{Eq. 2}$$

O erro da variância (denominador do teste *t*) é de difícil estimativa e está sujeito a flutuações erráticas quando os tamanhos amostrais são pequenos. Estimativas mais estáveis podem ser obtidas pela combinação dos dados de todos os genes, mas este procedimento está sujeito a erro quando a presunção de homogeneidade da variância é violada. Em virtude disto, versões modificadas do teste *t* são propostas como um ponto médio entre poder da análise e redução do viés [58], [59]. Na modificação do teste t proposta por TUSHER *et al.* [60], uma pequena constante positiva é adicionada ao denominador do teste *t* gene específico, removendo o problema da estabilidade da variância. Este teste é empregado no método denominado análise de significância de microarranjos (SAM), sendo a versão do teste *t* denominada teste *S* [60], que utiliza a adição de uma pequena constante positiva ao teste *t* gene específico, desta forma genes com razões de expressão menores não serão selecionados como significativos.

O teste *t* regularizado proposto por BALDI e LONG [61], combina a média ponderada das estimativas específicas de cada gene e da variância média global para

obter o denominador para o teste *t* gene-específico, substituindo a estimativa usual da variância por um estimador Bayesiano baseado em uma distribuição hierárquica a priori. LONNSTEDT e SPEED [62] propuseram a estatística B, que é uma razão de chances pós-logarítmica da expressão diferencial versus a expressão não diferencial. Esta é uma abordagem Bayesiana empírica que considera a variância específica de cada gene e também combina informação através dos genes. O que unifica estas abordagens é a utilização da informação entre genes ao invés de apenas a informação dos genes individualmente [59].

Outra abordagem em análise de dados de microarranjos se dá através do uso de modelos lineares com o método de análise de variância (ANOVA). KERR *et al.* [63] propuseram um modelo ANOVA de efeitos fixos para as intensidades log-transformadas incluindo os termos arranjo (A), marcador (D), tratamento (V) e gene (G), para os efeitos principais, e as interações entre estes. A idéia básica da ANOVA é construir um modelo que considere as diferentes fontes de variação que afetam as medidas e então usar os dados para estimar a variância individual de cada variável incluída no modelo. Neste trabalho, os autores consideram a variância comum entre os genes, aumentando assim os graus de liberdade para estimar o erro da variância, porém, ao acomodar as interações ocorre uma considerável perda de graus de liberdade. Para tentar superar o problema de assumir a variância comum, WOLFINGER *et al.* [34] propuseram um modelo misto, realizando uma análise gene-a-gene e tratando fatores como arranjo, marcador e/ou poço como aleatórios. A principal vantagem da ANOVA é que cada fonte de variabilidade é considerada, por isto é possível distinguir o efeito de interesse, como tratamento, dos efeitos secundários, como a diferença na incorporação dos corantes, por exemplo. O modelo ANOVA proposto primeiramente por KERR *et al* [63] e modificado por WOLFINGER *et al* [34]

foi empregado em um pacote para o programa/linguagem R, o pacote MAANOVA, que pode ser obtido gratuitamente através de repositórios do programa. Recentemente, o pacote MAANOVA foi utilizado por VIDAVALUR *et al.* [64] na investigação das alterações transcricionais desencadeadas pelo uso do Sildenafil, um inibidor da fosfodiestearase-5 em modelo murino de isquemia. Neste trabalho, os autores empregaram como método de normalização empregado o *lowess* e os limiares de significância para os valores *p*, corrigidos para falsa descoberta, de 0,005. Aproximadamente, 156 genes foram identificados como significativamente regulados demonstrando razão de expressão > 1,5.

Todos os métodos estatísticos citados dependem do número de replicações para melhorar o poder de análise. A fim de contornar o problema relacionado à falta de replicações, alguns autores propuseram métodos de análise que, em geral, utilizam o ordenamento dos valores de expressão associado a alguma estatística não-paramétrica na identificação de genes diferencialmente expressos [18], [19], [20], [21]. Entre estes podemos citar o produto de ordenação proposto por BREITLING *et al.* [21]. Este método baseia-se no cálculo do produto da posição dos genes em uma lista ordenada de acordo com os valores de razão de expressão, sendo definido como a probabilidade combinada de um gene estar no topo da lista em todas as replicações dado que ele não é diferencialmente expresso. Esta abordagem assume a independência dos genes e pode gerar subestimativas quando as medidas de expressão de muitos genes do topo da lista são dependentes, por exemplo, devido à hibridização cruzada [65].

ABERCROMBIE *et al.* [66] utilizaram o produto de ordenação na caracterização da resposta transcricional da planta *Arabidopsis thaliana* ao estresse por arsênico (As(V)). Neste trabalho foram utilizadas três replicações biológicas (três para controle,

três para tratamento), com uma lâmina de marcação reversa para cada grupo. Os dados foram normalizados e log-transformados (base 2) antes da análise estatística e o ajuste de intensidade foi realizado utilizando o método *lowess*. Os resultados foram confirmados utilizando qRT-PCR e ensaio de atividade Superóxido dismutase (SOD).

3.2.1. AJUSTE DE TESTES MÚLTIPLOS

Microarranjos envolvem o problema de testes múltiplos, ou seja, testam-se muitas hipóteses dentro de um único estudo, o que representa um grande desafio estatístico. Ou seja, ao testar 10.000 transcritos, empregando um nível de significância alfa de 0,05, normalmente, poderão ser produzidos 500 falsos positivos (10.000 X 0,05 = 500). Numerosos métodos foram sugeridos para contornar este problema, entre eles os métodos que controlam a taxa de erro *family-wise* tipo I (*family-wise error rate - FWER*), como por exemplo, a correção de Bonferroni. Estes métodos limitam a probabilidade de se cometer um ou mais erros tipo I em uma família de hipóteses, que corresponde ao conjunto de hipóteses testadas em um experimento. Entretanto, neste tipo de estudo biológico, onde são utilizados estudos confirmatórios, pode-se, ainda, controlar o número de falsos positivos através do controle da taxa de falsa descoberta (*false discovery rate*, FDR) definida por BENJAMINI e HOCHBERG [67]. A FDR é equivalente, ou relacionada, a muitas outras métricas que quantificam a confiança de que um gene em particular é diferencialmente expresso e pode ser controlada, em um determinado nível, ou estimada. A estimativa da FDR (*p*FDR) pode ser interpretada como a probabilidade de que um gene declarado como diferencialmente expresso seja um falso positivo [57].

3.2.3. TESTES ENTRE CLASSES E GENES (GENE-CLASS TESTING)

Identificar listas de genes diferencialmente expressos é uma etapa importante da análise, porém, há uma necessidade crescente de realizar análises mais elaboradas do ponto de vista biológico. Ao invés de enumerar genes que mostram alterações de expressão significativas é importante demonstrar como estes genes interagem como parte de complexos, vias ou redes, identificando as alterações na expressão gênica de relevância biológica coordenada em conjuntos gênicos que agem de modo sinérgico dentro de vias [68].

Uma forma de abordar esta questão é categorizar um conjunto selecionado de genes baseado no conhecimento de grupos funcionalmente relacionados, realizando testes entre classes e genes. Tais grupos podem ser definidos de diversas formas, entre elas a anotação de ontologia genética (*Gene Ontology*, GO), que fornecer pistas sobre questões biológicas relacionadas a estes genes. Identificar um pequeno número de classes gênicas relevantes entre um grande conjunto de classes candidatas requer ferramentas estatísticas [68]. Diversos métodos empregam testes estatísticos para comparar o número de genes significantes em uma classe, com o número esperado sob uma hipótese nula particular. Entre os métodos disponíveis estão o *Gene Set Enrichment Analysis* (GSEA) e o teste hipergeométrico, disponível no pacote *GOstats* para o programa R , ambas as abordagens apresentam características em comum: (1) ambos assumem grupos de genes definidos *a priori* (*gene ontology*, por exemplo), (2) consideram como universo gênico relevante, obtido, geralmente, a partir de métodos de filtragem e por fim, (3) ambos consideram a anotação de categorias particulares na

análise dos dados de microarranjos. A diferença entre elas está no fato de que o GSEA não requer uma separação rígida entre os genes que são diferencialmente expressos e os que não são. Desta forma, o GSEA utiliza os valores p, determinados a partir de um teste estatístico como o teste *t* ou teste *F* e verifica se estes genes estão associados com conjuntos gênicos de interesse.

O teste hipergeométrico, assim como o GSEA, é uma técnica empregada na analise de perfis moleculares dos dados de expressão *gênica*. Esta técnica é utilizada para classificar genes em uma determinada categoria, *Gene Ontology* (GO), por exemplo, e baseia-se no conceito de que genes fenotipicamente relacionados tendem se apresentar agrupados [69]. O teste hipergeométrico é realizado da seguinte forma: define-se o universo gênico, usando dados de microarranjos e se decide quais genes no universo são significativamente diferencialmente expressos e quais não são, e para uma ontologia específica no GO, identifica quais genes estão anotados para aquela ontologia e quais não estão. A seguir, usa um teste hipergeométrico para verificar se há mais sobreposições de entre os genes DE e a anotação GO do que o que seria esperado por chance.

Este capítulo apresentou uma revisão da literatura sobre o problema abordado neste trabalho, a talidomida e suas propriedades e de alguns dos métodos disponíveis para análise de dados de microarranjos, entre estes o modelo ANOVA e o produto de ordenação. O modelo ANOVA, apresenta características interessantes do ponto de vista da análise de microarranjos. Entre elas, está a possibilidade de separar as diversas fontes de variabilidade nestes experimentos, como os efeitos do marcador, da lâmina, do tecido, dos poços, dos genes bem como suas interações. Além disto, o modelo ANOVA é adequado para análise de mais de duas condições experimentais. Entretanto, sua utilização é condicionada pelo desenho experimental e número de

réplicas, que devem ser estabelecidos durante o planejamento experimental. Um método mais flexível, que permite maior liberdade tanto no desenho quanto no número de réplicas é o Produto de Ordenação. Este método requer poucas suposições sobre os dados, permitindo desenhos experimentais diversos, além de ser eficaz em experimentos com pouca ou nenhuma replicação. Apesar de não separar as fontes de variabilidade, este método apresenta a vantagem de poder ser empregado em experimentos realizados em plataformas diferentes. Por serem o modelo ANOVA e Produto de Ordenação os métodos empregados nas análises estes serão abordados mais detalhadamente na sessão seguinte.

4. BASES METODOLÓGICAS

4.1. MODELO DE ANÁLISE DE VARIÂNCIA (ANOVA)

O modelo de análise de variância (ANOVA) é usado para detectar a diferenças entre médias de três ou mais populações quando as variâncias são iguais. Este se baseia na decomposição da variância total de uma determinada resposta (variável dependente) em duas partes, que podem ser atribuídas ao (1) modelo de regressão e aos (2) resíduos (erros). A magnitude numérica destas variâncias é comparada através do teste de Fisher (teste *F*).

A análise de experimentos de microarranjo através de técnicas de Análise de Variância (ANOVA) permite a estimativa dos efeitos experimentais pelo particionamento da variabilidade experimental. Em geral, um experimento de microarranjos apresenta vários fatores que contribuem para a variabilidade sistemática: efeito do arranjo (A), do marcador (D), do tratamento (V) e do gene (G), acrescido das interações da interação entre estes fatores. Cada um desses efeitos principais e suas interações podem ser contabilizados em um modelo ANOVA.

Os dados são, geralmente, transformados para a escala logarítmica a fim de permitir a utilização de um modelo linear, ao invés de um multiplicativo. O teste F é usado para detectar qualquer padrão de expressão diferencial entre as diversas condições ao comparar a variação entre as amostras replicadas dentro e entre as condições [58].

Denominando *y* os valores de intensidade, pode-se escrever o modelo ANOVA para microarranjos como:

$$\log y_{ijkgr} = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \epsilon_{ijkg} \quad \text{Eq. 3}$$

onde, os termos do modelo (μ, A, D e AD) englobam a média de intensidade global (μ) e as variações em relação a arranjos (A), marcadores (D) e interação arranjo-marcador (AD). O termo G consiste na média de intensidade associada a um gene particular. VG são as variações dos níveis de expressão dos genes. O termo DG representa os efeitos gene-específicos dos marcadores e o termo AG a variação do gene em cada arranjo. Quando o modelo ANOVA é ajustado, obtêm-se estimativas para cada um dos termos individuais. De particular interesse, neste contexto, são os valores estimados de VG que são os valores relativos de expressão (nível de expressão de um gene g numa amostra k relativa a expressão média deste gene em todas as amostras do experimento) e denotamos seus valores estimados como $\overline{VG}$.

WOLFINGER *et al.* [34] propuseram o ajuste consecutivo de modelos de ANOVA para normalização (modelo do arranjo) e para a identificação de genes diferencialmente expressos (modelo gênico). No modelo de normalização,

$$Y_{ijkgr} = \mu + A_i + D_j + AD_{ij} \quad \text{Eq. 4}$$

onde, μ captura a média global enquanto os termos A e D capturam os efeitos de arranjo e reação de marcação, respectivamente. Os resíduos deste modelo (r_{ijkr}) são utilizados como entrada para o modelo gênico. Assim, o restante do modelo é ajustado iterativamente gene-a-gene. Para cada gene g, é ajustado o modelo:

$$r_{ijkr} = G + VG_k + DG_j + AG_i + \epsilon_{ijkr} \quad \text{Eq. 5}$$

onde, G captura o efeito médio do gene, AG a variação do arranjo com relação ao gene, DG a variação do gene com relação ao marcador, e VG a variação do gene com relação às condições experimentais.

