**UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
DEPARTAMENTO DE GENÉTICA**

***Pipeline* para análise *in silico* de dados de expressão de
miRNAs e mRNAs em células de mamíferos**

Luiz Fernando Martins Pignata

**Ribeirão Preto - SP
2012**

LUIZ FERNANDO MARTINS PIGNATA

***Pipeline* para análise *in silico* de dados de expressão de
miRNAs e mRNAs em células de mamíferos**

Dissertação de Mestrado apresentada ao
Programa de Pós-graduação em Genética da
Faculdade de Medicina de Ribeirão Preto,
para obtenção do Título de Mestre em
Ciências.

Área de Concentração: Genética.

Orientadora: Prof^a. Dr^a. Silvana Giuliatti

Ribeirão Preto - SP

2012

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Pignata, Luiz Fernando Martins

Pipeline para Análise In Silico de Dados de Expressão de miRNAs e mRNAs em Células de Mamíferos. Ribeirão Preto, 2012

81p.: il.;; 30cm.

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP – Área de concentração: Genética.

Orientadora: Giuliatti, Silvana

1.bioinformática, 2.*microarray*, 3.expressão gênica, 4.redes gênicas, 6.normalização

FOLHA DE APROVAÇÃO

Luiz Fernando Martins Pignata

Pipeline para Análise *In Silico* de Dados de Expressão de miRNAs e mRNAs em Células de Mamíferos

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Genética da Faculdade de Medicina de Ribeirão Preto/USP para obtenção do Título de Mestre em Ciências.

Área de concentração: Genética

Orientador: Prof^a. Dr^a. Silvana Giuliatti

Aprovado em: ____/____/____

Banca Examinadora

Prof.Dr. _____

Instituição: _____ Assinatura: _____

Prof.Dr. _____

Instituição: _____ Assinatura: _____

Prof.Dr. _____

Instituição: _____ Assinatura: _____

Dedicatória

*Dedico aos meus pais, J osé Augusto e
Maria L uiza, e ao meu irmão André que sempre
me incentivaram para que eu buscasse por meus sonhos,
com carinho e compreensão.*

*À minha amada esposa e companheira
Marcela, pelo seu amor incondicional, pela enorme
paciência me auxiliando a superar as dificuldades
encontradas e por sua compreensão.*

Amo vocês!

Agradecimientos

Agradeço a Deus em primeiro lugar, por tudo em minha vida.

Agradeço à Prof.^a Dr.^a Silvana Giuliatti por me aceitar no Grupo de Bioinformática e por todo seu apoio durante o desenvolvimento deste trabalho.

Agradeço ao Prof. Dr. Geraldo A. S. Passos pela colaboração e apoio para a realização deste trabalho.

Agradeço ao Departamento de Genética e à Faculdade de Medicina de Ribeirão Preto pela estrutura oferecida para a minha formação na pós-graduação.

Agradeço aos funcionários do Departamento de Genética, sempre atenciosos e dispostos a ajudar.

Agradeço ao CNPq pelo apoio financeiro.

Agradeço a todos os funcionários da FUNPEC-RP, pela compreensão, e especialmente ao Prof. Duarte pelo reconhecimento e ao lulo pelas traduções.

Agradeço a todos do Laboratório de Imunogenética Molecular, em especial à Cláudia, Adriane, Márcia e Shirley, pelo auxílio durante o desenvolvimento do projeto.

Agradeço a todos os amigos do Grupo de Bioinformática, Daniel, Saulo, André, Nilson, Pablo, Luciano, Iuliana, Roberta, Gabriela, Fernanda e Beatriz, pelo apoio e convívio durante esses anos.

Agradeço aos grandes amigos Daniel e Luiza, Saulo e Lara, Nilson e Mariana, André e Juliane, pela força nos momentos difíceis, companheirismo e amizade.

Agradeço aos amigos da Faculdade de Ciências Farmacêuticas, Anna Beatriz, Simone, Rodrigo, Silvia, Luciana, Cenzi, Luís Alexandre, Wellington, Samuel, Fábio, Eduardo e Fúlvia, pela amizade e incentivo.

Agradeço aos meus amigos de longa data, Luís Henrique, Anelisa, Francisco, Maria Eugênia, Flávia, Marília, Gustavo e Leandra pelos vários momentos

difíceis divididos, pela força, e por estarem há tantos anos comigo comemorando as nossas vitórias.

Agradeço aos queridos Flávia, José Antônio, Antônio Luís e Camila pelos conselhos, pela ajuda e pelo carinho de me aceitarem como parte de sua família.

Agradeço aos meus queridos pais José Augusto e Maria Luiza, e ao meu irmão André, pela força em todos os momentos, pela presença sempre necessária, e pelo amor que sempre houve em nossa família.

Agradeço à minha esposa Marcela, pela presença em todos os momentos fáceis ou difíceis, pelo amor compartilhado e incentivo total durante toda essa fase de nossa vida.

Enfim, agradeço a todos aqueles que, de alguma forma, me deram força e apoio para seguir em frente e concluir mais essa etapa em minha vida.

Muito obrigado!

*“É preciso força pra sonhar e perceber
que a estrada vai além do que se vê”.*

Marcelo Camelo

RESUMO

PIGNATA, L. F. M. ***Pipeline para análise in silico de dados de expressão de miRNAs e mRNAs em células de mamíferos.*** 2012. 81f. Dissertação (Mestrado). Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2012.

Os microRNAs estão envolvidos no processo de regulação da expressão gênica da célula, onde a molécula de microRNA se liga com o RNA mensageiro interrompendo, assim, a expressão do respectivo gene pela interrupção da tradução. A bioinformática tem auxiliado na identificação de vários genes codificadores de microRNAs em plantas e animais, incluindo mamíferos, por meio de análises de dados de *microarray*; assim como na predição de estruturas. Os dados de expressão de microRNAs e RNAs mensageiros foram obtidos por meio de cooperação firmada entre o Laboratório de Bioinformática do Departamento de Genética da Faculdade de Medicina de Ribeirão Preto - USP, coordenado pela orientadora desse projeto, e o Laboratório de Imunogenética Molecular do mesmo departamento, coordenado pelo Professor Doutor Geraldo A. S. Passos. Durante o desenvolvimento e os testes realizados, foram utilizados dados (valores numéricos de dados de expressão coletados por *microarrays*) provenientes da comparação da expressão de microRNAs e RNAs do timo de camundongos *non obese diabetic* que reproduzem diabetes melitus do tipo 1, e dados provenientes da comparação da expressão de microRNAs e RNAs de outros experimentos. O presente projeto teve como objetivo o desenvolvimento de um *pipeline* para a análise *in silico* de dados de expressão gênica de microRNAs e mRNAs obtidos por *microarray*. Com base em dados de expressão de microRNAs e RNAs mensageiros, foi possível a análise de diversas ferramentas e o desenvolvimento e ajuste de *scripts* para que seja possível a análise sequencial de tais dados. Dessa forma, o *pipeline* desenvolvido inclui a quantificação dos dados de expressão gênica a partir das lâminas de *microarray*, a normalização dos dados, as análises estatísticas das sequências diferencialmente expressas utilizando o *Multi Experiment Viewer*, a construção de redes de interação microRNAs-RNAs mensageiros e a busca de alvos de microRNAs baseada nesta rede, ambos pelo *GenMir++*. O *pipeline* desenvolvido é executado com facilidade e possibilitou a correta análise dos dados, evitando desperdício de tempo em análises de bancada. A partir dos resultados obtidos, novos alvos de miRNA foram encontrados com o uso do *pipeline* e comprovados em bancada. Tais resultados apresentados no 55º Congresso Brasileiro de Genética com o resumo intitulado “MicroRNA-mRNA Network Controlling the Promiscuous Gene Expression in the Thymus of NOD (Non Obese Diabetic) Mice: Implications in the Emergence of Type 1 Diabetes Mellitus”.

Palavras-chave: bioinformática, *microarray*, expressão gênica, redes gênicas, normalização.

ABSTRACT

PIGNATA, L. F. M. ***Pipeline for in silico analysis of miRNAs and mRNAs expression data in mammals cells***. 2012. 81f. Thesis (Master's Degree). Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2012.