Em um experimento de microarranjos típico, são ensaiados os mesmo genes em todos os arranjos, então o fator gene (G) é dito equilibrado com relação aos fatores arranjo (A) e marcador (D), consequentemente, as estimativas obtidas pelo ajuste do modelo em dois estágios são iguais às obtidas pelo ajuste do modelo em um estágio.

Os parâmetros de ajuste do modelo ANOVA são estimados e sujeitos a incerteza e erro. Por isto, métodos de inferência são necessários para a interpretação adequada dos resultados.

A inferência estatística requer a especificação de um modelo para os dados e, para o teste de hipótese, são necessários dois modelos: um nulo e outro alternativo. Considerando que cada amostra de RNA está associada a um tratamento (ou variedade), sob a hipótese nula (H_0), de não expressão diferencial, todas as amostras serão consideradas uma única variedade. No modelo alternativo são consideradas duas ou mais variedades.

Métodos de inferência para modelo ANOVA tradicionalmente consideram a normalidade do erro ou utilizam a teoria de grandes amostras para estabelecer os limites de confiança e/ou os valores *p* usando distribuições tabuladas tais como χ*2, t* ou *F*. Porém, dados de microarranjos não seguem uma distribuição de erro normal, nem possuem grande tamanho amostral (considerando gene-a-gene). Por isto, métodos de randomização como permutação e *Bootstrap*, são utilizados para estabelecer os níveis de significância para os testes estatísticos e intervalos de confiança. O modelo ANOVA, empregado no pacote MAANOVA para o programa R, para análise de microarranjos utiliza a permutação dos resíduos do ajuste do modelo

ANOVA para a hipótese nula, enquanto os intervalos de confiança são determinados pelo método *Bootstrap*. Ambos os procedimentos podem ser aplicados globalmente ou restrito a cada gene. O procedimento não-restrito assume a variância comum do erro, diferentemente do procedimento restrito que não requer esta presunção sendo, contudo, aplicável apenas a experimentos com grande número de replicações. O método ainda oferece uma terceira opção de permutação: por arranjo. Neste caso, os arranjos que apresentam a mesma variedade são permutados livremente. Este método não faz suposições sobre a distribuição dos dados, mas não é adequado a pequenas amostras. O pacote MAANOVA apresenta a possibilidade de obter os valores *p* tabulados, além dos valores obtidos por permutação.

4.1.1. TESTE DE HIPÓTESE

O MAANOVA disponibiliza quatro estatísticas de teste (F_1, F_2, F_3 e F_s). Todos os testes são baseados na soma quadrática dos resíduos gene-específicos, denominados rss_g e os graus de liberdade dos resíduos, denominados de *df*. Os testes de hipótese envolvem a comparação de dois modelos e os testes estatísticos são calculados gene-a-gene. Consideramos rss_0 e df_0 a notação para o modelo nulo e rss_1 e df_1, para o modelo alternativo.

A estatística F1 corresponde à estatística F tradicional, onde são utilizados os dados de um único gene.

$$F_1 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{rss_1/df_1} \qquad \text{Eq. 6}$$

Este teste corresponde a uma generalização do teste *t* [70] e não requer o erro comum da variância. Porém, tem baixo poder estatístico devido ao pequeno tamanho amostral e pode ser sensível a variações nas estimativas do resíduo da variância, rss_1.

O teste F3 assume o erro comum da variância, o teste estatístico é:

$$F_3 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{s^2_{pool}} \qquad \text{Eq. 7}$$

Onde, $s^2_{pool} = \frac{1}{N}\sum_{g=1}^{N} rss_{1g}/df_1$, é a estimativa da variância comum. O teste F3 é poderoso e pode ser utilizado para pequenas amostras. Porém, a igualdade das variâncias deve ser testada e caso seja necessário, uma transformação estabilizadora da variância deve ser aplicada aos dados antes da utilização do teste.

O teste F2 é um híbrido dos dois testes anteriores. O denominador F2 usa estimativa da variância gene específica que é um encolhimento em direção à variância média global,

$$F_2 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{\left(s^2_{pool} + rss_1/df_1\right)/2} \qquad \text{Eq. 8}$$

Este teste se aplica bem a experimentos independentemente replicados.

O teste Fs proposto por CUI *et al* [71] utiliza no denominador um estimador de encolhimento, para estimar os componentes da variância para cada gene. O estimador de encolhimento não faz nenhuma afirmação sobre a distribuição da variância através os genes. A estatística Fs se comporta bem tanto quando a variância é realmente constante ou quando varia através dos genes.

$$F_s = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{\tilde{s}_g^2} \quad \text{Eq. 9}$$

onde:

$$\tilde{s}_g^2 = \left(\prod_{g=1}^{G} \left(X_g/v\right)^{1/G}\right) B \times exp\left[\left(1 - \frac{(G-3)V}{\Sigma(\ln X_g - \overline{\ln X_g})^2}\right)_+ \times (\ln X_g - \ln X_g)\right] \quad \text{Eq. 10}$$

O estimador $\tilde{s}_g^2$, é baseado no estimador de encolhimento de James-Stein ($\hat{\theta}_{JS}$) que é dado por:

$$\hat{\theta}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|y\|^2}\right) y \quad \text{Eq. 11}$$

4.2. PRODUTO DE ORDENAÇÃO (RP)

O Produto de Ordenação [21] é baseado na razão de expressão, uma abordagem biológica onde são comparados níveis de mRNA sob diferentes as condições (A e B) em uma lâmina. Com apenas uma lâmina, é possível identificar genes induzidos ou reprimidos em cada uma das condições, entretanto, estes dados são muito ruidosos e pouco confiáveis. O método utiliza o conceito de que probabilidade aleatória (p) de um gene específico ocupar uma posição mais extrema (r) em uma lista de n genes ordenados por valores de razão de expressão é igual a $p = \frac{r}{n}$, em um experimento com *k* replicações a probabilidade (P) de um gene estar no topo de uma lista é exatamente

$$P_g = \frac{1}{n^k}$$

se as listas são independentes e inteiramente aleatórias. Então, para cada gene g, em k replicações, pode-se calcular o produto de ordenação:

$$RP_g^{up} = \prod_{i=1}^{k} \frac{r_{g,i}^{up}}{n} \qquad \text{Eq. 12}$$

onde r_g^{up} é a posição do gene g na *i*-ésima replicação ordenada por valores decrescentes de razão de expressão. De modo análogo, r_g^{down} é calculado a partir da lista de genes ordenados por valores crescentes de razão de expressão. Quanto menor o valor RP, menor a probabilidade de que a posição do gene no topo da lista seja por chance. A Figura 8 montra um desenho esquemático as posições possíveis de um gene em diferentes replicações.

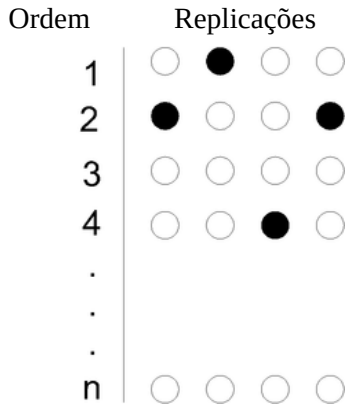

Figura 8: Representação do método produto de ordenação.

A distribuição dos valores RP pode ser obtida por permutação onde, para um determinado número de experimentos aleatórios, simulados e com o mesmo número

de replicações (k) e genes (g) que o experimento real, são realizadas permutações aleatórias dos números $1,\ldots,n$ e, para estes, os valores RP são calculados como descrito na equação 10. O valor *p* corresponde à determinação da probabilidade de se observar um determinado valor RP, ou melhor, em um experimento aleatório e pode ser determinado pela contagem do número de simulações em que os valores RP, para um dado gene, menores ou iguais ao obtido no experimento dividido pelo número de experimentos simulados aleatórios.

O método produto de ordenação está disponível como pacote para R e pode ser obtido através do Bioconductor. O Bioconductor é um projeto que fornece ferramentas para análise e compreensão de dados genômicos [72].

4.3. TESTE ENTRE CLASSE E GENE

A crescente de quantidade de dados em genômica tem impulsionado o desenvolvimento de ferramentas para auxiliar na tarefa de representação e processamento de informações sobre genes, seus produtos e suas funções. Uma destas ferramentas é o Gene Ontology (GO), desenvolvido em paralelo com o trabalho em uma variedade de bases de dados de bioinformática.

A ontologia gênica é, segundo o Consórcio Gene Ontology que a desenvolveu, um vocabulário controlado. A ontologia pode, neste caso, ser entendida como uma terminologia nos quais axiomas e definições estão associadas de forma a permitir utilização análise de computacionais de suas relações. O foco principal da ontologia gênica é direcionado para estabelecer uma ferramenta útil que permite o estabelecimento de estruturas que permitam se manter a par das anotações biológicas que são aplicadas aos produtos de genes.

A análise funcional baseada na associação dos termos de ontologia gênica em uma lista de genes selecionados pode ser considerada então como uma ferramenta bastante informativa na análise de expressão gênica, permitindo identificar a significancia biológica dos resultados [73]. Há diferentes pacotes de análise que se propõe a realizar esta tarefa, entre eles o pacote *GOstats* disponível no Bioconductor, proposto por FALCON e GENTLEMAN [73]. Neste pacote é implementado um teste condicional que utiliza o teste hipergeométrico para analisar as relações hierárquicas entre os termos GO e uma lista de genes diferencialmente expressos, semelhante ao apresentado por ALEXA *et al.* [74].

Um experimento hipergeométrico é um experimento estatístico onde uma amostra de tamanho *M* é selecionada aleatoriamente, sem reposição de uma população de *N* elementos. Na população, *r* itens podem ser classificados como sucessos e *n – r* itens podem ser classificados como fracassos. Considerando um experimento estatístico, onde se tem uma urna contendo 10 bolas, das quais cinco são brancas e cinco pretas a seleção aleatória de duas bolas, sem reposição, e a contagem do número de bolas brancas selecionadas representa o experimento hipergeométrico.

A função densidade de probabilidade da distribuição hipergeometrica pode ser então definida por,

$$P(X = r) = \frac{{}^{M}C_{r} \times {}^{N-M}C_{n-r}}{{}^{N}C_{n}} \qquad \text{Eq. 13}$$

Para realizar uma análise usando o teste hipergeométrico é necessário, definir um universo de genes (geralmente definido como o número de bolas em uma urna) e uma lista de genes selecionados como diferencialmente expressos a partir deste universo. Embora seja claro que a lista de selecionados gene determina o resultado da análise, a idéia de que o universo tenha um grande efeito sobre as conclusões é,

talvez, menos óbvia e a correta especificação do universo é de grande importância na análise. A seguir, identifica-se o subconjunto do universo que é considerado de interesse. Este conjunto pode ser construído a partir de um teste estatístico ou a qualquer outro método para determinação de genes diferencialmente expressos. No modelo hipergeométrico, cada termo é tratado como um independente na classificação. Cada gene é classificado de acordo com sua seleção ou não, numa determinada classe. A probabilidade hipergeométrica é calculada para avaliar se o número de genes selecionados associado ao termo é maior do que esperado por chance [73].

No pacote *GOstats*, a função *hyperGTest* fornece uma implementação para o cálculo da distribuição hipergeométrica de super ou sub-representação de termos em uma lista de genes especificada. Este cálculo ignora a estrutura dos termos GO e trata cada termo como independente de todos os outros termos. Muitas vezes, esta análise de associação gera uma considerável sobreposição de genes, pois cada termo GO herda todas as anotações de seus antecedentes mais próximos. Para minimizar este problema o pacote *GOstats* implementa um método que condiciona a significância de uma determinada categoria a significância de todos os termos filhos anotados nela, a um determinado *p*-valor de corte [73].

5. MATERIAIS E MÉTODOS

Os dados dos microarranjos foram obtidos através de uma parceria entre o Instituto de Biofísica Carlos Chagas Filho (UFRJ), o Laboratório de Hanseníase da FIOCRUZ e a Unidade de Microarranjos (UNAM – México). Os experimentos consideram um pool de células PBMC de bolsas de sangue arterial coletadas de três indivíduos saudáveis como referência. Para a amostra de controle, células de mesma origem foram tratadas com DMSO (dimetil sulfóxido), para amostra de tratamento, as células foram tratadas com talidomida diluída em DMSO.

Para o microarranjos, os oligonucleotídeos (50 mer) representantes de 9984 genes humanos envolvidos em processos celulares como metabolismo, morte e crescimento celular, adquiridos da empresa MWG Biotech Company, foram aplicados em duplicata (replicação interna) sobre lâminas de vidro revestidas com superamina (Telechem International) usando o sistema Virtek ChipWriter, resultando em um total de 19968 spots.

Para minimização do viés de intensidade, associado à diferença na incorporação dos marcadores empregados nos microarranjos de cDNA, foi realizada a marcação reversa (*reverse labeling*), na qual as amostras de controle são marcadas com o fluoróforo antes empregado nas amostras de teste, e vice-versa. Foi fabricada uma lâmina por tratamento para esse fim. O esquema de marcação empregado para cada lâmina está representado na Tabela 1, onde cada experimento é composto por três lâminas no total.

Após a pré-hibridização (5x SSC, 0.1% SDS, 1% BSA) por uma hora a 42 °C, as amostras obtidas das células não tratadas (controle) e tratadas com DMSO e

DMSO + Talidomida (teste) foram adicionadas ao tampão de hibridização Hybit 2 (Telechem International) e hibridizadas durante a noite em três arranjos para cada experimento (2 réplicas e um dye-swap) a 42 °C em uma câmara umidificada (Corning). As lâminas hibridizadas foram então lavadas, secas e escaneadas usando o Scan Array 4000 (Packard Biochips). A intensidade de spot para cada arranjo foi obtida pelo software Array Pro Analyzer.

Tabela **1**: Esquema de marcação das lâminas em cada experimento.

	Marcadores	
	Cy3	Cy5
Experimento 1		
H10KA-01-39	Controle (PBMC)	Tratamento (DMSO)
H10KA-02-02	Tratamento (DMSO)	Controle (PBMC)
H10KA-02-06	Controle (PBMC)	Tratamento (DMSO)
Experimento 2		
H10KA-01-40	Controle (PBMC)	Tratamento (Talidomida)
H10KA-02-03	Tratamento (Talidomida)	Controle (PBMC)
H10KA-02-07	Controle (PBMC)	Tratamento (Talidomida)

Os dados utilizados nas análises são referentes ao *background* e ao *foreground* das lâminas escaneadas.

Para dar início à análise dos dados, foi realizada a correção de *background* subtraindo-se os valores de *foreground* dos valores de *background* em todas as lâminas. Esta correção pode gerar valores menores ou iguais a zero e, neste caso, estes valores eram substituídos pelos de sua réplica. A Tabela utilizada para análise foi construída utilizando-se os valores de intensidade (com correção de *background*) de cada lâmina; o identificador dos genes (GeneId); o número do gene, correspondente à sua posição na tabela, variando de 1 a 19968); e número de acesso do gene (Acc), que corresponde ao código de identificação único de uma sequência

polimérica (DNA ou proteína) quando submetido a uma base de dados de sequências; além da informação correspondente à posição de impressão de cada gene na lâmina, meta row, meta column, row e column. A Tabela 2 apresenta um trecho da tabela empregada na análise dos dados.

Tabela **2**: Representação da tabela utilizada na análise de expressão diferencial. Sendo mostradas apenas as lâminas utilizadas na análise.

Num	metarow	metacol	row1	col	Acc	UGCluster	Symbol	R1	G1	R2	G2	R4	G4	R5	G5
1	1	1	1	1	NM_003090	Hs.434901	SNRPA1	1298,04	531,81	599,11	842,71	1155,32	1450,78	548,51	492,65
2	1	1	1	2	NM_003090	Hs.434901	SNRPA1	1143,96	639,32	654,06	953,24	790,26	945,93	508,82	450,02
3	1	1	1	3	NM_018933	Hs.283803	PCDHB13	74,61	155,94	129,06	482,03	51,42	55,94	130,42	125,19
4	1	1	1	4	NM_018933	Hs.283803	PCDHB13	70,03	179,94	99,56	504,32	62,58	64,45	93,53	109,52
5	1	1	1	5	NM_000607	Hs.528635	ORM1	704,29	333,92	286,06	666,52	866,56	1484,94	279,45	321
6	1	1	1	6	NM_000607	Hs.528635	ORM1	578,46	283,53	223,11	616,72	885,61	1256,59	273,44	333,68
7	1	1	1	7	NM_002380	Hs.153647	MATN2	367,68	134,98	131,16	475,82	425,54	469,2	146,36	102,91
8	1	1	1	8	NM_002380	Hs.153647	MATN2	289,48	123,39	123,14	486,07	356,83	395,7	113,4	119,54
9	1	1	1	9	NM_012062	Hs.180628	DNM1L	152,37	275,88	209,4	613,26	166,62	302,21	182,6	256,14
10	1	1	1	10	NM_012062	Hs.180628	DNM1L	149,14	288,55	181,67	638,24	160,95	291,9	179,13	238,05

Acc = número de acesso GenBank; UGCluster = identificador de acesso UniGene; Symbol = símbolo oficial (NCBI).

A etapa seguinte consistiu da avaliação da qualidade dos dados. O critério de avaliação escolhido foi a razão sinal-ruído (RSR), obtida pela razão entre sinal, corrigido para *background*, e *background*. Poços que apresentavam RSR menor que dois foram considerados de baixa qualidade. A partir da avaliação da RSR foram empregadas duas abordagens para análise de dados: na primeira, os poços com baixa qualidade receberam uma marcação na lâmina e continuaram na análise, este conjunto de dados foi denominado conjunto completo; na segunda, se um mesmo gene apresentava mais de seis poços de baixa qualidade considerando as seis lâminas (totalizando 12 poços por gene), este gene era eliminado do experimento. A fim de equilibrar o desenho de análise, foi eliminada uma lâmina de marcação direta

em cada grupo (controle e tratamento). A eliminação foi baseada no critério de qualidade estabelecido anteriormente, sendo eliminadas aquelas lâminas que apresentaram maior número de poços de baixa qualidade.

Para a etapa de inferência estatística, foram utilizadas apenas duas lâminas, para cada grupo, sendo uma de marcação direta e uma de marcação inversa. O desenho experimental, utilizado nesta análise foi o de referência, onde as células mononucleares de sangue periférico sem tratamento foram consideradas como referência [33].

Para estabilização da variância, os dados foram transformados aplicando-se o logaritmo de base 2, e em seguida normalizados pelo método *lowess*, que consiste numa regressão linear localmente ponderada. Este procedimento requer a utilização de um parâmetro “span” que determina quais dados são locais. Como este parâmetro relaciona-se à amplitude, valores muito altos não corrigem adequadamente a curvatura e valores muito pequenos, levam a um super-ajuste [75]. Neste trabalho, foi utilizado 20% como *span*, o que se mostrou adequado na correção da curvatura (Figura 05). Estes dados normalizados também são utilizados como entrada para o método produto de ordenação. A seguir, foi calculada a média dos poços de replicação interna para cada lâmina.

A qualidade do modelo ajustado foi avaliada por gráficos de dispersão e da normalidade dos resíduos. A expressão diferencial foi determinada utilizado o teste Fs. A escolha do teste foi motivada pela sua eficiência quando há informação limitada para se estimar a variância gene-específica, como no caso dos dados deste experimento. Como o teste *Fs* não segue a distribuição F tabulada, seus valores críticos foram determinados por permutação [76], sendo utilizadas 100 permutações.

A visualização dos resultados do teste Fs do MAANOVA é realizada através do gráfico Volcano, um gráfico de espalhamento onde eixo vertical representa o logaritmo decimal negativo do valor p e o eixo horizontal representa as razões de expressão.

A razão de expressão relativa (*fold change*) foi calculada para todos os genes, utilizando os resultados fornecidos pelo pacote MAANOVA, segundo a equação:

$$\textbf{\textit{Razão de Expressão}}\ (FC) = 2^{\log_2 controle - \log_2 tratamento} \qquad \text{Eq. 14}$$

A análise empregando o método produto de ordenação utilizou os dados log-transformados e normalizados pelo método *lowess*. Esta análise, o desenho empregado foi o de referência. Foram selecionados genes diferencialmente expressos aqueles com valor p menor que 0,001.

Após análise de significância das alterações de expressão para os genes no conjunto reduzido e obtenção dos valores p para cada gene foi realizada a testagem entre classes e genes. Para isto foi definido com conjunto de interesse os genes que apresentam classificação de ontologia gênica (*Gene Ontology*) [77]. Os dados de filtragem foram utilizados para definir o conjunto universo e os genes cujo valor *p* foi inferior a 0,05 foram utilizados como conjunto de teste. O método empregado para esta análise foi o teste hipergeométrico, condicional, disponível no pacote GOstats/R [73].

6. RESULTADOS

Os valores de intensidade foram calculados pela subtração do *foreground* do *background*. Em diferentes lâminas, alguns genes apresentaram valores menores ou iguais a zero visto que as intensidades de *foreground* eram menores que *background* e, nestes casos, os valores de suas replicatas foram utilizados. Uma das lâminas de marcação direta de DMSO (lâmina de número 3) apresentou o maior número de poços substituídos, 17 poços. Após a transformação dos dados pelo logaritmo na base dois, foi realizada uma avaliação da qualidade dos dados utilizando a razão sinal ruído, onde 11524 poços (58%) apresentaram sinal de baixa qualidade, de acordo com o critério adotado (sinal/ruído < 2). Para que pudéssemos equilibrar o desenho de estudo, uma lâmina de marcação direta em cada grupo (controle e tratamento) foi eliminada, com base na razão sinal ruído. Ou seja, duas lâminas de marcação direta, uma em cada grupo, que apresentaram maior número de poços com sinal de baixa qualidade foram excluídas nas etapas de análise subsequentes. Ainda, baseado na avaliação de qualidade foram formados dois conjuntos de dados: no primeiro, os genes que apresentaram baixa razão sinal ruído receberam apenas uma marcação, sendo este conjunto denominado conjunto completo, com 19968 poços (9984 genes); o segundo conjunto foi formado apenas pelos genes considerados com sinal de boa qualidade (RSR >2), este conjunto foi denominado reduzido, contendo 8392 poços (ou 4196 genes).

A estrutura dos dados foi avaliada graficamente utilizando o logaritmo da razão das intensidades (M) versus a média de intensidades (A), para os canais vermelho e verde (gráfico MA ou MA-plot). Este gráfico apresentou um padrão não linear,

mostrando uma estrutura intensidade-dependente nos dados, indicando a necessidade de normalização. Após a normalização utilizando o gráfico MA verificou-se maior linearidade dos dados. A Figura 9 mostra gráficos MA para os conjuntos completo e reduzido, antes e depois da transformação *lowess*.

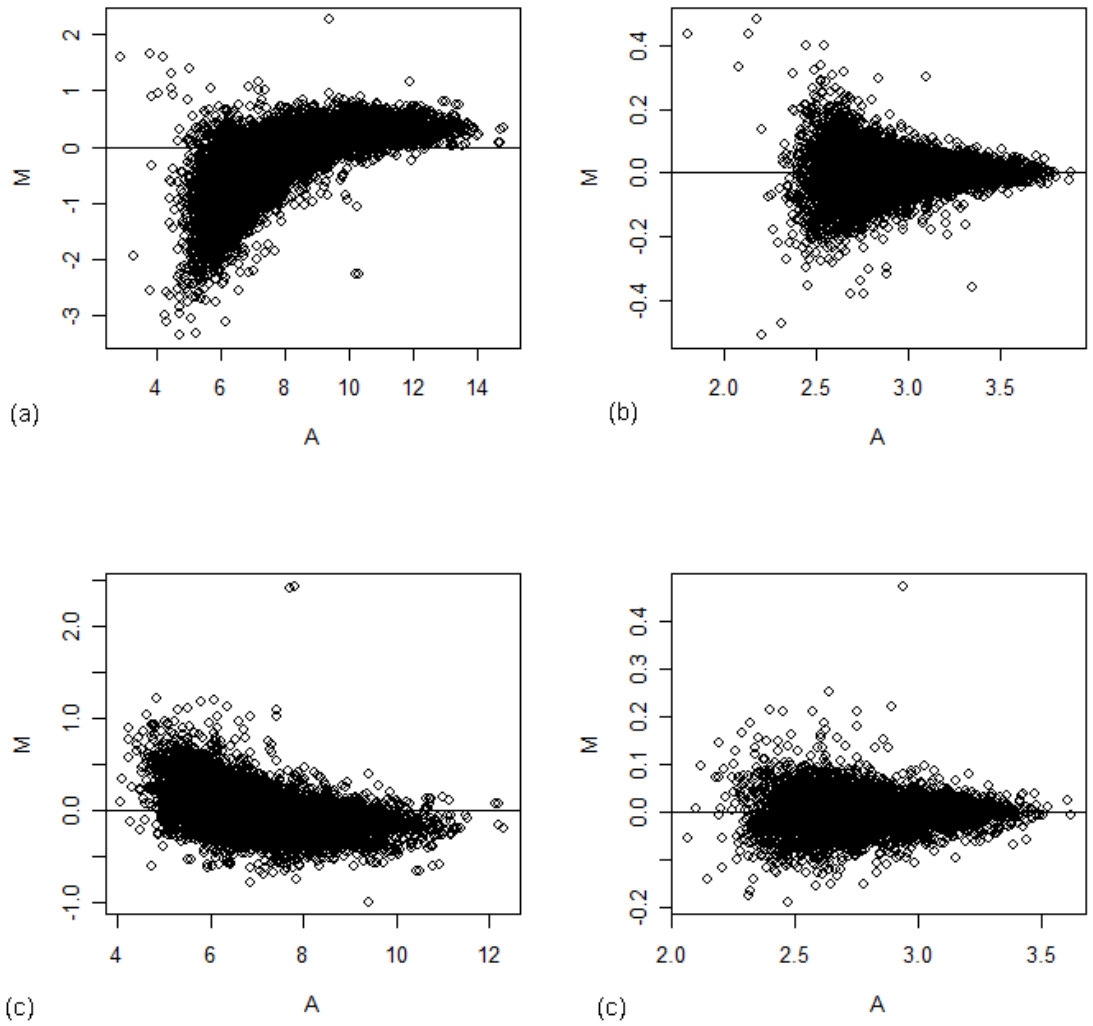

Figura 9: Gráfico MA para dados de corte antes e após a normalização pelo método lowess (a e b), os dados de conjunto completo (c e d), direta os gráficos para conjunto reduzido. Representação apenas de um arranjo.

O modelo utilizado na análise de variância do conjunto completo incluiu os termos arranjo, marcador e amostra e suas interações, utilizando o termo amostra como termo aleatório. Esta escolha foi baseada na inspeção gráfica dos resíduos. A Figura 10 apresenta os gráficos *boxplot* para os dados do conjunto completo para os dados pré-normalização. A Figura 11 apresenta os *boxplot* para os dados pós-normalização. Os gráficos de resíduos para o modelo escolhido são apresentados na Figura 12. Para o conjunto reduzido foi utilizado um modelo que incluiu os termos arranjo, marcador e amostra, sem empregar o termo aleatório. O gráfico de resíduos para este modelo é apresentado na Figura 13.