The microRNAs are involved in the regulation of gene expression of the cell. The miRNA molecule binds to the messenger RNA and interrupts the gene expression by disrupting the translation. Through microarray data analysis, bioinformatics is a valuable aid for the identification of several genes that encode miRNAs in plants and animals, including mammals. It is also very useful for predicting structures. Data of miRNA and mRNA expression were obtained by the collaboration the Bioinformatics Laboratory and the Molecular Immunogenetics Laboratory of the Department of Genetics of the Faculty of Medicine of Ribeirão Preto - USP, coordinated by professors Silvana Giuliani and Geraldo A. S. Passos, respectively. During the development and tests of the research, microarrays data (numerical values os the expression) were obtained from the comparison between the expression of miRNA and mRNA of the thymus of *non obese diabetic* mice with diabetes mellitus type 1, as well as from comparisons of their expression in other experiments. The present study is aimed at the development of a *pipeline* for *in silico* analysis of the data of miRNAs and mRNA gene expression obtained by microarray. Based on miRNAs and mRNA expression, it was possible to analyze several tools, develop and adjust *scripts* that allowed the sequential analysis of such data. The *pipeline* includes the quantification of gene expression data from microarray, the normalization of the data, the statistical analysis of differentially expressed sequences using *Multi Experiment Viewer*, the construction of networks of interaction of miRNA-mRNAs, and the search for targets of miRNAs based on such network using *GenMir++*. The *pipeline* was performed easily and allowed the correct analysis of the data, avoiding waste of time in bench analysis. From the results, new targets of miRNA were found using the *pipeline* and were verified further in bench analysis. The results were presented in the 55th Brazilian Genetics Congress in the paper entitled "MicroRNA-mRNA Network Controlling the Promiscuous Gene Expression in the Thymus of NOD (Non Obese Diabetic) Mice: Implications in the Emergence of Type 1 Diabetes Mellitus"

Key words: bioinformatics, *microarray*, gene expression, gene networks, normalization.

LISTA DE FIGURAS

- Figura 1. Processamento do miRNA, partindo do gene dentro do núcleo até a repressão da tradução ou a quebra do mRNA (Traduzido de WINTER et al., 2009) 20
- Figura 2. Fluxograma das etapas realizadas no presente trabalho, ressaltando as etapas iniciais realizadas pelo Laboratório de Imunogenética Molecular. ... 28
- Figura 3. Esquema de desenvolvimento do pipeline. 31
- Figura 4. Exemplo de um spot e as diversas interações que podem ocorrer em uma lâmina de microarray (FERREIRA FILHO, 2009). 32
- Figura 5. Imagens de um experimento de microarray tratado por DUDOIT et al.(2002). A) Leitura do scanner com o comprimento de onda para a cor verde (Cy3). B) Leitura do scanner com o comprimento de onda para a cor vermelha (Cy5). O Banco de dados deste experimento , bem como suas imagens está disponível em <http://www.stat.berkeley.edu/users/terry/zarray/Html/apodata.html>. 33
- Figura 6. Resultado da sobreposição das imagens geradas pelo scaneamento da lâmina de microarray pelos dois canais de onda. (DUDOIT et al., 2002). 33
- Figura 7. na lâmina parece ser de um dedo, trazendo um grande problema para a análise deste microarray (ESTEVES, 2007). 36
- Figura 8. Esquema exemplificando matrizes de entrada para a ferramenta *GenMir++*. A) uma matriz Z com valores de expressão para os tecido/indivíduos analisados. B) uma matriz X com os valores de expressão de mRNA para os tecidos/indivíduos analisados. C) uma matriz C onde os números “1” (um) representam uma possível interação a ser estudada entre o miRNA e o mRNA, e os números “0” (zero) representam a não existência desta relação. 43
- Figura 9. Esquema representando as etapas do projeto: Aquisição das Imagens de microarray; Processamento das Imagens; Normalização; Agrupamento e Redes de Interação. 46
- Figura 10. Imagens exibindo a hibridação em microarrays. (A) com 4500 mRNAs de TSAs. (B) com 640 miRNAs. 48
- Figura 11. Parte do esquema exibindo a etapa de Processamento das imagens de *microarray*, sendo as lâminas de miRNA e mRNA processadas separadamente. 49

Figura 12. Exemplo de uma tela do software Spottfinder onde é realizada a definição da grade (que pode ser observada pelos quadriculados azuis em cada bloco) que demarca a posição de cada spot na lâmina (John C. Braisted, 2008).	50
Figura 13. Arquivos no formato TAV, com as informações de expressão de duas lâminas de um experimento utilizado neste trabalho.	50
Figura 14. Esquema representando a etapa de Normalização dos dados, onde os Dados Brutos de miRNA e mRNA são importados para montagem dos arquivos de saída com os dados normalizados.	52
Figura 15. Tela inicial do aplicativo R, onde o ponteiro em vermelho “aguarda” pelo código a ser executado.	53
Figura 16. Início da execução do script de normalização, após os avisos a solicitação do caminho para a “Pasta de Trabalho”.	54
Figura 17. Gráfico de densidade dos canais verde e vermelho, sem normalização.	55
Figura 18. Linhas de código executando as funções de normalização pré-definidas pelas bibliotecas da plataforma R.	55
Figura 19. Gráfico de densidade dos canais verde e vermelho, após normalização.	56
Figura 20. Esquema da etapa de Agrupamento, com os dendogramas gerados pelo <i>Software</i> MEV, a esquerda para miRNA e a direita para mRNA.	58
Figura 21. Dendograma do experimento com miRNA, exibindo o agrupamento entre os indivíduos pré-auto imunes e os indivíduos auto imunes.	60
Figura 22. Esquema representando a etapa de geração das Redes de Interação.	61
Figura 23. Gráfico com os valores de beta_gk considerando em um experimento, todos os mRNAs como possíveis alvos de todos os miRNAs. Neste experimento gerado se tem 22 miRNAs e para cada miRNA se tem 90 mRNAs (colunas).	64
Figura 24. Gráfico com os valores de beta_gk considerando em um experimento, apenas os mRNAs assumidos como possíveis alvos de miRNAs. Neste experimento gerado se tem 22 miRNAs e para cada miRNA se tem 90 mRNAs (colunas)	66

Figura 25. Rede Gênica com 18 miRNAs (círculos) e 58 mRNAs (quadrados) codificadores de TSA em camundongos NOD	67
--	----

LISTA DE ABREVIATURAS E SIGLAS

Cy3	Marcador <i>Cyanine 3</i>
Cy5	Marcador <i>Cyanine 5</i>
DNA	<i>Deoxyribonucleic acid</i>
Kb	Kilobase
LIMMA	<i>Linear models for microarray data</i>
MEV	<i>MultiExperiment Viewer</i>
miRNA	Micro RNA
mRNA	RNA mensageiro
nm	Nanômetros
NOD	<i>Non obese diabetic</i>
nt	Nucleotídeos
Pri-miRNA	Micro RNA primário
RNA	<i>Ribonucleic acid</i>
SAM	<i>Significance Analysis of Microarray</i>
TAV	Formato de arquivo “ <i>TIGR Array Viewer</i> ”
TIFF	Formato de arquivo de imagem
TSAs	<i>Tissue specific antigen</i>
USP	Universidade de São Paulo
UTR	<i>Untranslated Regions</i>

SUMÁRIO

RESUMO.....	xi
ABSTRACT	xii
LISTA DE FIGURAS	xiii
LISTA DE ABREVIATURAS E SIGLAS	xvi
1. Introdução	19
1.1. microRNA.....	19
1.2. <i>Microarray</i>	22
1.3. Bioinformática	23
2. Objetivos	26
2.1. Objetivo Geral	26
2.2. Objetivos Específicos:	26
3. Material e Métodos.....	28
3.1. Obtenção dos dados de expressão gênica por <i>microarrays</i>	29
3.2. <i>Pipeline</i> para análise <i>in silico</i> dos dados obtidos através do experimento de <i>microarray</i>	30
3.2.1. Processamento das imagens.....	31
3.2.2. Normalização dos dados	36
3.2.3. Agrupamento dos dados de expressão	39
3.2.4. Redes Gênicas	41
4. Resultados e Discussão.....	46
4.1. Obtenção dos dados de expressão gênica por <i>microarray</i>	47
4.2. Processamento das imagens	48
4.3. Normalização dos Dados	51
4.4. Agrupamento dos dados de expressão.....	57
4.5. Redes Gênicas.....	60
4.6. Considerações Finais.....	68
5. Conclusão	71
REFERÊNCIAS BIBLIOGRÁFICAS	73
APÊNDICE	78

/ ntrodução

1. Introdução

1.1. microRNA

Descritos pela primeira vez por Lee e colaboradores, em 1993 (LEE; FEIRAUM; AMBROS, 1993), os microRNAs (miRNAs) pertencem a uma classe descoberta de pequenas moléculas de RNA, as quais possuem, aproximadamente, 22 nucleotídeos (nt) e são transcritos como partes de moléculas maiores de vários kilobases (kb).