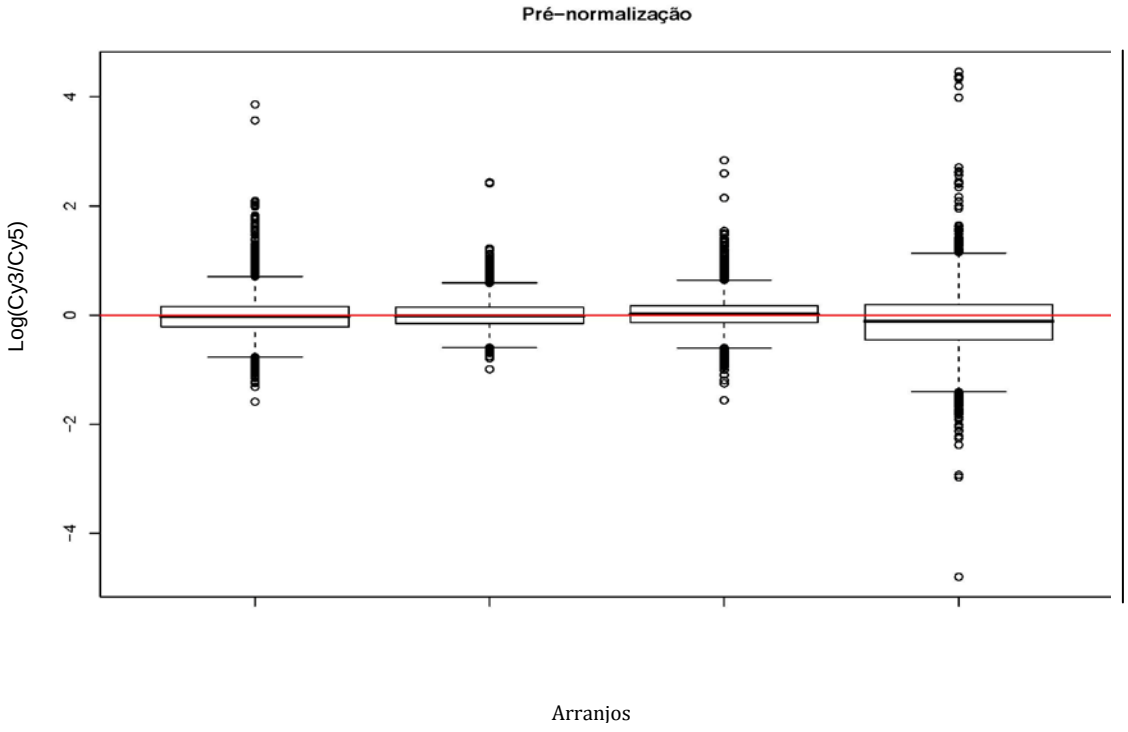

Figura 10: Boxplot dos arranjos pré-normalização. Dados de corte

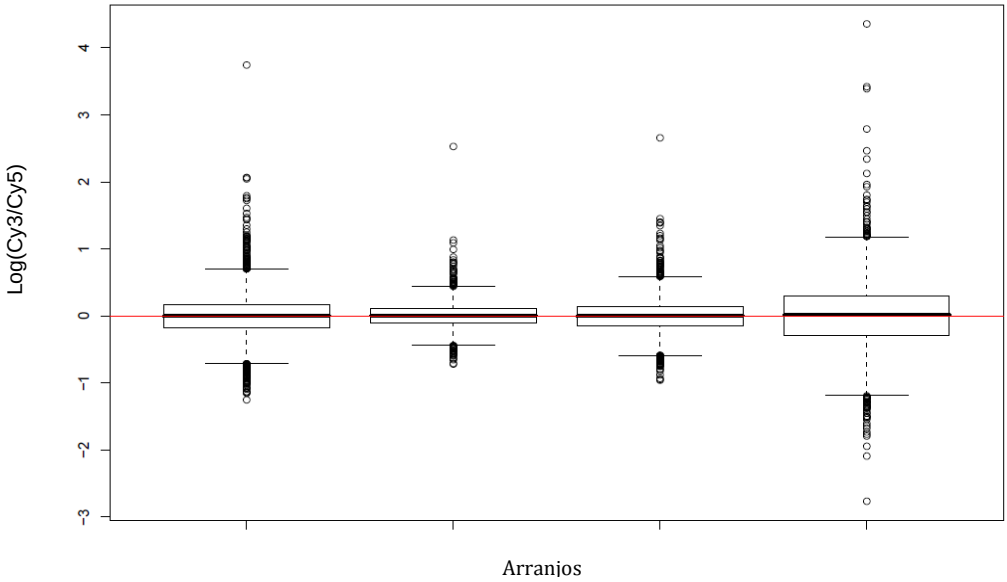

Figura 11: Boxplot dos arranjos pós normalização, dados de corte.

O efeito de interesse corresponde ao da interação entre gene e amostra, que corresponde à expressão diferencial. Para avaliar a significância estatística da expressão diferencial foi utilizado o teste de Fs<(0,0001 para os resultados do conjunto completo e α < 0,005, para o conjunto reduzido). Este resultado pode ser avaliado pelo gráfico Volcano, que permite a visualização dos efeitos biológicos e estatísticos.

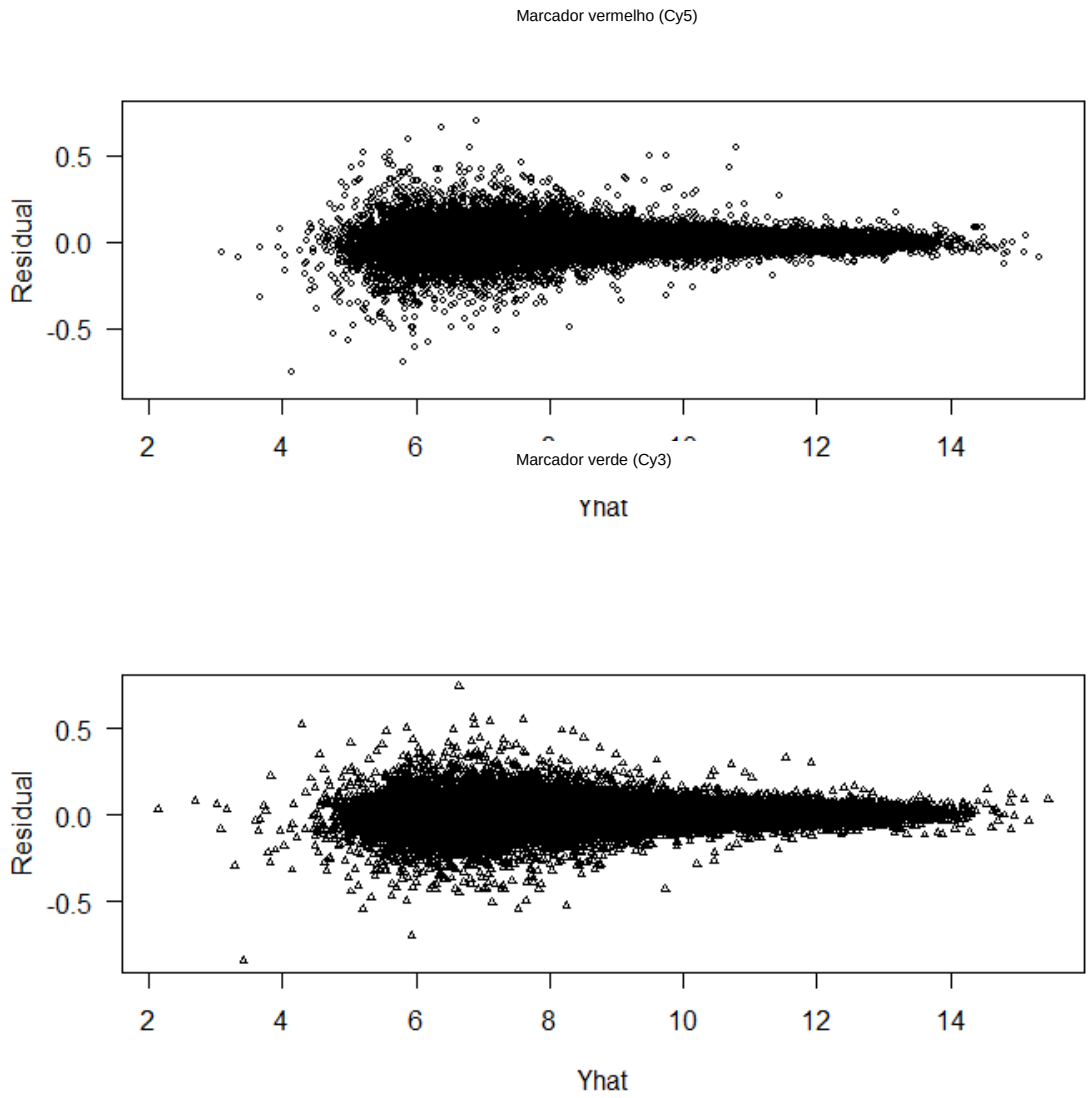

Figura 12: Gráfico de resíduos do modelo ANOVA para os dados do conjunto completo.

Para o conjunto completo, foram identificados 13 genes diferencialmente expressos. Entre estes, havia genes marcados na etapa de avaliação como de baixa qualidade. A Tabela 3 mostra os genes diferencialmente expressos, os respectivos valores p e razão de expressão, para os genes do conjunto completo. Para o conjunto reduzido (Tabela 4), foram encontrados oito genes diferencialmente expressos, entre

eles apenas o gene MRPL46 foi identificado diferencialmente expresso na análise de ambos os conjuntos.

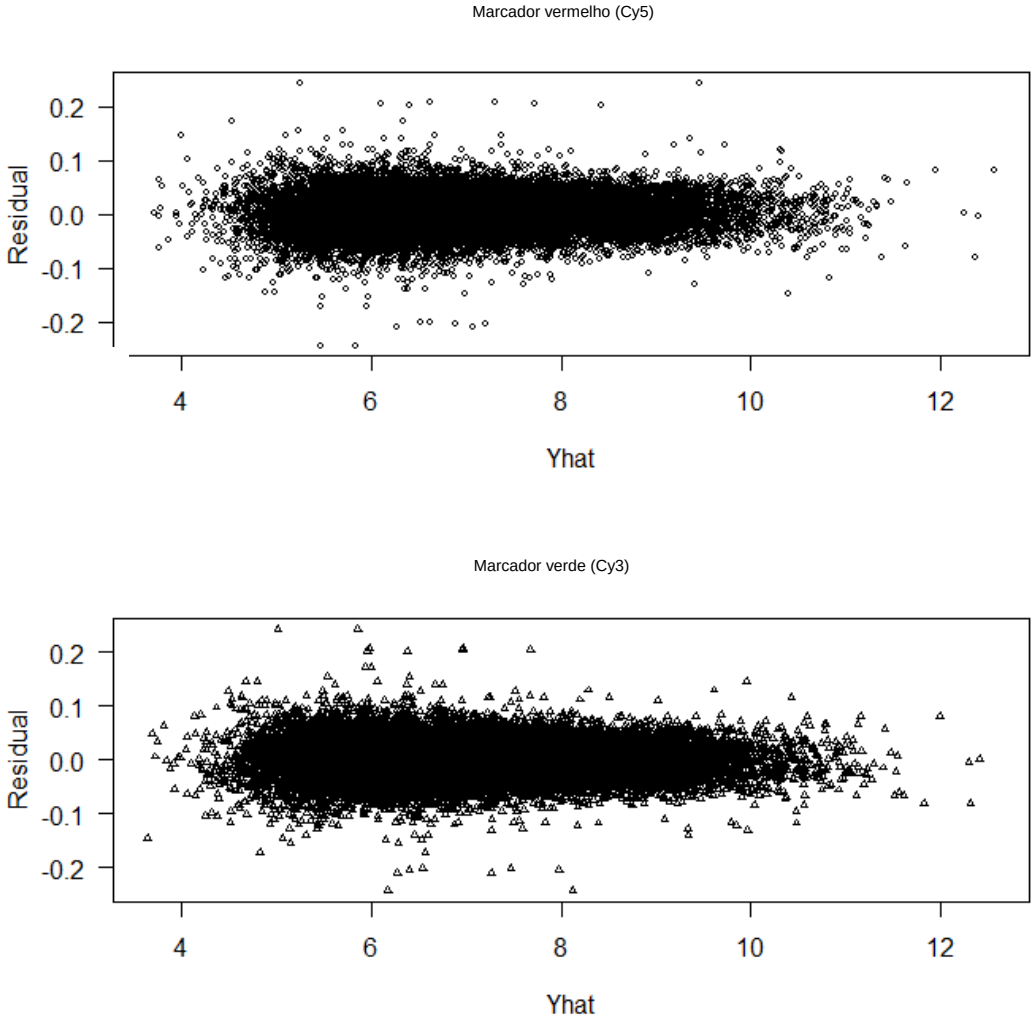

Figura 13: Gráfico dos resíduos do modelo ANOVA para os dados de corte.

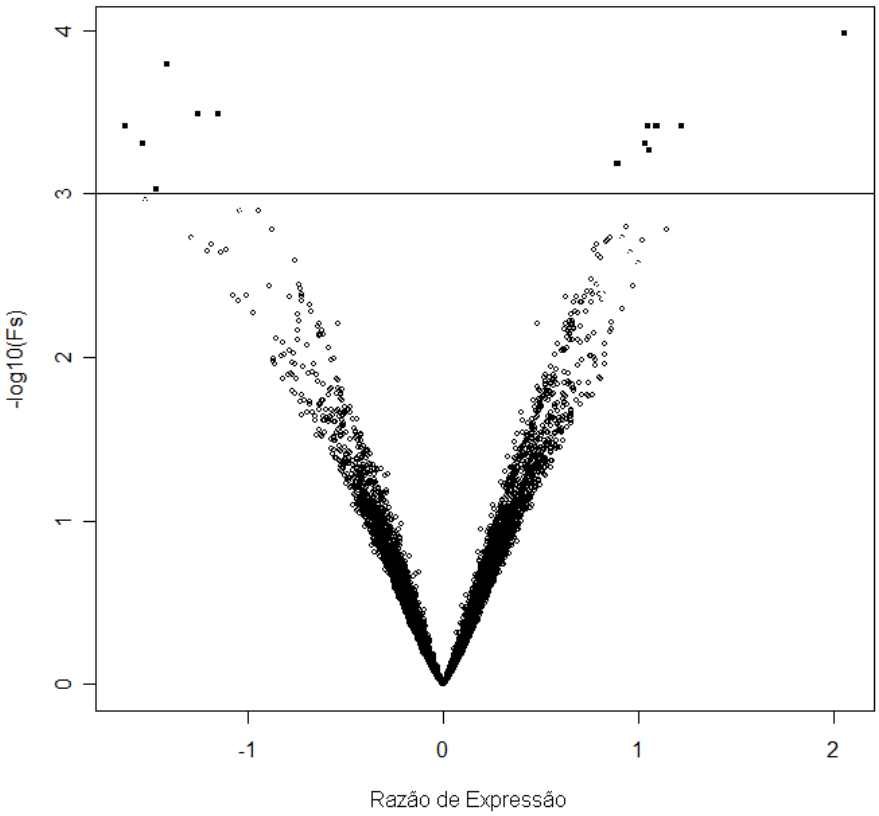

Figura 14: Gráfico Volcano para os dados do conjunto completo utilizando a estatística de teste Fs. Os genes acima da linha de corte são considerados diferencialmente expressos.

A análise empregando o segundo método, o produto de ordenação, utilizou também os dois conjuntos, completo e reduzido. Os genes diferencialmente expressos foram identificados como induzidos ou reprimidos de acordo com sua posição na Tabela de ordenação. Para o conjunto completo foram identificados 37 genes diferencialmente expressos, sendo 19 genes induzidos e 18 genes reprimidos (Tabela 06). Para o conjunto reduzido foram identificados 13 genes induzidos e cinco reprimidos (Tabela 8).

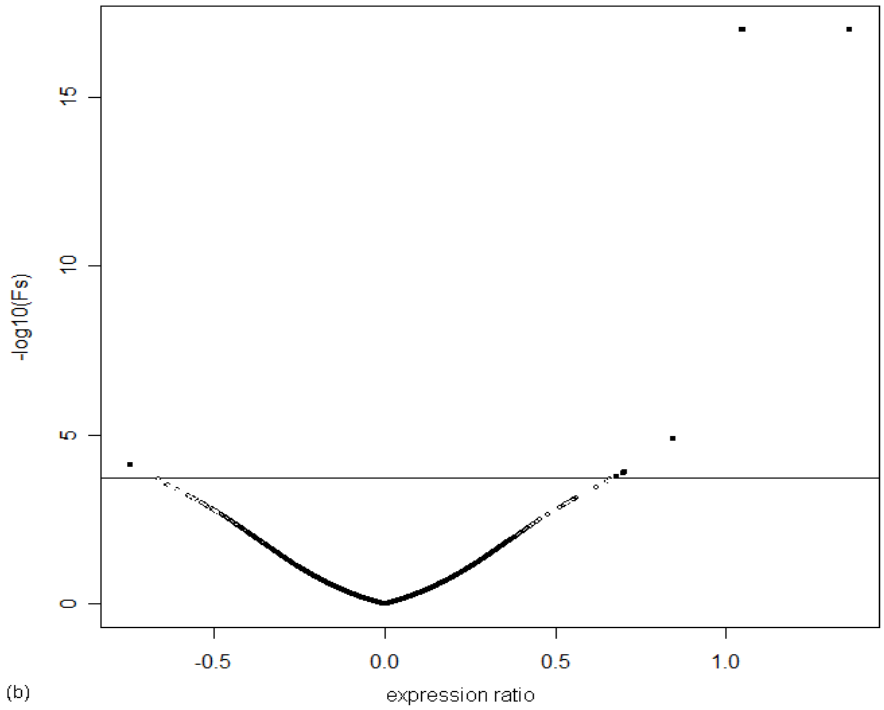

Figura 15: Gráfico Volcano para os dados de conjunto reduzido. Acima da linha de corte estão os genes diferencialmente expressos

Tabela **3**: Resultados de análise pelo modelo ANOVA, dados do conjunto completo

Número de Acesso	Símbolo	*Valor p*	FC
NM_018378	FBXL8	0.00016290	2.0
NM_002718	PPP2R3A	0.00032570	2.2
NM_000665	ACHE	0.00032570	2.1
NM_014499	P2RY10	0.00038440	-1.7
NM_000264	PTCH1	0.00038440	-2.3
NM_014230	SRP68	0.00038440	-2.1
NM_000673	ADH7	0.00038440	2.9
NM_022118	RBM26	0.00038440	-1.7
NM_003678	C22orf19	0.00049520	-2.1
NM_022163	MRPL46	0.00054730	2.7
NM_025094	hypothetical protein FLJ22184	0.00054730	-1.6
NM_004455	EXTL1	0.00065150	-1.9
NM_003192	TBCC	0.00093840	2.9

Tabela **4**: Genes diferencialmente expressos obtidos pela análise empregando o modelo ANOVA. Dados de conjunto reduzido

Número de Acesso	Símbolo	Valor *p*	FC
NM_022163	MRPL46	0,000000	2,6
NM_002032	FTH1	0,000000	2,1
XM_052761	Hypothetical protein 52761	0,000013	1,8
NM_000336	SCNN1B	0,000078	-1,7
NM_000581	GPX1	0,000128	1,6
NM_001020	RSP16	0,000133	1,6
XM_069860	similar to similar to ribosomal protein l31 (h, sapiens) loc136380	0,000168	1,6
NM_024597	MAP7D3	0,000198	-1,6

Tabela **5**: Teste condicional para representação de ontologia genética (função molecular) dos genes indicados como diferencialmente expressos utilizando o modelo ANOVA.

GOBPID	Pvalor	O.R.	ExpCount	Count	Size	Termo
GO:0016337	0,002	9,609	1	4	80	Adesão celular
GO:0002037	0,007	∞	0	1	1	Regulação negativa do transporte de L-glutamato
GO:0002439	0,007	∞	0	1	1	Resposta inflamatória crônica a estímulos antigênicos
GO:0002740	0,007	∞	0	1	1	Regulação negativa da secreção de citocinas durante a resposta
GO:0006927	0,007	∞	0	1	1	Apoptose de células transformadas
GO:0021568	0,007	∞	0	1	1	Desenvolvimento do Rombômero 2
GO:0021658	0,007	∞	0	1	1	Morfogênese do Rombômero 3 r
GO:0032891	0,007	∞	0	1	1	Regulação negativa do transporte de ácido orgânico
GO:0034116	0,007	∞	0	1	1	Regulação positiva da adesão celular heterotópica
GO:0035284	0,007	∞	0	1	1	Segmentação cerebral
GO:0045994	0,007	∞	0	1	1	Regulação positiva da iniciação traducional pelo ferro
GO:0050828	0,007	∞	0	1	1	Regulação do liquido de tensão superficial
GO:0051953	0,007	∞	0	1	1	Regulação negativa do transporte de aminas
GO:0022610	0,008	4,870	1	5	197	Adesão biológica

Tabela **6**: Resultado da análise do produto de ordenação. Dados do conjunto completo

	Número de acesso	Símbolo	FC	*p* valor
Genes Reprimidos	NM_022052	NXF3	0.35	0.00E+000
	NM_005568	LHX1	0.34	0.00E+000
	NM_000673	ADH7	0.30	0.00E+000
	NM_003457	ZNF207	0.23	0.00E+000
	NM_018976	SLC38A2	0.46	1.00E-004
	mwgaracontrol#006-06		0.30	1.00E-004
	mwghuman10K#00849_hk2		0.43	1.00E-004
	NM_007199	IRAK3	0.48	1.00E-004
	NM_018442	PC326	0.49	2.00E-004
	mwghuman20K#3553_hk1		0.50	2.00E-004
	NM_018207	FLJ10759	0.42	3.00E-004
	NM_002165	ID1	0.53	3.00E-004
	NM_003647	DGKE	0.54	4.00E-004
	NM_000142	FGFR3	0.48	4.00E-004
	NM_020243	C22orf2	0.54	5.00E-004
	NM_000567	CRP	0.54	5.00E-004
	NM_005654	NR2F1	0.53	5.00E-004
	NM_000965	RARB	0.52	6.00E-004
	NM_005337	HEM1	0.48	7.00E-004
	NM_003192	TBCC	0.39	8.00E-004
Genes Reprimidos	NM_022458	C7orf2	2.64	0.00E+000
	NM_003757	EIF3S2	2.68	0.00E+000
	NM_018510		3.04	1.00E-004
	NM_014766	SCRN1	2.01	3.00E-004
	NM_018499		2.85	3.00E-004
	NM_016006	ABHD5	1.91	4.00E-004
	NM_022116	FIGNL1	1.86	6.00E-004
	NM_005480	TROAP	1.94	6.00E-004
	XM_012405		2.10	7.00E-004
	NM_003946	NOL3	2.41	7.00E-004
	NM_018971	GPR27	1.87	7.00E-004
	NM_000264	PTCH	1.80	8.00E-004
	NM_000840	GRM3	2.09	8.00E-004
	NM_017896	C20orf11	1.98	8.00E-004
	NM_006871	RIPK3	1.88	9.00E-004
	NM_002562	P2RX7	1.82	1.00E-003
	NM_002705	PPL	2.03	1.00E-003
	NM_023007	FLJ12517	1.80	1.00E-003
	NM_005400	PRKCE	1.93	1.20E-003
	NM_005507	CFL1	1.73	1.30E-003

FC = Razão de expressão

Tabela **7**: Teste condicional para representação de ontologia genética (processo biológico) dos genes indicados como diferencialmente expressos utilizando o Produto de Ordenação.

GOBPID	Valor p	O.R.	ExpCount	Count	Size	Termo
GO:0006488	0.003	∞	0	1	1	Processo biossintético de oligossacarídeo ligado ao dolicol
GO:0006457	0.005	23.242	0	2	34	Conformação protéica
GO:0008535	0.007	334.000	0	1	2	Construção do complexo IV da cadeia respiratória
GO:0006405	0.013	111.259	0	1	4	Exportação do RNA a partir do núcleo
GO:0006878	0.013	111.259	0	1	4	Homeostase celular do íon cobre
GO:0006825	0.016	83.417	0	1	5	Transporte de íon cobre
GO:0009312	0.016	83.417	0	1	5	Processo biossintético de oligonucleotídeo
GO:0051028	0.042	27.731	0	1	13	Transporte de mRNA
GO:0050657	0.046	25.590	0	1	14	Transporte de ácido nucléico
GO:0051236	0.046	25.590	0	1	14	Determinação da localização do RNA
GO:0006986	0.064	17.474	0	1	20	Reposta a proteína não “dobrada”
GO:0030001	0.067	5.578	0	2	131	Transporte de íon metálico
GO:0051170	0.083	13.253	0	1	26	Importação nuclear
GO:0006486	0.098	11.026	0	1	31	Glicosilação proteína aminoácido
GO:0070085	0.098	11.026	0	1	31	Glicosilação

7. DISCUSSÃO

Neste trabalho, foi abordada a utilização de dois diferentes métodos de identificação de alteração na expressão gênica em resposta a utilização de uma droga. Esta análise representa uma exploração dos possíveis genes alterados em resposta à utilização da talidomida, empregando células PBMC, extraídas de indivíduos saudáveis.

O objetivo neste trabalho foi identificar alterações na expressão de genes humanos em resposta ao tratamento com talidomida, uma droga cujo mecanismo molecular de ação ainda permanece desconhecido. Para isto, dados de expressão de 9.984 oligonucleotídeos representativos de genes humanos obtidos a partir de experimentos de microarranjos de DNA foram utilizados. A análise destes dados de microarranjos foi realizada através dos métodos, MAANOVA e Produto de Ordenação, que utilizam a inferência estatística para determinar genes candidatos à expressão diferencial.

Em geral, dados de microarranjos são muito ruidosos e, quase sempre, não é possível identificar e eliminar todas as fontes de ruído. Uma fonte de ruído que pode influenciar sobremaneira a análise é a experimental. Uma forma de tentar controlar o ruído experimental é avaliar a qualidade das lâminas, numa etapa anterior à inferência estatística, o que pode aumentar a confiabilidade dos resultados. Há na literatura diversas formas de se avaliar a qualidade dos dados [78], [79], [80]. Neste trabalho optamos por avaliar a razão sinal ruído em cada poço, pois uma baixa razão sinal-ruído implica em aumento na variabilidade dos dados e redução na precisão dos resultados [79]. Neste estudo, a avaliação da qualidade dos dados empregando a RSR

identificou que 58% dos poços em todas as lâminas eram de baixa qualidade. De acordo com COOMBES *et al.* [42] em um experimento de microarranjos usando dois marcadores (microarranjos de DNA) espera-se que 30% a 60% dos poços apresentem RSR menor que dois.