Os miRNAs estão envolvidos no processo de regulação da expressão gênica da célula. Essa regulação ocorre através da ligação da molécula de miRNA com o RNA mensageiro (mRNA) interrompendo, assim, sua tradução e impedindo a expressão de genes nos organismos que possuem essa estrutura (LEE; FEIRAUM; AMBROS, 1993). Eles têm sido identificados numa variedade de organismos (BARTEL, 2004; HE E HANNON, 2004; SEVIGNANI et al., 2006). Em células de animais, os miRNAs simples hélice, se ligam a mRNAs por meio de complementaridade, principalmente nas regiões 3' UTR (*untranslated regions*).

Os miRNAs são estruturas codificadas por sequências que são transcritos como transcritos primários (pri-miRNA), possuem cauda de poli-A e *cap*, mas que não são traduzidos em proteínas e sim, processados. No caso de animais, os pri-miRNA são processados por meio de uma ribonuclease RNA-específica chamada *Drosha*, dando origem ao pré-miRNA. No caso de plantas, esse procedimento é realizado pela proteína *Pasha* (LIN et al., 2010; TORRES et al., 2011).

Os pré-miRNAs (70-100nt) são, então, processados para miRNAs maduros (21-23nt) no citoplasma. Esse processo é realizado por uma ribonuclease específica denominada *Dicer*. O miRNA maduro representa a estrutura final e ativa. Em animais, os miRNAs atuam, negativamente, na regulação pós transcricional de um determinado gene formando um duplex de RNA entre o mRNA e o miRNA. O miRNA, através da sua complementaridade de bases, interrompe a tradução do mRNA, em geral, pela região 3'UTR. Um miRNA pode ser complementar a um ou mais mRNAs (LEE; FEIRAUM; AMBROS, 1993; LIN et al., 2010; TORRES et al., 2011). O processamento dos miRNAs pode ser observado no esquema exibido na Figura 1.

Figura 1. Processamento do miRNA, partindo do gene dentro do núcleo até a repressão da tradução ou a quebra do mRNA.
Fonte: Traduzido de WINTER et al. 2009.

Estas estruturas, de acordo com determinadas características, podem ser usadas para identificar novos miRNAs a partir de miRNAs já conhecidos (CHIROMATZO et al., 2007). Diversas pesquisas têm mostrado a importância dessas moléculas no organismo, desde o início do seu desenvolvimento

(ROMERO et al., 2008), pois uma característica fundamental do seu estudo é a sua conservação, mesmo em organismos geneticamente distantes (AMBROS, 2004).

O estudo dessas estruturas tem contribuído bastante para compreensão dos processos análogos que são regulados pelos miRNAs na maioria dos organismos vivos existentes (ENRIGHT et al., 2003; LIM et al., 2003; HE et al., 2005; WINTER et al., 2009).

A função do miRNA parece estar na regulação gênica. Por um mecanismo ainda não totalmente entendido, os mRNAs anelados (chamados alvos) não podem ser traduzidos nas suas respectivas proteínas, ou seja, a expressão do gene é controlada. Alternativamente, o mRNA alvo, anelado ao seu miRNA, pode ser degradado, reduzindo os níveis deste transcrito.

Há um grande interesse no entendimento do mecanismo de ação e na identificação de alvos de miRNAs. A pesquisa de identificação de alvos de miRNAs, em células de mamíferos, atrai grupos de pesquisa, principalmente na área de câncer (LIM et al., 2005; RAJEWSKI, 2006). O estudo realizado por He L. e colaboradores (HE et al., 2005) mostra que a expressão do *cluster-17-92* atua em conjunto com o gene *cmyc* para diminuir a velocidade do desenvolvimento de tumores em células B de ratos. Entretanto, na área de autoimunidade o assunto ainda é bastante original.

Uma maneira de entender como os miRNAs funcionam é determinando quando e onde eles estão expressos, da mesma forma como é feito com genes que codificam proteínas.

Uma variedade de adaptações tem sido feita com as tecnologias existentes para a detecção de pequenos RNAs. Dentre elas, um número de

métodos baseados em *microarray* tem sido desenvolvido (KRICHEVSKY et al., 2003).

Tecnologia de arranjos de expressão gênica tem sido adaptada para examinar miRNAs e tem revelado padrões de expressão distintos em diferentes estágios de desenvolvimento ou, como por exemplo, de regiões do cérebro de mamífero (KRICHEVSKY et al., 2003; CROUSER et al., 2011)

1.2. Microarray

Os *microarrays*, também conhecidos como “biochips de DNA”, são arranjos de DNA distribuídos sequencialmente sobre uma superfície sólida - no caso uma lâmina de vidro, semelhante às lâminas de microscópio – em posições específicas, onde cada ponto é chamado de *spot* e cada *spot* contém uma sequência representativa de um gene, sendo possível a utilização desta tecnologia para analisar perfis de expressão gênica em larga escala, podendo haver de dezenas a centenas de milhares de *spots* em uma lâmina.

Amostras de mRNA são copiadas em cDNA, que recebem marcações por fluorocromos (Cy3 e Cy5) e são hibridados com o *microarray*. Na sequência, as lâminas são levadas a um *scanner*, onde suas imagens são adquiridas. Baseando-se na intensidade da excitação do fluorocromo em cada *spot*, é possível quantificar o valor da expressão encontrada para aquele determinado gene (HARRINGTON et al., 2000; AGBAVWE et al., 2011).

Com a tecnologia de *microarray* é possível obter uma identidade molecular da expressão gênica de um determinado tecido em um determinado momento, ou seja, com ela, é possível determinar padrões de expressão para tecidos doentes ou saudáveis, sendo extremamente valiosa para o

entendimento a nível molecular de certas patologias. Essa identidade molecular é conhecida como “assinaturas de expressão gênica”.

1.3. Bioinformática

Métodos computacionais têm sido desenvolvidos para auxiliarem na análise e estudo de dados de expressão gênica e para complementarem métodos experimentais na identificação de miRNAs e de seus alvos.

A bioinformática tem auxiliado na identificação de vários genes codificadores de miRNAs em plantas e mamíferos (ENRIGHT et al., 2003; WINTER et al., 2009). Buscas de homólogos têm revelado ortólogos e parálogos de conhecidos miRNAs.

Assim como para análise de *microarray*, várias ferramentas de bioinformática têm sido desenvolvidas para auxiliarem nos estudos que envolvem a busca por novos miRNAs e seus alvos. Vários algoritmos têm sido desenvolvidos para predição de miRNAs e seus alvos em animais, como miRanda, TargetScan, entre outros (ENRIGHT et al., 2003; WINTER et al., 2009). Há também uma variedade de *softwares* para predição de estruturas tais como MirFold (BILLOUD et al., 2005).

A bioinformática tem também auxiliado na construção de redes gênicas, como no caso da ferramenta GeneNetwork (WU et al., 2004). As redes gênicas auxiliam na compreensão do complexo sistema de organismos vivos. Entretanto, com a chegada dos miRNAs e seus alvos, viu-se a necessidade de estudar a expressão em pares, ou seja, dados de pares de expressão gênica de miRNAs e mRNAs. Perfis de expressão de miRNAs e mRNAs devem revelar uma relação inversa entre o perfil de expressão gênica do miRNA e a de seus alvos.

A ferramenta *GenMir++*, desenvolvida por Huang e colaboradores (HUANG et al., 2007), mostra que perfis de pares de expressão gênica de miRNAs-mRNAs podem ser usados em larga escala para melhorar a precisão da predição de alvos de miRNAs.

Como dito acima, a bioinformática tem auxiliado muito os pesquisadores em suas análises, principalmente naquelas que envolvem grande quantidade de dados, como em análises de expressão gênica. Esse também é o caso dos projetos desenvolvidos no Laboratório de Imunogenética Molecular da Faculdade de Medicina de Ribeirão Preto - USP, coordenado pelo Professor Doutor Geraldo A. S. Passos.

Em um de seus projetos, intitulado “Avaliação do papel dos microRNAs no controle da expressão gênica promíscua no timo de camundongos NOD (*Non Obese Diabetic*)” (2006/56741-5), o grupo realiza um estudo comparativo da expressão gênica promíscua de antígenos específicos de tecido (TSAs) no timo de camundongos da linhagem NOD que reproduz o diabetes melitus do tipo 1, durante a transição do estado pré auto imune para auto imune. A hipótese levantada pelo grupo é a de que existem mensageiros de TSAs que são alvos de miRNAs e que, uma vez desregulados no timo, poderão influenciar o surgimento da autoimunidade.

Para isso, os pesquisadores do grupo geraram dados de expressão gênica tanto de miRNAs como de mRNAs, com o objetivo de avaliar a participação dos miRNAs na autoimunidade como controladores da expressão gênica promíscua. Sendo necessário para isso o desenvolvimento de um *pipeline* para realizar a análise dos dados gerados.

Objetivos

2. Objetivos

2.1. Objetivo Geral

O presente trabalho teve como objetivo geral o desenvolvimento de um *pipeline* para a análise *in silico* de dados de expressão gênica de microRNAs e mRNAs obtidos por *microarray*.

2.2. Objetivos Específicos:

- Desenvolvimento do *pipeline* para a análise dos dados de expressão gênica de microRNA;
- Desenvolvimento e ajuste de *scripts* de normalização;
- Geração de uma rede de interação entre microRNAs e RNAs mensageiros;
- Utilização da rede como uma ferramenta de predição de novos alvos de microRNAs;
- Criação do *pipeline* para a análise dos dados tanto de microRNA quanto de RNA mensageiro desde a aquisição da imagem até a montagem da rede de interação.

Material e Métodos

3. Material e Métodos

Neste item será apresentada a metodologia utilizada para a realização deste trabalho. Um esquema simplificado das etapas realizadas é apresentado na Figura 2.

Figura 2. Fluxograma das etapas realizadas no presente trabalho, ressaltando as etapas iniciais realizadas pelo Laboratório de Imunogenética Molecular.

A partir de uma parceria firmada entre o Laboratório de Bioinformática do Departamento de Genética da Faculdade de Medicina de Ribeirão Preto - USP, coordenado pela orientadora deste projeto, Professora Doutora Silvana Giuliatti, e o Laboratório de Imunogenética Molecular do mesmo departamento, coordenado pelo Professor Doutor Geraldo A. S. Passos, o presente projeto pôde ser dividido em duas etapas principais.

Como pode ser observada no esquema representado na Figura 2, a etapa de “Obtenção dos dados de expressão por *microarray*” foi realizada pelo Laboratório de Imunogenética Molecular, onde o grupo deste laboratório realizou as etapas de montagem das lâminas e experimentos de *microarray*, bem como a aquisição de suas imagens, sendo fornecidos para este trabalho os dados resultantes das etapas acima mencionadas.

O “*Pipeline* para análise *in silico* dos dados obtidos através do experimento de *microarray*” foi desenvolvido no Laboratório de Bioinformática, a partir dos dados concedidos pelo Laboratório de Imunogenética Molecular. Assim, foi possível o desenvolvimento de novas ferramentas e o ajuste de ferramentas já utilizadas para a conclusão do presente trabalho.

3.1. Obtenção dos dados de expressão gênica por *microarrays*

Em um dos projetos do Laboratório de Imunogenética Molecular, intitulado “Avaliação do papel dos microRNAs no controle da expressão gênica promíscua no timo de camundongos NOD (*Non Obese Diabetic*)” (2006/56741-5), o grupo realizou um estudo comparativo da expressão gênica promíscua de antígenos específicos de tecido (*tissue-specific antigen*, TSAs) no timo de camundongos da linhagem NOD que reproduz o diabetes melitus do tipo 1, durante a transição do estado pré auto imune para auto imune. A hipótese

levantada pelo grupo é a de que existem mRNAs de TSAs que são alvos de miRNAs e que, uma vez desregulados no timo, poderão influenciar o surgimento da autoimunidade.

Os pesquisadores do grupo geraram dados de expressão gênica tanto de miRNAs como de mRNAs, com o objetivo de avaliar a participação dos miRNAs na autoimunidade como controladores da expressão gênica promíscua.

Utilizando dos dados de expressão gênica deste experimento e de alguns outros experimentos, foi possível dar início ao desenvolvimento deste trabalho. Sendo os dados obtidos de valores de expressão de experimentos em lâminas de vidro, com marcadores *bicolor* e números variados de amostras. A seguir serão expostas as etapas da análise *in silico* realizadas já com o auxílio do grupo do Laboratório de Bioinformática.

3.2. Pipeline para análise *in silico* dos dados obtidos através do experimento de *microarray*

Os valores numéricos (quantificação dos valores de expressão) apresentados pelos experimentos de miRNA são considerados muito baixos comparados com os de mRNA, isso ocorre, principalmente porque os miRNAs são cadeias de aproximadamente 22 nucleotídeos, enquanto que os mRNAs possuem de milhares a milhões de nucleotídeos, diminuindo drasticamente a intensidade de hibridação das amostras de miRNA, tornando os valores quantificados muito mais baixos.

Devido a este fato, as análises dos dados de *microarray* de miRNA não devem ser realizadas utilizando-se dos mesmos protocolos utilizados para os dados de *microarray* de mRNA. Em cada etapa serão discutidas as

ferramentas do protocolo já utilizado para experimentos com mRNA e as ferramentas propostas por este trabalho para o novo protocolo para experimentos com miRNA.

A análise *in silico* foi dividida em algumas etapas, como pode ser visto no esquema mostrado na Figura 3. Em cada uma destas etapas foram estudadas mais de uma solução, sendo que todos os programas utilizados são de domínio público, e puderam ser executados localmente ou via *web*.

Figura 3. Esquema de desenvolvimento do pipeline.

3.2.1. Processamento das imagens

Durante a hibridação, as moléculas marcadas com Cy3 / Cy5 se anelam às moléculas complementares fixadas nos *spots* da lâmina. A região do *spot* onde a hibridação acontece é chamada de *foreground*, e a área da lâmina onde não se encontram os *spots* é chamada de *background*. Algumas vezes as moléculas nos *spots* podem se ligar a outros genes, dando origem a uma

hibridação cruzada. A intensidade deste tipo de hibridação é chamada de *ruído*. Em outros casos pode ocorrer a hibridação fora da área delimitada pelo *spot*, ou seja, no *background*. Este tipo de interação é chamado de *artefato*. Estes termos são melhores exemplificados na Figura 4.

Figura 4. Exemplo de um spot e as diversas interações que podem ocorrer em uma lâmina de microarray.

Fonte: FERREIRA FILHO, 2009.

A presença de *ruídos* e *artefatos* são considerados problemas comuns em experimentos de *microarray*, porém tais problemas podem levar a resultados indesejados no decorrer do experimento, existindo a possibilidade de um *spot* ser super ou supra quantificado devido a presença destes problemas.

Após as lâminas terem sido marcadas e hibridadas, foram levadas ao *scanner* para a digitalização da imagem do *microarray*. O *scanner* faz a varredura da lâmina com um raio laser gerando uma imagem para cada um dos canais de cores (Figura 5A e 5B), sendo emitido o comprimento de onda de excitação do laser de 532nm para o canal verde e 633nm para o canal vermelho (AGBAVWE et al., 2011).

Figura 5. Imagens de um experimento de microarray tratado por DUDOIT et al.(2002). A) Leitura do scanner com o comprimento de onda para a cor verde (Cy3). B) Leitura do scanner com o comprimento de onda para a cor vermelha (Cy5).

Fonte: DUDOIT et al., 2002. O Banco de dados deste experimento , bem como suas imagens está disponível em <http://www.stat.berkeley.edu/users/terry/zarray/Html/apodata.html>.

Após o processo de aquisição das imagens em ambos os canais, as imagens são sobrepostas, computacionalmente, para que assim seja gerada a imagem colorida de expressão do *microarray* (Figura 6).

Figura 6. Resultado da sobreposição das imagens geradas pelo scaneamento da lâmina de microarray pelos dois canais de onda.

Fonte: DUDOIT et al., 2002.

As imagens obtidas devem ser analisadas e quantificadas para que sejam encontrados os valores de expressão. A análise da imagem faz-se necessária também para a remoção de problemas comuns, como a detecção de *ruídos*, *artefatos* e correção do *background*. Nessa análise também são levados em consideração problemas na impressão do *chip* e sinais fracos. Portanto, a definição da grade dos *spots* e a correção de *background* devem ser realizadas.

Como dito anteriormente, devido à grande diferença nos valores de expressão, a análise de dados provenientes de experimentos de *microarray* de miRNA não deve ser realizada seguindo os protocolos adotados para a análise de dados de mesmo experimento com mRNA. Sendo assim, para a análise das imagens geradas e quantificação dos experimentos foram testados dois *softwares*: *BZScan* (LOPEZ et al., 2004) e *Spotfinder* (SAEED et al., 2003).

O *software* BZScan é uma plataforma completa para extração e quantificação de dados de *microarray*. Além da análise baseada em uma figura do tipo TIFF, o *software* BZScan permite o processamento de diversas imagens através da criação de *batches* (sequência de execução), podendo assim processar até 300 imagens em sequência.

A princípio, o *software* BZScan mostrou-se uma boa opção para quantificar as imagens provenientes de miRNA, pois seu desenvolvimento visa a quantificação de imagens com baixa expressão, porém, apesar de sua alta sensibilidade na detecção das imagens, rapidez e versatilidade, o BZScan se mostrou especialmente desenvolvido para dados de *microarray* obtidos em *nylon* combinado com marcação radioativa de alvos de mRNA (LOPEZ et al.,

2004), dessa forma, não sendo a melhor opção para análise das lâminas montadas para o experimento.