Após a avaliação da qualidade dos poços, foi realizada a análise dos dados empregando duas abordagens: na primeira, o conjunto de análise, composto por quatro lâminas, recebeu uma marcação nos genes em cada lâmina quando a qualidade de sinal era considerada duvidosa. Na segunda, os genes de qualidade questionável foram eliminados. As duas abordagens foram consideradas porque este estudo representa uma análise exploratória que visa identificar o perfil de expressão dos genes nas células tratadas com talidomida. Genes com baixa razão sinal ruído podem ser considerados pouco informativos em uma análise em que um determinado perfil é esperado, ou seja, onde se espera que, por exemplo, a maioria dos genes na lâmina não se expresse. Porém, este não é o caso desta análise, onde desconhecemos o perfil esperado. A análise utilizando o conjunto de dados reduzido foi baseada no conceito de que a filtragem do banco de dados, eliminando genes com sinal de expressão baixo e, entre eles genes possivelmente não informativos, poderia reduzir o número de hipóteses testadas, minimizando o problema da testagem múltipla [81].

O desenho experimental é determinado pela necessidade de minimizar os efeitos das variações indesejáveis, enquanto aumenta-se a precisão das estimativas dos parâmetros de interesse: a alteração da expressão gênica de uma condição em relação à outra [82]. De modo geral, o desenho experimental é definido de acordo com a pergunta a ser respondida e com a disponibilidade de recursos. Este estudo utilizou

como desenho experimental o desenho de referência, que consiste na hibridização de cada amostra de teste com um tipo comum de amostra [33]. Para empregar este desenho, nós consideramos as células PBMC sem tratamento como a referência constituída de um "*pool*" células extraídas de sangue periférico de três doadores saudáveis. O desenho de referência apresenta algumas vantagens, entre elas a flexibilidade na análise, possibilitando a inclusão de outras lâminas posteriormente. Este trabalho utilizou o conceito de desenho de referência replicado na análise dos experimentos, considerando a amostra PBMC sem tratamento como tempo zero [33]. Desta forma conseguimos maior flexibilidade na análise, permitindo a incorporação de outras lâminas ao estudo, caso necessário, e permite a análise empregando o modelo ANOVA.

Estudos de microarranjos empregando dois marcadores estão sujeitos à presença de viés devido às diferenças na incorporação dos marcadores, este viés pode ser estimado pela introdução de uma lâmina de marcação reversa. Este estudo empregou a marcação reversa em cada grupo, entretanto havia apenas uma lâmina de marcação reversa para duas de marcação direta nos grupos, o que poderia dificultar a estimativa do viés de marcador, em virtude do desequilíbrio entre marcadores. Para equilibrar as análises foi eliminada uma lâmina de marcação direta em cada grupo. A escolha da lâmina a ser eliminada foi feita com base na avaliação da qualidade do sinal, pela relação sinal ruído, onde a lâmina que apresentou maior número de poços com razão sinal ruído baixa foi eliminada.

A estrutura dos dados de microarranjos, que apresenta muitas fontes de variabilidade, introduz a necessidade do uso de transformações e normalizações antes da inferência estatística, pois a maioria dos métodos estatísticos requer suposições

sobre a homocedasticidade dos dados. Neste estudo foi usada a transformação logarítmica dos dados corrigidos para *background*. Esta transformação permite a estabilização da variância para valores de expressão elevados. Contudo, para valores de expressão mais baixos, principalmente valores próximos ao de *background*, ocorre um aumento da variabilidade [83]. Ao examinarmos o gráfico MA, um gráfico de espalhamento, dos dados log-transformados, verificamos a estrutura intensidade dependente, indicando ainda a necessidade de normalização. Para isto, foi empregado, o método *lowess*, um método de normalização que adota uma estratégia de ajuste de curva usando a distribuição local dos dados (no eixo da intensidade) para recentralizá-los [58].

Conforme já mencionado, experimentos de microarranjos tendem a apresentar um pequeno número de replicações [21] e os diferentes métodos estatísticos usados para interpretação dos resultados apresentam limitações na consideração deste problema. Métodos que se mostram hábeis na identificação de genes diferencialmente expressos em condições de pouca ou nenhuma replicação em geral apresentam resultados inconsistentes e de difícil [65]. Levando-se em conta esta questão, decidimos utilizar dois métodos estatísticos na identificação de genes diferencialmente expressos. Apesar dos resultados dos métodos não serem comparáveis quanto à sua eficiência, os resultados podem ser considerados complementares e a identificação de genes diferencialmente expressos em comum por ambos os métodos pode ser indício de verdadeiro positivo.

A escolha do método de análise não é trivial. Há diferentes métodos de análise propostos, enfocando as diferentes características experimentais. Porém, a maioria destes não apresenta desempenho adequado em experimentos com poucas

replicações e/ou mais de duas condições experimentais. Um método que se adapta às condições deste trabalho é o que utiliza um modelo de análise de variância (ANOVA) [84]. O modelo ANOVA permite a separação das diversas fontes de variabilidade que devem ser consideradas na construção de testes de expressão diferencial [85]. Desta forma, os níveis de intensidade são expressos como um somatório de componentes conforme a equação 03 [MODELO ANOVA], possibilitando também a análise de experimentos multifatoriais. Um fator é um grupo de tratamentos ou condições. Neste trabalho, é considerado apenas um fator, o tratamento com três níveis: sem tratamento (PBMC, apenas), DMSO, DMSO+Talidomida.

Outro método identificado como adequado às condições experimentais foi o produto de ordenação [21] Neste método, poucas suposições são necessárias à respeito da natureza dos dados. Considera-se apenas que haja poucos genes diferencialmente expressos no conjunto de análise, que as medições são independentes entre os arranjos, que a maioria das alterações de expressão é independente e que a variância é igual para todos os genes [21]. Na estatística clássica, aplicações onde a suposição de uniformidade da variância é necessária uma prática comum é buscar transformações de dados que estabilizem a variância [44]. Após o emprego das técnicas de normalização, pode-se assumir a homocedasticidade dos dados, Figura 10 (boxplot dos dados normalizados). Para que os resultados pudessem ser comparáveis, as técnicas de normalização usadas neste método foram as mesmas empregadas no modelo ANOVA.

As técnicas de normalização são importantes para redução ou eliminação dos desvios sistemáticos. A transformação logarítmica foi empregada, pois possibilita a conversão das razões entre os dois canais em diferenças para cada poço e, também,

estabiliza a variância dos poços de alta intensidade [44]. Após a log-tranformação, os dados apresentavam características intensidade dependente e, por isto, foram normalizados utilizando o método *lowess* um procedimento de ajuste de curva. Este método de regressão local reduziu a variabilidade dependente da intensidade como visto na Figura 10, permitindo que as diferenças biológicas fossem mais bem visualizadas.

A análise estatística foi realizada empregando dois métodos, modelo ANOVA e Produto de Ordenação. O primeiro método, que utiliza um modelo ANOVA em duas etapas, leva em consideração o desenho experimental empregado, os marcadores, as replicações, os genes, além das suas interações. Dessa forma, a contribuição dos diferentes fatores pode ser considerada para que os efeitos genes específicos sejam estimados e a significância dos termos individuais do modelo pode ser testada. Para cada gene, foi possível testar se a expressão era alterada pelo tratamento, gerando uma lista de estatística F. O modelo ANOVA é empregado no pacote MAANOVA, que permite a utilização de adaptações da estatística F, possibilitando que a estatística de teste seja escolhida de acordo com a característica dos dados. Neste trabalho utilizou-se a estatística Fs, pois o pequeno número de replicações empregado não permite uma estimativa precisa dos parâmetros amostrais e os pressupostos de normalidade dos resíduos e das variâncias, que devem ser assumidos ao utilizar a estatística F tradicional não precisaram ser assumidos aqui. A significância da estatística F (valor *p*) foi determinada pela análise de permutação dos dados originais. Este procedimento, apesar de menos robusto é necessário, pois a estatística *Fs* não segue uma distribuição tabular.

A visualização dos resultados foi realizada considerando-se a significância estatística com relação à magnitude da alteração de expressão empregando-se o gráfico Volcano (Figuras 11a e 11b), um gráfico de espalhamento dos valores de expressão relativa contra o p valor para cada gene, que permite a inspeção visual dos resultados.

Um dos problemas decorrentes da utilização de métodos estatísticos na tecnologia de microarranjos é equilibrar os erros tipo I e II, minimizando o erro tipo I, aquele que consiste em rejeitar a hipótese nula quando ela é verdadeira, controlando a detecção de falsos positivos (erro tipo II).

Apesar de inúmeros trabalhos utilizando microarranjos indicarem a necessidade de correção de testes múltiplos, neste trabalho esta correção não foi realizada. Alguns motivos para isto podem ser listados. O objetivo deste estudo é criar uma hipótese que explique o mecanismo de ação da talidomida, estudo exploratório. Os genes selecionados pelos testes estatísticos serão confirmados em análises posteriores. Estudos exploratórios normalmente requerem flexibilidade de desenho e análise, sendo possível desconsiderar a correção de múltiplos testes [86].

A lista de genes selecionados como genes induzidos e genes reprimidos pelo método produto de ordenação apresentou concordância com a lista de genes indicados como diferencialmente expressos pelo modelo ANOVA. Este resultado é bastante satisfatório, pois apesar dos métodos não serem comparáveis entre si, a concordância indica maior possibilidade de que os genes encontrados sejam realmente diferencialmente expressos.

Para conjuntos de dados ruidosos com um pequeno número de replicações, após normalização adequada, o produto de ordenação fornece uma ferramenta adequada para detecção de genes diferencialmente expressos.

GENES CANDIDATOS À EXPRESSÃO DIFERENCIAL

Poucos estudos foram identificados na literatura envolvendo uso de microarranjos para analise da talidomida. Entre estes está o estudo de VACCA *et AL.* [56]sobre o efeito antiangiogênico da Talidomida. Para isto, os autores, utilizando células endoteliais de medula óssea, estudaram o efeito da Talidomida em pacientes portadores de diferentes doenças cujo mecanismo de angiogênese é determinante do prognóstico clínico. Este estudo identificou genes relacionados à apoptose, como o BCL-2, além de genes realcionados à angiogênese. Nesse trabalho, apesar do modelo empregado não avaliar a ação da Talidomida *in vivo* foram identificados genes relacionados ao processo de apoptose. Além de genes relacionados ao processo de remodelamento vascular.

A talidomida possui diversas atividades que podem ser definidas como: antiangiogênica, anti-inflamatória e imunomodulatória, embora o mecanismo de ação molecular que rege estas atividades não tenha sido elucidado. Neste trabalho foram utilizadas células tratadas e não tratadas com talidomida. Por ser uma droga com atividade óptica, instável em meio aquoso, a talidomida foi dissolvida em dimetil-sulfóxido (DMSO). O DMSO é um composto químico orgânico, com elevada capacidade higroscópica que apresenta propriedades farmacológicas e terapêuticas resultantes de sua capacidade de interagir ou combinar com substâncias orgânicas e muitas drogas sem alterar sua configuração molecular [87]. Seus efeitos são classificados como: antiinflamatórios, antioxidantes e anestésicos locais [88], com

larga utilização como medicamento veterinário. Tem seu uso aprovado pelo FDA, em humanos, desde 1978 [87]. Devido às propriedades do DMSO serem em alguns aspectos semelhantes as da Talidomida, foi necessária a utilização de um controle apenas com DMSO para que se evidenciar os efeitos de interesse da talidomida.

A análise utilizando o pacote MAANOVA e, também o método produto de ordenação, comparou o controle DMSO com o tratamento Talidomida (diluída em DMSO), gerando listas de genes diferencialmente expressos.

A função dos genes foi obtida no Entrez Gene, uma base de dados que permite a integração da informação obtida a partir do sequenciamento genômico gerando uma chave de anotação [89]. Dos genes selecionados, alguns podem fornecem informações sobre o mecanismo de ação da talidomida.

O gene F-box (FBXL8), está ligado ao sistema ubiquitina, um sistema relacionado à via de degradação de proteínas celulares, que desempenha importantes papéis em uma variedade de vias incluindo caquexia, como a que ocorre na AIDS [90]. Adicionalmente, esse gene está relacionado a respostas imune e antiinflamatória, através do processamento de antígenos MHC classe I. Nesta análise o gene está superexpresso, indicando que a via ubiquitina pode estar ativada na resposta antiinflamatória induzida pela talidomida.

A Talidomida também foi identificada como capaz de reduzir a replicação do HIV-1 in vitro, mas os mecanismos ainda não foram elucidados. Esta análise identificou o gene PP2A como diferencialmente expresso, com *fold-change* de 2,30 (valor p de 0,0003). Este gene está implicado no controle negativo do crescimento e divisão celular, além de estar envolvido no controle da transcrição do RNA viral e pode

estar relacionado à regulação da replicação do HIV- 1 [91]. Em trabalho recente DIAS {[92]} analisando as alterações na expressão gênica da talidomida em células tratadas com Talidomida identificou o gene PP2A subunidade regulatória B. Neste estudo foi analisado o perfil de expressão de células tratadas com LPS (modelo de inflamação) em comparação com células tratadas com LPS e Talidomida, utilizando o modelo ANOVA. Este resultado pode indicar o efeito da talidomida sobre a expressão do gene PP2

Os resultados também indicaram o gene PTCH1 como reprimido (*fold-change* menor que um). Este gene age na formação de estruturas embrionárias [93] e, o mecanismo através do qual a talidomida exerce sua ação teratogênica talvez possa ser explicado pela inibição deste gene.