O *Spotfinder* é um *software* que faz parte de um conjunto de aplicativos para análise e manipulação de dados de *microarray* desenvolvido pelo Instituto de Pesquisa Genômica - *The Institute for Genomic Research*, TIGR. Tem como entrada arquivos de imagem - no formato TIFF - gerados pela maioria dos *scanners* de *microarray*, e como saída, um arquivo no formato TAV - TIGR ArrayViewer.

Após a aquisição da imagem, é possível fazer o alinhamento manual ou automático das grades de leitura dos pontos – *spots* - e, a partir dessa etapa, é realizada a remoção do *background* e a quantificação dos dados adquiridos (SAEED et al., 2003).

Os experimentos realizados pelo Laboratório de Imunogenética Molecular, com *microarrays* de mRNA, utilizam em seu protocolo de análise o *software Spotfinder*, fato este que reforçou a escolha do mesmo *software* para a análise dos dados de miRNA, apesar da grande diferença nos valores de expressão.

Outro ponto que foi levado em consideração para a escolha do *Spotfinder*, como *software* de análise de imagens, foi o fato de que sendo utilizado o mesmo *software* para a quantificação dos experimentos de miRNA e de mRNA, os dados de saída são apresentados no mesmo formato e, com isso, a comparação e análise entre os mesmos é facilitada, evitando conversões entre formatos.

Com ambos os dados, tanto de miRNA quanto de mRNA, os arquivos com valores brutos de expressão, quantificados e não normalizados, dos dois experimentos foram utilizados para a próxima etapa deste trabalho.

3.2.2. Normalização dos dados

A etapa de normalização dos dados visa a remoção de variações técnicas, não referentes à hibridação dos genes, a fim de tornar os dados comparáveis. Durante a realização do experimento, fatores físicos, químicos, internos ou externos ao experimento podem causar diferenças entre uma lâmina de *microarray* e outra, ou até mesmo dentro de uma mesma lâmina, como por exemplo, as diferenças da incorporação dos corantes, falhas nas impressões das lâminas, imprecisão dos equipamentos, entre outros (SMYTH e SPEED, 2003).

A Figura 7 mostra uma lâmina com problemas em sua distribuição, possivelmente com a marca de uma impressão digital causando uma grande mancha sobre as reações hibridadas nos *spots*.

Figura 7. na lâmina parece ser de um dedo, trazendo um grande problema para a análise deste microarray.

Fonte: ESTEVES, 2007.

O que se pretende com a realização da normalização, é fazer com que todos os pontos dos experimentos sejam levados ao mesmo nível de variações, o que proporcionará a correção de problemas como o do tipo ocorrido na Figura 7. Para isso a utilização da bioinformática e de processos computacionais e matemáticos é de extrema importância.

Os dados gerados pela saída do *software Spotfinder* foram tratados de acordo com seu experimento: os dados gerados pelo *microarray* de miRNA, e, separadamente, os dados gerados pelo *microarray* de mRNA. Esta separação deve ser realizada na etapa de normalização devido a grande diferença entre os valores de expressão de miRNA e os de mRNA (RAO et al., 2008).

Para a normalização dos dados foi utilizada a plataforma R ([HTTP://www.r-project.org](http://www.r-project.org)), baseada nas bibliotecas para análise de dados de experimentos de *microarray* que são: KTH ([HTTP://www.biotech.kth.se/molbio/microarray](http://www.biotech.kth.se/molbio/microarray)), Aroma (BENGTSSON, 2004) e LIMMA (GENTLEMAN et al., 2004). Cada uma destas bibliotecas possui funções básicas específicas para a realização da normalização do experimento.

Os dados provenientes das imagens do *microarray* são transformados, no ambiente R, em matrizes numéricas de M versus A, sendo:

$$M = \log_2(R/G) \quad (1)$$

$$A = \frac{1}{2} \cdot \log_2(R \cdot G) \quad (2)$$

onde R é a intensidade de expressão medida pelo *scanner* no canal “vermelho” e G é a intensidade de expressão medida pelo *scanner* pelo canal “verde”. Com isso o valor de M representa a diferença entre os valores de

expressão de cada canal para cada *spot*, e A representa o valor médio da expressão.

Estes dados, já são a representação do experimento, uma vez que já é possível observar a relação entre genes diferencialmente e não diferencialmente expressos. Sendo assim, através destes valores de M e A, os dados são submetidos à normalização intra e entre *arrays*.

Diversos métodos são utilizados para normalização de dados de *microarray*, incluindo o método Quantil, *print-tip lowess* (BOLSTAD et al., 2003), média e mediana (CHURCHILL et al., 2003; DAVISON et al., 2006), dentre outros.

O protocolo padrão para análise de dados de *microarray* utilizado pelo Laboratório de Imunogenética Molecular, traz como normalização para seus experimentos com mRNA dois métodos: “*print-tip Lowess*” e “*median absolute deviation*” (MAD). O método “*print-tip Lowess*” é utilizado para normalização dentro de cada *array*, corrigindo erros que possam ter sido gerados durante o experimento aplicando regressões lineares aos dados, e o método MAD para a normalização entre os *arrays* fazendo com que todos os *arrays* apresentem seus valores dentro de um mesmo padrão.

Baseado nos experimentos anteriores do Laboratório de Imunogenética Molecular e nas informações encontradas na literatura para a normalização de dados de *microarray* de mRNA (SMYTH e SPEED, 2003; DAVISON et al., 2006), foi mantida a normalização já utilizada anteriormente, através de “*print tip lowes*” e MAD.

Diferente dos métodos utilizados para a normalização de mRNA, o método quantil apresenta resultados mais adequados para normalização de dados de *microarray* de miRNA, de acordo com a literatura (RAO et al., 2008).

A normalização baseada no protocolo do Laboratório de Imunogenética Molecular é feita através de um *script* escrito para a plataforma R. Com base na metodologia adotada por Rao, Y. e colaboradores (RAO et al., 2008), foi desenvolvido um novo *script*, para a normalização dos dados de miRNA, utilizando a normalização por Quantil para tais dados.

Para agilizar e facilitar a normalização, o *script* original utilizado pelo Laboratório de Imunogenética Molecular foi adaptado. Unindo o novo *script* de normalização de dados de miRNA ao *script* original, foi possível realizar a normalização de ambos os dados em um mesmo *script*, cada um com o método mais adequado ao tipo de dado.

Então, o *script* de normalização recebe como entrada os dados dos dois experimentos, sendo solicitado ao usuário a informação do local onde os arquivos gerados pelo *Spotfinder* com os dados quantificados, tanto de miRNA quanto de mRNA, se encontram.

3.2.3. Agrupamento dos dados de expressão

Após a normalização dos dados, tanto de miRNA quanto de mRNA, as informações foram levadas ao *software Multi Experiment Viewer* (MEV), também desenvolvido pela TIGR.

O MEV é um *software* com licença gratuita desenvolvido especialmente para realizar diversas análises através de dados genômicos em larga escala. Possui diversos módulos de algoritmos desenvolvidos para análises

estatísticas, visualização e mineiração de dados de *microarray*, descoberta de padrões biológicos e agrupamento de dados.

A escolha do MEV para esta etapa do trabalho foi baseada na versatilidade apresentada pelo *software*. Fazendo parte do mesmo pacote de aplicativos para análise de *microarray* da TIGR, a sequência de análise vinda do *Spotfinder* (também da TIGR) se mostra mais coesa, sendo os formatos de dados trabalhados mais compatíveis.

Outro ponto para a utilização do MEV é a facilidade de uso e amplitude de funções para análise apresentadas pelo *software*. É possível executar diversas ferramentas pelo MEV simplesmente a partir de botões, sem a necessidade da utilização de linhas de comando, ferramentas importantes como agrupamento hierárquico, Teste *t*, *Significance Analysis of Microarray* (SAM), dentre outros.

Nesta etapa foi realizado o agrupamento dos dados de *microarray*, já normalizados. Este agrupamento é de grande importância para a análise dos dados gerados, uma vez que é montado um dendograma baseado nos dados de expressão, seus genes e tecidos.

Já utilizado pelo protocolo do Laboratório de Imunogenética Molecular para dados de mRNA, o MEV se mostrou também adequado aos dados de miRNA, uma vez que o aplicativo se adéqua ao escopo dos dados com os quais está trabalhando.

O MEV foi empregado para a análise estatística de padrões de expressão, onde, através do algoritmo *Significance Analysis of Microarray* (SAM), foi aplicado o método Cluster & TreeView, utilizando tanto para os dados de mRNA quanto para os de miRNA um valor de “*false discoverde rate*”

(FDR) maior ou igual a 5%. Com base nestes dados, o aplicativo construiu um dendograma agrupando genes e amostras, sendo assim possível a visualização de tais padrões de expressão.

Vale a pena ressaltar que os dados de mRNA e miRNA foram analisados separadamente, gerando assim dois dendogramas diferentes, porém seguindo os mesmos padrões como resultados.

3.2.4. Redes Gênicas

Como sequência deste projeto foi dado início ao trabalho com redes gênicas, focando na interação entre os miRNAs e seus possíveis alvos (mRNA).

O programa GeneNetwork (WU et al., 2004) é amplamente utilizado para a geração de redes gênicas de mRNAs. Através de dados resultantes de experimentos de *microarray*, normalizados e tratados, é possível montar uma rede de interação entre estes mRNA. Porém, o intuito deste trabalho é a montagem de uma rede de interação de miRNA com mRNA. Dessa forma, alguns testes foram realizados sem sucesso, visto que o GeneNetwork gera as redes baseadas em valores numéricos, e comparativamente, como já explicado, os valores de expressão de miRNA, mesmo quando altamente expressos, são muito menores que os de mRNA.

Dessa forma buscou-se um novo programa para o desenvolvimento desta rede, e com base no estudo de Huang e colaboradores foi selecionado o programa *GenMir++* (HUANG et al., 2007).

A ferramenta *GenMir++*, desenvolvida por Huang e colaboradores (HUANG et al., 2007), é um algoritmo matemático e estatístico executado a partir do ambiente MatLab. O *GenMir++* mostra que perfis de pares de

expressão gênica de miRNAs-mRNAs podem ser usados em larga escala para melhorar a precisão da predição de alvos de miRNAs.

O MatLab é um ambiente de alto nível computacional desenvolvido para a análise de dados, desenvolvimento de algoritmos, computação matemática e estatística. No caso deste projeto, o ambiente computacional do MatLab foi utilizado para a execução do algoritmo matemático *GenMir++*.

Através de dados de expressão gênica de miRNA e mRNA, o *GenMir++* é capaz de inferir relações entre miRNA e seus possíveis alvos. Baseando-se na literatura sobre a ferramenta (HUANG et al., 2007), o *script* de normalização foi adaptado de forma a fornecer como saída as três matrizes principais de entrada para a inferência das relações pelo *GenMir++*.

Para relacionar os miRNAs com seus possíveis alvos, o *GenMir++* necessita de três dados principais: uma matriz X contendo os valores de expressão, já normalizados, de mRNA por tecido ou indivíduo analisado; uma matriz Z contendo os valores de expressão, já normalizados, de miRNA por tecido ou indivíduo analisado, e uma matriz C de valores um e zero. A matriz C relaciona os miRNAs e os mRNAs, tendo em seu eixo x os miRNAs e em seu eixo y os mRNAs, sendo que os valores das células desta matriz é “1” (um), quando houver uma possível relação miRNA-mRNA e “0” (zero) quando não houver essa relação (HUANG et al., 2007).

Um esquema exemplificando estas matrizes pode ser observado na Figura 8 a seguir:

	TECIDO 01	TECIDO 02		TECIDO 01	TECIDO 02
micro RNA 01	Valor(1,1)	Valor(1,2)	mRNA 01	Valor(1,1)	Valor(1,2)
micro RNA 02	Valor(2,1)	Valor(2,2)	mRNA 02	Valor(2,1)	Valor(2,2)
micro RNA 03	Valor(3,1)	Valor(3,2)	mRNA 03	Valor(3,1)	Valor(3,2)

A) Matriz Z

B) Matriz X

	micro RNA 01	micro RNA 02	micro RNA 03
mRNA 01	1	0	0
mRNA 02	0	0	1
mRNA 03	0	1	0

C) Matriz C

Figura 8. Esquema exemplificando matrizes de entrada para a ferramenta *GenMir++*. A) uma matriz Z com valores de expressão para os tecido/indivíduos analisados. B) uma matriz X com os valores de expressão de mRNA para os tecidos/indivíduos analisados. C) uma matriz C onde os números “1” (um) representam uma possível interação a ser estudada entre o miRNA e o mRNA, e os números “0” (zero) representam a não existência desta relação.

Com base nessa matriz C, o *GenMir++* analisa os valores de expressão de miRNA e mRNA para cada tecido, sendo que serão avaliados apenas as relações que forem marcadas com o valor 1 (um). Sendo assim, esta matriz (C) traz a relação dos miRNAs e seus possíveis alvos que devem ser testada pelo *GenMir++*.

O *GenMir++* tem como saída uma matriz “beta”, que traz os valores de probabilidade de um dado miRNA, marcado na matriz C com o valor de 1 (um) para um certo mRNA, ter uma interação real com este mRNA. Os valores dados em “beta” devem passar por um limiar, sendo que, os valores acima deste limiar serão considerados como indicadores de interação miRNA-mRNA, os valores abaixo deste limiar, serão desconsiderados (HUANG et al., 2007).

A partir dos valores da matriz “beta”, que foram considerados indicadores de interação miRNA-mRNA, foi utilizado o *software* gráfico *CytoScape* para a montagem de uma rede de interação. Tal *software* é

utilizado apenas para o desenho da rede propriamente dita, montando-a a partir de uma matriz de valores, no caso a matriz “beta”.

A análise e utilização da ferramenta *GenMir++* será melhor discutida no item 4.5 (Resultados e Discussões – Redes Gênicas).

*R*esultados e *D*iscussão

4. Resultados e Discussão

Neste item serão expostos os resultados obtidos em cada uma das etapas do desenvolvimento deste trabalho. Na Figura 9 é apresentado um esquema com estas etapas em detalhes.

Figura 9. Esquema representando as etapas do projeto: Aquisição das Imagens de microarray; Processamento das Imagens; Normalização; Agrupamento e Redes de Interação.

Cada uma das etapas descritas no esquema, apresentado pela Figura 9, será discutida neste item separadamente, dando maior ênfase às etapas de Normalização e de Redes de Interação, devido ao desenvolvimento da ferramenta de normalização e da utilização das Redes de Interação na busca de novos alvos.

4.1. Obtenção dos dados de expressão gênica por *microarray*

Após a hibridação, as lâminas de *microarray* foram levadas ao *scanner* para a geração das imagens de hibridação das lâminas, que podem ser observadas na Figura 10.

Esta pode ser considerada a etapa inicial para o *pipeline* para a análise *in silico* dos dados obtidos a partir dos experimentos de *microarray*, pois a partir das imagens que foram adquiridas no *scanner* é possível realizar a quantificação dos valores de expressão para o experimento em andamento.

A partir das imagens das lâminas “*scaneadas*” (Figura 10A e 10B) é possível observar alguns *spots* hibridados, onde os *spots* representados por um ponto vermelho indicam a hibridação mais efetiva com a sonda marcada com Cy5, os *spots* representados por um ponto verde indicam a hibridação mais efetiva com a sonda marcada com Cy3, e os outros tons de laranja e amarelo indicam a hibridação por ambas as sondas, em concentrações diferentes.

Aquisição das Imagens

Figura 10. Imagens exibindo a hibridação em microarrays. (A) com 4500 mRNAs de TSAs. (B) com 640 miRNAs.

4.2. Processamento das imagens

A partir de discussões entre os pesquisadores e grupos de pesquisa, e analisando os resultados dos dois *softwares* - *Spotfinder* e *BZScan* -, a utilização do *software Spotfinder* foi definida como sendo a melhor opção para as análises tanto de mRNA quanto de miRNA.

Apesar das intensidades de hibridação dos *spots* nas lâminas de miRNA serem mais baixas que a hibridação dos *spots* nas lâminas de mRNA, o *software Spotfinder* foi capaz de quantificá-las perfeitamente.

A utilização do *software BZScan* não se mostrou mais eficaz para tal quantificação, além do fato de que sendo ambas as lâminas quantificadas pelo mesmo *software*, o resultado das quantificações são apresentados no mesmo formato e padrão, tornando mais fácil a análise dos experimentos em conjunto.

A partir das lâminas de *microarray* “scaneadas” foi possível, utilizando o *software Spotfinder*, realizar a análise e quantificação dos valores de expressão, separadamente, das lâminas de miRNA e mRNA (Figura 11).

Figura 11. Parte do esquema exibindo a etapa de Processamento das imagens de *microarray*, sendo as lâminas de miRNA e mRNA processadas separadamente.

As lâminas de miRNA e mRNA foram analisadas, delimitando-se a grade dos *spots* e a remoção do “background”. Um exemplo de delimitação da grade dos *spots*, utilizando o *software Spotfinder*, pode ser observado na Figura 12.