Os genes identificados em comum entre a análise utilizando o pacote MAANOVA e o produto de ordenação foram o PTCH1, ADH7, TBCC.

O TBCC (cofator C da Tubulina *folding*) é um gene codificador de proteína. O cofator C é uma das quatro proteínas (cofatores A, D, E e C) envolvidos na via que leva à conformação final da β -tubulina correta a partir de seus intermediários. Os cofatores A e D acredita-se que desempenhem um papel na captura e estabilização de intermediários da β -tubulina numa conformação quase nativa. O cofator E se liga ao cofator D/complexo β-tubulina; a interação com o cofator C leva a dos polipeptídios de β-tubulina que são levados ao seu estado nativo [94].

O ADH7, gene codificador da enzima álcool desidrogenase 7 (classe IV), a subunidade μ ou σ, que é membro da família álcool desidrogenase. Os membros desta família metabolizam uma ampla variedade de substratos que incluem etanol, retinol e

outros alcoóis alifáticos, hidroxiesteróides e produtos de pré-oxidação lipídica. A enzima codificada por este gene é ineficiente na oxidação do etanol, mas ativa como retinol desidrogenase; participando então na síntese do ácido retinóico, um importante hormônio da diferenciação celular [95].

O MRPL46 foi o único identificado como diferencialmente expresso na análise empregando o modelo ANOVA tanto no conjunto completo quanto reduzido. Este gene codifica a subunidade 39S da proteína ribossomal de mamíferos

O método produto de ordenação identificou um número mais elevado de genes como induzidos ou reprimidos. De acordo com BREITLING *et al. [65]*, a violação da suposição da homogeneidade das variâncias, pode levar a uma superestimativa dos genes diferencialmente expressos. Após a utilização do corte pelo percentual de falsos positivos (uma estimativa da taxa de falsa descoberta), a quantidade de genes considerada como diferencialmente expressa foi reduzida. Este redução ocorreu apenas na lista de genes considerados com reprimidos. Uma explicação para isto pode estar na estrutura intensidade-dependente dos dados, que permaneceu mesmo após a transformação pelo método *lowess*. Apesar disto, os resultados são satisfatórios para uma análise exploratória.

A obtenção de listas de genes diferencialmente expressos não fornece informação consistente sobre os mecanismos biológicos envolvidos na alteração de expressão celular, pois a análise de um único gene pode perder informação importante sobre os efeitos de outros genes na mesma via metabólica, por exemplo. Para superar esta questão, algumas abordagens foram propostas para agregar às listas de genes diferencialmente expressos informação biológica relevante baseando-se no agrupamento dos genes em conjuntos, como classes ou vias, por exemplo. Estes

conjuntos podem ser definidos com base em um conhecimento prévio a respeito de vias bioquímicas ou expressão em experimentos anteriores. Uma das informações amplamente utilizadas é a ontologia gênica. Neste trabalho a utilização do teste hipergeométrico para se testar alguma categoria ontológica era super-representada na lista de genes diferencialmente expressos permitiu a identificação de processos biológicos nos quais a talidomida exerce seus efeitos. Foram identificadas categorias como resposta inflamatória crônica a estímulos antigênicos, apoptose celular, regulação negativa da secreção de citocina na resposta imune, que correspondem a processos biológicos nos quais a talidomida exerce efeito.

8. CONCLUSÃO

Neste trabalho, os dados de expressão apresentavam grande variabilidade e sinal de baixa intensidade. A escolha do método de análise considerou a estrutura dos dados e a quantidade de replicações, priorizando métodos que requeriam poucas suposições. Foram empregados dois métodos, o modelo ANOVA e o Produto de Ordenação. O modelo ANOVA permite a separação das fontes de variabilidade e identificação dos efeitos especificamente relacionados aos genes. Desta forma, variações devidas ao tratamento podem ser identificadas. O segundo método empregado, o produto de ordenação, utiliza o conceito biológico de razão de expressão identificando aqueles genes que tem razão de expressão elevada devido ao tratamento e não por chance. As fontes de variabilidade não são tratadas separadamente neste método, por isto ele requer transformação e normalização cuidadosas. Técnicas de transformação e normalização são de grande importância em estudos envolvendo microarranjos, principalmente devido às variações experimentais, incluindo as relacionadas a amostra biológica e etapas experimentais. Nestes estudos as técnicas aplicadas se mostraram adequadas na minimização destas variações. A partir dos dados transformados e normalizados, a análise da expressão gênica diferencial pode ser determinada. As transformações e normalizações foram comuns aos dois métodos, sendo as diferenças na identificação dos genes devidas apenas a características dos métodos. Na literatura não há identificação de um método padrão-ouro, sendo assim a aplicação de dois métodos pode ser considerada complementar. Neste trabalho, a aplicação dos dois métodos identificou duas listas de genes, havendo uma pequena interseção entre elas, o que esta de acordo com o identificado na literatura que mostra que a aplicação de

métodos diferentes apresenta resultados divergentes. As listas de genes compreendem genes quem quando avaliados individualmente, podem ser relacionados a ações da talidomida. Estas listas são importantes resultados, mas acrescentam pouca informação a respeito do mecanismo de ação da talidomida. A complementação da análise avaliando estas listas com relação aos conjuntos gênicos, baseados em informações biológicas já conhecidas, identificou possíveis sítios de ação da droga, entre eles a apoptose celular e regulação negativa da secreção de citocinas. Apesar destes resultados requerem confirmação por estudos moleculares eles fornecem pistas importantes sobre os mecanismos de ação da talidomida.

Desta forma, os métodos escolhidos se mostraram adequados na triagem dos genes diferencialmente expressos que deverão ser confirmados em análises posteriores.

9. REFERÊNCIAS BIBLIOGRÁFICAS

1. CALABRESE, L. and FLEISCHER, A.B., *Thalidomide: current and potential clinical applications.* **The American Journal of Medicine**, 2000. 108(9): p. 487-495.
2. ALLEN, B.R., *Thalidomide.* **British Journal of Dermatology**, 2001. 144 (2): p. 227-229.
3. MUJAGIC, H., CHABNER, B.A., and MUJAGIÆ, Z., *Mechanisms of Action and Potential Therapeutic Uses of Thalidomide.* **croatian medical journal**, 2002. 43(3): p. 274-285.
4. SAMPAIO, E.P., CARVALHO, D.S., NERY, J.A.C., LOPES, U., and SARNO, E., *Thalidomide: an overview of its pharmacological mechanisms of action.* **anti-inflamatory and anti-allergy in medicinal chemistry**, 2006. 5 (1): p. 71-77.
5. D'AMATO, R.J., LOUGHNAN, M.S., FLYNN, E., and FOLKMAN, J., *Thalidomide is an inhibitor of angiogenesis.* **Proceedings of the National Academy of Sciences of the United States of America**, 1994. 91(9): p. 4082-4085.
6. MINCHINTON, A.I., FRYER, K.H., WENDT, K.R., CLOW, K.A., and HAYES, M.M., *The effect of thalidomide on experimental tumors and metastases.* **Anti-Cancer Drugs**, 1996. 7(3): p. 339-343.
7. GULLESTAD, L., UELAND, T., FJELD, J.G., HOLT, E., GUNDERSEN, T., BREIVIK, K., FOLLING, M., HODT, A., SKARDAL, R., KJEKSHUS, J., ANDREASSEN, A., KJEKSHUS, E., WERGELAND, R., YNDESTAD, A., FROLAND, S.S., SEMB, A.G., and AUKRUST, P., *Effect of Thalidomide on Cardiac Remodeling in Chronic Heart Failure: Results of a Double-Blind, Placebo-Controlled Study.* **Circulation**, 2005. 112(22): p. 3408-3414.
8. PENNA, G.O., PINHEIRO, A.M.C., and HAJJAR, L.A., *Talidomida: mecanismo de ação, efeitos colaterais e uso terapêutico.* **Anais Brasileiros de Dermatologia**, 1998. 75: p. 501-514.
9. MOSMANN, T.R. and SAD, S., *The expanding universe of T-cell subsets: Th1, Th2 and more.* **immunology today**, 1996. 17 (3): p. 138-146.
10. DEBOUCK, C. and GOODFELLOW, P.N., *DNA microarrays in drug discovery and development.* **Nature genetics**, 2000. 21: p. 48-50.

11. HELLER, M.J., *DNA MICROARRAY TECHNOLOGY: Devices, Systems, and Applications.* **Annu. Rev. Biomed. Eng.**, 2002. 4: p. 129-153.
12. CHAUDHURI, J.D., *Genes arrayed out for you: the amazing world of microarrays.* **Med Sci Monit**, 2005. 11(2): p. RA52-62
13. CROWTHER, D.J., *Applications of microarrays in the pharmaceutical industry.* **Current Opinion in Pharmacology**, 2002. 2: p. 551-554.
14. KUDOH, K., RAMANNA, M., RAVATN, R., ELKAHLOUN, A.G., BITTNER, M.L., MELTZER, P.S., TRENT, J.M., DALTON, W.S., and CHIN, K.-V., *Monitoring the Expression Profiles of Doxorubicin-induced and Doxorubicin-resistant Cancer Cells by cDNA Microarray.* **Cancer Res**, 2000. 60(15): p. 4161-4166.
15. QUACKENBUSH, J., *Computational analysis of microarray data.* **Nat Rev Genet**, 2001. 2: p. 418-427.
16. LEUNG, Y.F. and CAVALIERI, D., *Fundamentals of cDNA microarray data analysis.* **Trends in Genetics**, 2003. 19(11): p. 11.
17. DUDOIT, S., SHAFFER, J.P., and BOLDRICK, J.C., *Multiple Hypothesis Testing in Microarray Experiments.* **Statistical Science**, 2003. 3(1): p. 71-100.
18. DRUMMOND, R.D., PINHEIRO, A., ROCHA, C.S., and MENOSSI, M., *ISER: selection of differentially expressed genes from DNA array data by non-linear data transformations and local fitting.* **BIOINFORMATICS**, 2005. 21(24): p. 4427-4429.
19. MUTCH, D., BERGER, A., MANSOURIAN, R., RYTZ, A., and ROBERTS, M.-A., *The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data.* **BMC Bioinformatics**, 2002. 3(1): p. 17
20. MANSOURIAN, R., MUTCH, D.M., ANTILLE, N., AUBERT, J., FOGEL, P., LE GOFF, J.-M., MOULIN, J., PETROV, A., RYTZ, A., VOEGEL, J.J., and ROBERTS, M.-A., *The Global Error Assessment (GEA) model for the selection of differentially expressed genes in microarray data.* **BIOINFORMATICS**, 2004. 2(16): p. 2726-2737.
21. BREITLING, R., ARMENGAUD, P., AMTMANN, A., and HERZYK, P., *Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.* **FEBS Lett**, 2004. 573: p. 83-90.
22. ALBERTS, B., *Molecular Biology of the Cell*. 2002, New York: Garland Science. 4301
23. LEHNINGER, *Principles of Biochemical.* 2000.

24. HARRINGTON, C.A., ROSENOW, C., and RETIEF, J., *Monitoring gene expression using DNA microarrays.* **Current Opinion in Microbiology** 2000. 3: p. 285-291.
25. REECE, R.J., *Analysis of Genes and Genomes*. 2004, UK: University of Manchester.
26. HOWBROOK, D.N., VALK, A.M.V.D., O'SHAUGHNESSY, M.C., SARKER, D.K., BAKER, S.C., and LLOYD, A.W., *Developments in microarray technologies.* **Drug Discovery Today**, 2003. 8 p. 643-651.
27. OLEKSIAK, M.F., CHURCHILL, G.A., and CRAWFORD, D.L., *Variation in gene expression within and among natural populations.* **Nat Genet**, 2002. 32 p. 261-266
28. DOBBIN, K., SHIH, J., and SIMON, R., *Statistical design of reverse dye microarrays.* **BIOINFORMATICS**, 2003. 19(7): p. 803-810.
29. KERR, M.K. and CHURCHILL, G.A., *Statistical design and the analysis of gene expression microarray data.* **Genet Res**, 2001. 77: p. 123-128.
30. KERR, M., AFSHARI, C., BENNETT, L., BUSHEL, P., MARTINEZ, J., WALKER, N., and CHURCHILL, G., *Statistical Analysis of a gene expression microarray experiment with replication.* **Statistica Sinica**, 2002. 12 %6: p. 203-217 %&.
31. YANG, Y. and SPEED, T., *Design issues for cDNA microarray experiments.* **Nat Rev Genet**, 2002. 3 (8): p. 579-588.
32. GLONEK, G.F.V. and SOLOMON, P.J., *Factorial and time course designs for cDNA microarray experiments.* **Biostatistics**, 2004. 5(1): p. 89-111.
33. STEIBEL, J.P. and ROSA, G.J.M., *On Reference Designs For Microarray Experiments.* **Statistical Applications in Genetics and Molecular Biology**, 2005. 4: p. 36.
34. WOLFINGER, R.D., GIBSON, G., WOLFINGER, E.D., BENNETT, L., HAMADEH, H., BUSHEL, P., AFSHARI, C., and PAULES, R.S., *Assessing gene significance from cDNA microarray expression data via mixed models.* **J Comput Biol**, 2001. 8 p. 625-637
35. CAUSTON, H.C., QUACKENBUSH, J., and BRAZMA, A., *A beginner's guide microarray gene expression data analysis*. 2003: Blackwell Science.
36. SCHENA, M., SHALON, D., HELLER, R., CHAI, A., BROWN, P.O., and DAVIS, R.W., *Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.* **Proc Natl Acad Sci USA**, 1996. 93 p. 10614-10619.
37. PASANEN, T., SAARELA, J., SAARIKKO, L., TOIVANEN, T., TOLVANEN, M., VIHINEN, M., and WONG, G., *DNA Microarray Data Analysis*, ed. J. Tuimala and M.M. Laine. 2003.