Com base nas imagens quantificadas foi possível obter os arquivos com os valores numéricos de expressão. Estes arquivos, no formato TAV, fornecem diversas informações além dos valores de expressão.

Fonte: John C. Braisted, 2008.

Número da Linha	Número da Coluna	Linha do Bloco	Coluna do Bloco	Intensidade Cy3	Intensidade Cy5
1	1	1	1	38295	38295
1	1	1	1	44155	44155
1	2	1	1	36161	36161
1	3	1	1	36797	36797
1	4	1	1	42851	42851
1	5	1	1	267795	267795
1	6	1	1	54195	54195
1	7	1	1	44006	44006
1	8	1	1	81552	81552
1	9	1	1	75635	75635
1	10	1	1	30138	30138
1	11	1	1	85570	85570
1	12	1	1	235518	235518
1	13	1	1	123885	123885
1	14	1	1	80761	80761
1	15	1	1	30425	30425
1	16	1	1	12880	12880
1	17	1	1	38902	38902
1	18	1	1	33360	33360
1	19	1	1	30979	30979
1	20	1	1	36067	36067
1	21	1	1	26399	26399
1	22	1	1	25088	25088
1	23	1	1	0	0
1	24	1	1	48951	48951
1	25	1	1	0	0
1	26	1	1	231564	231564
1	27	1	1	22920	22920
1	28	1	1	19020	19020
1	29	1	1	57403	57403
1	30	1	1	76656	76656
1	31	1	1	70798	70798
1	32	1	1	22818	22818
1	33	1	1	208	208
1	34	1	1	152	152
1	35	1	1	116	116
1	36	1	1	85	85
1	37	1	1	108	108
1	38	1	1	250	250
1	39	1	1	235	235
1	40	1	1	132	132
1	41	1	1	76	76
1	42	1	1	124	124
1	43	1	1	108	108
1	44	1	1	306	306
1	45	1	1	281	281
1	46	1	1	264	264
1	47	1	1	295	295
1	48	1	1	245	245
1	49	1	1	101	101
1	50	1	1	78	78
1	51	1	1	161	161
1	52	1	1	0	0
1	53	1	1	107	107
1	54	1	1	70	70
1	55	1	1	94	94
1	56	1	1	133	133
1	57	1	1	131	131
1	58	1	1	157	157
1	59	1	1	212	212

O próximo passo, após a quantificação das lâminas dos experimentos de *microarray*, foi a normalização dos dados. Os arquivos no formato TAV gerados na etapa anterior serviram de entrada para a etapa seguinte, sendo importados pelo *script* de normalização.

4.3. Normalização dos Dados

Os primeiros testes realizados, com os dados de expressão quantificados e armazenados no formato TAV, se basearam na normalização dos dados pelo *script* original utilizado pelo Laboratório de Imunogenética Molecular.

Com base neste *script* observou-se que o método de normalização realizado não poderia ser utilizado para normalizar os dados de *microarray* de miRNA. Com isso, buscou-se novas alternativas para a normalização destes dados, e foi encontrada a solução com a normalização Quantil, como discutido em Materiais e Métodos.

No fragmento do esquema exibido na Figura 14 são apresentados, resumidamente, os passos realizados pelo *script* desenvolvido para a normalização. Neste *script* os dados brutos de miRNA e mRNA são importados, separadamente, para o ambiente de trabalho do *script*. A partir destes dados são gerados gráficos de densidade dos valores de expressão. Em seguida, os dados são normalizados e, novamente, são gerados gráficos de densidade, com estes novos dados. A partir destes dados normalizados são gerados, também, os arquivos necessários para as próximas etapas do projeto. Estes passos serão discutidos mais profundamente ainda neste item.

Figura 14. Esquema representando a etapa de Normalização dos dados, onde os Dados Brutos de miRNA e mRNA são importados para montagem dos arquivos de saída com os dados normalizados.

O *script* para a normalização dos dados de miRNA e mRNA foi desenvolvido na plataforma R. Esta plataforma é um *software* que emula um ambiente estatístico e pode ser executado em uma variedade de plataformas UNIX, Windows e MacOS, dessa forma, o *script* criado pode ser executado em qualquer sistema operacional em que o R esteja. O *script* pode ser visto na íntegra no APÊNDICE.

A plataforma R permite a importação de bibliotecas com funções pré-definidas para a análise de dados de microarray e, assim, facilitando no desenvolvimento do *script* de normalização. A execução do *script* e os principais pontos do desenvolvimento serão discutidos a seguir.

Após o carregamento do ambiente da plataforma R, ilustrado na Figura 15, o sistema exhibe ao usuário diversas informações básicas, como versão do R, ISBN, em qual plataforma está sendo executado, dentre outras. Ao final das

informações é exibido o cursor do R, onde o usuário deverá digitar comandos a serem executados pela plataforma.

Figura 15. Tela inicial do aplicativo R, onde o ponteiro em vermelho “aguarda” pelo código a ser executado.

Para o início de sua execução, o usuário deve apenas realizar a chamada do *script* através de um comando – ‘source(“normalizacao.R”)’ -, fazendo com que seja iniciada a execução do *script*.

Após o início da execução, como mostrado na Figura 16, é solicitado ao usuário que os arquivos no formato TAV, tanto de mRNA quanto de miRNA sejam colocados na “PASTA DE TRABALHO”, ou seja, em um diretório específico que deve ser definido para que os arquivos sejam corretamente analisados.

Inseridos os arquivos na “Pasta de Trabalho”, e informada a pasta ao *script* em execução, todo o processo de normalização e geração de gráficos e

arquivos de saídas é automatizado, sendo informado ao usuário ao final do processo, os gráficos de densidade de expressão dos dados brutos e dos dados normalizados e o local onde se encontram os arquivos de saída.

```
> source("normalizacao.R")
[1] "Coloque na PASTA DE TRABALHO os arquivos .tav a serem analisados (cDNA), assim co$
[1] "Coloque na PASTA DE TRABALHO/mirna os arquivos .tav de miRNA"
Qual o caminho para a PASTA DE TRABALHO de trabalho?
```

Figura 16. Início da execução do script de normalização, após os avisos a solicitação do caminho para a “Pasta de Trabalho”.

Com os dados importados para o ambiente R, o *script* prossegue com sua execução. Em uma primeira etapa da execução, o *script* apresenta um retorno visual dos dados com os quais irá trabalhar, para que haja uma maior facilidade no entendimento do processo de normalização.

O *script* gera, então, para os dados brutos de miRNA e para os dados brutos de mRNA, dois gráficos com as linhas de densidade de valores de expressão, representado por uma linha verde e outra vermelha, representando os valores de expressão dos marcadores Cy3 e Cy5, respectivamente. Um exemplo deste primeiro gráfico pode ser observado na Figura 17.

Seguindo com sua execução, o *script* realiza a normalização dos dados, separadamente, para cada tipo de molécula - miRNA e mRNA -, pois para a normalização dos dados de mRNA foi utilizado o método “*Print-tip Lowess*”, e para a normalização dos dados de miRNA foi utilizado o método “*Quantil*”.

Figura 17. Gráfico de densidade dos canais verde e vermelho, sem normalização.

Na Figura 18 são exibidos dois trechos do *script* responsáveis pela normalização propriamente dita, onde são chamadas as funções para normalização presentes nas bibliotecas utilizadas.

```
32  
33     ### Normalizando por Quantil  
34     Yn <- normalizeQuantile(Y)  
35  
    .  
    .  
213  
214     ### Normalizando por Print-tip lowess  
215     normalizeWithinSlide(ma, "p")  
216
```

Figura 18. Linhas de código executando as funções de normalização pré-definidas pelas bibliotecas da plataforma R.

Após a execução das funções de normalização, o *script* gera novos gráficos de densidade, para que, a partir deles, seja possível verificar se a normalização causou algum efeito visualmente significativo nos valores de intensidade.

Um exemplo deste gráfico pode ser observado na Figura 19, onde é apresentado o gráfico com os dados normalizados do mesmo experimento exibido na Figura 17.

Figura 19. Gráfico de densidade dos canais verde e vermelho, após normalização.

Como pode ser observado, comparando-se os gráficos exibidos nas Figuras 17 e 19, as linhas representando as densidades dos canais na Figura 19 estão, praticamente, sobrepostas, ao contrário do que foi apresentado na Figura 17, onde, visivelmente, os canais verde e vermelho estavam com intensidades muito diferentes. Dessa forma, é possível observar que os experimentos estão sendo normalizados dentro de um mesmo intervalo de intensidade de expressão.

Com os dados já normalizados e após a exibição do gráfico de densidades dos valores normalizados, o *script* de normalização desenvolvido auxilia na montagem dos arquivos necessários para as próximas etapas: a execução da ferramenta MEV e o *software* GenMir++.