38. SCHARPF, R.B., IACOBUZIO-DONAHUE, C.A., SNEDBON, J.B., and PARMIGIANI, G., *When should one subtract background fluorescence in 2-color microarrays?* **Biostatistics**, 2006. 0(0): p. 1-13.
39. WANG, X., GHOSH, S., and GUO, S.-W., *Quantitative quality control in microarray image processing and data acquisition.* **Nucl. Acids Res.**, 2001. 29 (15): p. e75-
40. REIMERS, M. and WEINSTEIN, J.N., *Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases.* **BMC Bioinformatics**, 2005. 6(166).
41. TRAN, P.H., PEIFFER, D.A., SHIN, Y., MEEK, L.M., BRODY, J.P., and CHO, K.W.Y., *Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals.* **Nucl. Acids Res.**, 2002. 30 %6(12): p. e54- %&.
42. COOMBES, K., WANG, J., and ABRUZZO, L.V. *Monitoring the quality of microarrary expreiments*. 2003: Springer US.
43. SAVIOZZI, S. and CALOGERO, R.A., *Microarray probe expression measures, data normalization and statistical validation.*, in *Comparative and Functional Genomics* 2003. p. 442-446.
44. CUI, X. and CHURCHILL, G.A., *Data transformation for cDNA microarray data.* 2002.
45. YANG, Y., DUDOIT, S., LUU, P., and SPEED, T., *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* **Nuclear Acids Res**, 2002. 30 (4): p. e15.
46. CLEVELAND, W.S., *Robust Locally Weighted Regression and Smoothing Scatterplots.* **Journal of the American Statistical Association**, 1979. 74(368): p. 829-836
47. SHANBAHG, P.S., VISWANATH, V., and TORSEKAR, R.G., *Thalidomide: Current status.* **Indian J Dermatol Venereol Leprol**, 2006. 72(1): p. 75-80.
48. SAMPAIO, E., SARNO, E., GALILLY, R., COHN, Z., and KAPLAN, G., *Thalidomide selectively inhibits tumor necrosis factor alpha production by stimulated human monocytes.* **J. Exp. Med.**, 1991. 173(3): p. 699-703.
49. BAUER, K.S., DIXON, S.C., and FIGG, W.D., *Inhibition of Angiogenesis by Thalidomide Requires Metabolic Activation, Which Is Species-dependent.* **Biochemical Pharmacology**, 1998. 55: p. 1827-1834.
50. KEIFER, J.A., GUTTRIDGE, D.C., ASHBURNER, B.P., and ALBERT S. BALDWIN, J., *Inhibition of NF-kB Activity by Thalidomide through Suppression of IkB Kinase*

Activity. **THE JOURNAL OF BIOLOGICAL CHEMISTRY**, 2001. 276(25): p. 22382-22387.

51. GILMORE, T.D., *Introduction to NF-κB: players, pathways, perspectives.* **Oncogene**. 25(51): p. 6680-6684.
52. MOREIRA, A., SAMPAIO, E., ZMUIDZINAS, A., FRINDT, P., SMITH, K., and KAPLAN, G., *Thalidomide exerts its inhibitory action on tumor necrosis factor alpha by enhancing mRNA degradation.* **J. Exp. Med.**, 1993. 177(6): p. 1675-1680.
53. STEPHENS, T.D., BUNDE, C.J.W., and FILLMORE, B.J., *Mechanism of Action in Thalidomide Teratogenesis.* **Biochemical Pharmacology**, 2000. 59: p. 1489-1499.
54. HASLETT, P.A.J., CORRAL, L.G., ALBERT, M., and KAPLAN, G., *Thalidomide Costimulates Primary Human T Lymphocytes, Preferentially Inducing Proliferation, Cytokine Production, and Cytotoxic Responses in the CD8+ Subset.* **J. Exp. Med.**, 1998. 187(11): p. 1885-1892.
55. HUANG, F., GU, J., ZHAO, W., ZHU, J., ZHANG, J., and YU, D.T.Y., *One-Year Open-Label Trial of Thalidomide in Ankylosing Spondylitis.* **Arthritis & Rheumatism (Arthritis Care & Research)**, 2002. 47(3): p. 249-254.
56. VACCA, A., SCAVELLI, C., MONTEFUSCO, V., DI PIETRO, G., NERI, A., MATTIOLI, M., BICCIATO, S., NICO, B., RIBATTI, D., DAMMACCO, F., and CORRADINI, P., *Thalidomide Downregulates Angiogenic Genes in Bone Marrow Endothelial Cells of Patients With Active Multiple Myeloma.* **J Clin Oncol**, 2005. 23(23): p. 5334-5346.
57. ALLISON, D.B., CUI, X., PAGE, G.P., and SABRIPOUR, M., *Microarray data analysis: from disarray to consolidation and consensus.* **nature reviews genetics**, 2006. 7 p. 55-66.
58. CUI, X. and CHURCHILL, G., *Statistical tests for differential expression in cDNA microarray experiments.* **Genome Biology**, 2003. 4(4): p. 210.
59. YANG, Y.H., XIAO, Y., and SEGAL, M.R., *Identifying differentially expressed genes from microarray experiments via statistic synthesis.* **BIOINFORMATICS**, 2005. 21(7): p. 1084-1093.
60. TUSHER, V., TIBSHIRANI, R., and CHU, C., *Significance analysis of microarrays applied to transcriptional response to ionizing radiations.* **PNAS**, 2001. 98: p. 5116-5121.

61. BALDI, P. and LONG, A., *A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes.* **BIOINFORMATICS**, 2001. 17: p. 509-519.
62. LONNSTEDT, I. and SPEED, T., *Replicated microarray data.* **Statistica Sinica**, 2002. 12 p. 31.
63. KERR, M.K., MARTIN, M., and CHURCHILL, G.A., *Analysis of Variance for Gene Expression Microarray Data*. 2000, The Jackson Laboratory. p. %&.
64. VIDAVALUR, R., PENUMATHSA, S.V., THIRUNAVUKKARASU, M., ZHAN, L., KRUEGER, W., and MAULIK, N., *Sildenafil augments early protective transcriptional changes after ischemia in mouse myocardium.* **Gene**, 2009. 430 p. 30-37.
65. BREITLING, R. and HERZYK, P., *Rank-based methods as a non-paramatric alternative of the t-statistic for the analysis os biological microarray data.* **Journal of Bioinformatics and Computational Biology**, 2005. 3(5): p. 1171-1189.
66. ABERCROMBIE, J., HALFHILL, M., RANJAN, P., RAO, M., SAXTON, A., YUAN, J., and STEWART, C.N., *Transcriptional responses of Arabidopsis thaliana plants to As (V) stress.* **BMC Plant Biology**, 2008. 8(1): p. 87-92.
67. BENJAMINI, Y. and HOCHBERG, Y., *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* **J R Stat Soc B**, 1995. 57: p. 289-300.
68. PAVLIDIS, P., QIN, J., ARANGO, V., MANN, J.J., and SIBILLE, E., *Using the Gene Ontology for Microarray Data Mining: A Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex* **Neurochemical Research**, 2004. 29(6): p. 1213-1222.
69. SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V.K., MUKHERJEE, S., EBERT, B.L., GILLETTE, M.A., PAULOVICH, A., POMEROY, S.L., GOLUB, T.R., LANDER, E.S., and MESIROV, J.P., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* **PNAS**, 2005. 102(43): p. 15545-15550.
70. DUDOIT, S., YANG, Y.H., CALLOW, M.J., and SPEED, T.P., *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* 2000.

71. CUI, X., HWANG, J.T.G., QIU, J., BLADES, N.J., and CHURCHILL, G.A., *Improved statistical tests for differential gene expression by shrinking variance components estimates.* **Biostatistics**, 2005. 6(1): p. 59-75.
72. GENTLEMAN, R.C., CAREY, V.J., BATES, D.M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A.J., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J.Y., and ZHANG, J., *Bioconductor: open software development for computational biology and bioinformatics.* **Genome Biol**, 2004. 5: p. R80.
73. FALCON, S. and GENTLEMAN, R., *Using GOstats to test gene lists for GO term association.* **Bioinformatics**, 2007. 23(2): p. 257-258.
74. ALEXA, A., RAHNENFUHRER, J., and LENGAUER, T., *Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.* **Bioinformatics**, 2006. 22(13): p. 1600-1607.
75. YANG, Y., DUDOIT, S., LUU, P., LIN, D., PENG, V., NGAI, J., and SPEED, T., *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* **Nucleic Acids Research**, 2002. 30: p. e15.
76. WU, H., KERR, M.K., CUI, X., and CHURCHILL, G.A., *MAANOVA: a software package for the analysis of spotted cDNA microarray experiments*.
77. GENE ONTOLOGY CONSORTIUM, *Gene Ontology: tool for the unification of biology.* **nature genetics**, 2000. 25 (1): p. 25-29.
78. JENSSEN, T.-K., LANGAAS, M., KUO, W.P., SMITH-SORENSEN, B., MYKLEBOST, O., and HOVIG, E., *Analysis of repeatability in spotted cDNA microarrays.* **Nucl. Acids Res.**, 2002. 30(14): p. 3235-3244.
79. SAUER, U., PREININGER, C., and HANY-SCHMATZBERGER, R., *Quick and simple: quality control of microarray data.* **BIOINFORMATICS**, 2005. 21(8): p. 1572-1578.
80. WANG, X., GHOSH, S., and GUO, S., *Quantitative quality control in microarray image processing and data acquisition.* **Nucleic Acids Research**, 2001. 19: p. e75.
81. HUBER, W., HEYDEBRECK, A.V., SÜLTMANN, H., POUSTKA, A., and VINGRON, M., *Variance stabilization applied to microarray data calibration and to the*

quantification of differential expression. **BIOINFORMATICS**, 2002. 18(s1): p. s96-s104.

82. VINCIOTTI, V., KHANIN, R., D'ALIMONTE, D., LIU, X., CATTINI, N., HOTCHKISS, G., BUCCA, G., DE JESUS, O., RASAIYAAH, J., SMITH, C.P., KELLAM, P., and WIT, E., *An experimental evaluation of a loop versus a reference design for two-channel microarrays.* **BIOINFORMATICS**, 2005. 21(4): p. 492-501.
83. DURBIN, B.P. and ROCKE, D.M., *Variance-stabilizing transformations for two-color microarrays. .* **BIOINFORMATICS**, 2004. 20(5): p. 660-667.
84. KERR, M., MARTIN, M., and CHURCHILL, G., *Analysis of variance for gene expression microarray data.* **J Comput Biol**, 2000. 7: p. 819-837.
85. CHURCHILL, G.A., *Using ANOVA to Analyze Microarray Data* **Biotechniques**, 2004. 37(2): p. 173-177.
86. BENDER, R. and LANGE, S., *Adjusting for multiple testing - when and how?* **Journal of Clinical Epidemiology**, 2001. 54 p. 343-349.
87. SANTOS, N.C., MARTINS-SILVA, J.F.-C.J., and SALDANHA, C., *Multidisciplinary utilization of dimethyl sulfoxide:pharmacological,cellular, and molecular aspects.* **Biochemical Pharmacology**, 2003. 65: p. 1035-1041.
88. SHIMIZUA, S., SIMONA, R.P., and GRAHAM, S.H., *Dimethylsulfoxide (DMSO) treatment reduces infarction volume after permanent focal cerebral ischemia in rats.* **Neuroscience letters**, 1997. 239 (2-3): p. 125-127.
89. MAGLOTT, D., OSTELL, J., PRUITT, K.D., and TATUSOVA, T., *Entrez Gene: gene-centered information at NCBI.* **Nucleic Acids Research**, 2005. 33.
90. GLICKMAN, M.H. and CIECHANOVER, A., *The Ubiquitin-Proteasome Proteolytic Pathway: Destruction for the Sake of Construction.* **Physiol. Rev.**, 2002. 82(2): p. 373-428.
91. AMMOSOVA, T., WASHINGTON, K., DEBEBE, Z., BRADY, J., and NEKHAI, S., *Dephosphorylation of CDK9 by protein phosphatase 2A and protein phosphatase-1 in Tat-activated HIV-1 transcription.* **Retrovirology**, 2005. 2: p. 1-15.
92. DIAS, B.C., *Efeito da Talidomida na Expressão Gênica de Células Humanas Tratadas com Talidomida*, in *Programa de Engenharia Biomédica*. 2009, UFRJ: Rio de Janeiro.
93. TANAKA, T.S., JARADAT, S.A., LIM, M.K., KARGUL, G.J., WANG, X., GRAHOVAC, M.J., PANTANO, S., SANO, Y., PIAO, Y., and NAGARAJA, R., *Genome-wide expression*

profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. **Proc Natl Acad Sci USA**, 2000. 97: p. 9127-9132

94. NOLASCO, S., BELLIDO, J., GONÇALVES, J., ZABALA, J.C., and SOARES, H., *Tubulin cofactor A gene silencing in mammalian cells induces changes in microtubule cytoskeleton, cell cycle arrest and cell death.* **FEBS letters**, 2005. 579(17): p. 3515-3524.

95. SATRE, M., ZGOMBIC-KNIGHT, M., and DUESTER, G., *The complete structure of human class IV alcohol dehydrogenase (retinol dehydrogenase) determined from the ADH7 gene.* **J. Biol. Chem.**, 1994. 269(22): p. 15606-15612.