A função para gerar os arquivos .MEV, utiliza-se de informações encontradas no arquivo TAV e nos valores de expressão normalizados para criar os arquivos a serem inseridos no *software* MEV. A função, chamada de “exportMev()”, foi baseada no *script* de normalização utilizado anteriormente pelo Laboratório de Imunogenética Molecular, onde a função gerava os arquivos .MEV de experimentos de mRNA normalizados.

Com base nos dados necessários para a execução do *GenMir++* e, conseqüentemente, geração da rede de interação entre miRNAs e mRNAs, o *script* gera os dois arquivos principais para a execução do *software* de redes: a “Matriz X” com os dados de expressão de mRNA por tecido/indivíduo analisado, e a “Matriz Z” com os dados de expressão de miRNA por tecido/indivíduo analisado.

As duas etapas seguintes – Agrupamento e Rede de Interação – podem ser realizadas na ordem que o usuário achar melhor, pois ambas são independentes, e necessitam apenas dos dados gerados pela normalização. Dessa forma optou-se pela etapa do Agrupamento antes da Rede de Interação.

4.4. Agrupamento dos dados de expressão

Para a visualização dos dados de expressão, foram gerados dois dendogramas: um baseado nos dados de expressão dos mRNAs e outro baseado nos dados de expressão de miRNAs, ambos conjuntos de dados gerados pelo *Spotfinder*, em seguida normalizados pelo *script*.

Como pode ser observado no trecho do esquema apresentado na Figura 20, os dendogramas foram gerados, separadamente, a partir do arquivo MEV gerado pelo *script*.

Os dendogramas foram gerados através do *software* MEV, utilizando-se a ferramenta *Significance Analysis of Microarray* (SAM). Os dados foram agrupados, tomando um FDR menor ou igual a 5% por meio de agrupamento hierárquico, gerando uma relação entre os indivíduos pré-auto imunes e autoimunes.

Figura 20. Esquema da etapa de Agrupamento, com os dendogramas gerados pelo *Software* MEV, a esquerda para miRNA e a direita para mRNA.

Os dendogramas são gerados levando-se em conta os valores de expressão para cada gene e cada tecido/indivíduo analisado. Através do *Software* MEV, a ferramenta SAM busca por similaridades em valores de expressão para cada conjunto gene-tecido, com base na similaridade encontrada é gerado o dendograma, agrupando tanto os genes quanto os tecidos por sua similaridade nos valores de expressão.

Tomando como base o experimento exibido na Figura 21, as linhas representam o gene – no caso o miRNA – analisado, e as colunas são as amostras – tecido ou indivíduo, por exemplo -, os retângulos marcados em cada combinação de miRNA-tecido representam os valores de expressão do miRNA para aquela amostra, sendo que o vermelho indica um valor de expressão positivo e o verde um valor de expressão negativo; as demais cores representam sobreposição entre estes valores.

Como pode ser observado nos dendogramas gerados, mostrados na Figura 21, o *software* foi capaz de agrupar tanto os dados de miRNA quanto os de mRNA, formando um grupo de indivíduos pré-auto imunes e um grupo de indivíduos auto imunes baseados nos valores de expressão encontrados para cada tipo de indivíduo. Logo, pode-se afirmar que o *software* MEV é capaz de realizar o agrupamento para ambos os tipos de dados estudados.

Figura 21. Dendograma do experimento com miRNA, exibindo o agrupamento entre os indivíduos pré-auto imunes e os indivíduos auto imunes.

4.5. Redes Gênicas

A partir do ambiente matemático MatLab, o GenMir++ foi executado e testado. Inicialmente seguindo os padrões informados pelo desenvolvedor do algoritmo, tendo como parâmetros as matrizes X, Z e C.

Na Figura 22 é possível observar o esquema da etapa da construção das Redes de Interação. Para esta etapa o *software* GenMir++ foi executado de duas maneiras diferentes. Primeiramente, assumindo que todos os mRNAs eram possíveis alvos de miRNA, e, após selecionados os valores mais expressivos, assumindo apenas alguns mRNAs como possíveis alvos de miRNAs. A seguir esta etapa é discutida mais detalhadamente.

Figura 22. Esquema representando a etapa de geração das Redes de Interação.

O *script* de normalização tem também a função de gerar os arquivos necessários para a execução do GenMir++. Sendo assim, com base nos dados normalizados de valores de expressão, o *script* gerou a matriz X e Z, com os dados de mRNA e de miRNA, respectivamente, por tecido/indivíduo estudado.

Uma terceira matriz, essencial para a execução do GenMir++, é a matriz C. Nesta matriz são marcadas com o número 1 as possíveis relações miRNA-mRNA que se deseja que o GenMir++ analise. Com base em bancos de dados de relações já definidas entre miRNA-mRNA, como por exemplo a base de dados miRBASE ([HTTP://http://www.mirbase.org/](http://www.mirbase.org/)), a matriz C foi criada, levando-se em consideração as relações já definidas pela literatura e pela base de dados miRBASE.

O GenMir++ foi então executado, e como retorno de seu algoritmo é fornecida uma nova matriz, *beta_gk*, onde o *software* estima uma probabilidade beta de que o miRNA k tem como alvo o mRNA g. A partir destes dados, é definido um limiar, para que os dados presentes acima deste limiar sejam considerados como uma verdadeira interação, e os demais sejam descartados.

Porém, a partir de considerações sobre o *software* GenMir++, percebeu-se que a montagem da matriz C estava fazendo com que os possíveis alvos de miRNA fossem encontrados diretamente através de buscas em bases de dados, realizando dessa forma a montagem de uma rede de relações que poderiam ter sido descobertas apenas buscando tais informações em bancos de dados.

Dessa forma, foram realizados alguns testes, e com base nestes testes foi definida uma nova metodologia para a utilização do GenMir++, na qual foi possível realizar uma busca de novos alvos de miRNA, e a montagem desta rede de interação.

Da mesma forma que antes, o *script* de normalização gerou as duas matrizes de dados de expressão, as matrizes X e Z. Porém, ao invés de se gerar a matriz C com apenas alguns dados marcados como possíveis relações, gerou-se uma matriz C onde foi informado que todos os mRNAs eram possíveis alvos de miRNA.

A partir deste conjunto de dados, o GenMir++, retornou uma matriz beta_gk, com os valores das possibilidades de interações. Com base nestes valores, foi definido um limiar, e a partir deste limiar foi realizado o corte dos valores considerados como possíveis alvos.

Na Figura 23 é exibido um gráfico contendo todos os valores de beta_gk para um experimento, onde todos os mRNAs foram tratados como sendo possíveis alvos de todos os miRNAs. Dessa forma, o gráfico gerado contém uma coluna para cada valor de beta_gk, separados de acordo com cada miRNA do experimento.

Como pode ser observado na Figura 23, a maioria dos valores segue um padrão, e alguns deles extrapolam este padrão. Tomando como base estes valores, foi realizado um corte a partir de um limiar, definido pela linha vermelha no gráfico da Figura 23.

Figura 23. Gráfico com os valores de beta_gk considerando em um experimento, todos os mRNAs como possíveis alvos de todos os miRNAs. Neste experimento gerado se tem 22 miRNAs e para cada miRNA se tem 90 mRNAs (colunas).

A partir deste limiar considerado, o *software* GenMir++ foi executado novamente, sendo que, desta vez, uma nova matriz C foi gerada, levando em consideração apenas os dados que apresentaram um valor de beta_gk maior que o limiar adotado.

Após a análise deste novo conjunto de dados, um novo gráfico foi gerado para Figura 24 e representa os valores de beta_gk para os possíveis alvos de miRNA.

Para uma melhor filtragem sobre estes valores, também foi aplicado o filtro de um limiar, para que apenas os valores acima deste limiar fossem considerados como possíveis interações entre miRNA e mRNA. Tal limiar é representado na Figura 24 pela linha vermelha.

Com base nessa última matriz beta_gk, foi possível a montagem das redes gênicas de interação miRNA-mRNA. Para tal foi utilizado o *software* CytoScape, que é capaz de montar grafos baseados em matrizes com as informações relevantes.

Sendo assim, a matriz beta_gk resultante das análises foi levada ao CytoScape, e dessa forma foi gerada uma rede gênica, que, baseada nos dados do projeto intitulado “Avaliação do papel dos microRNAs no controle da expressão gênica promíscua no timo de camundongos NOD (*Non Obese Diabetic*)” (2006/56741-5), pode ser observado na Figura 25.

Figura 24. Gráfico com os valores de beta_gk considerando em um experimento, apenas os mRNAs assumidos como possíveis alvos de miRNAs. Neste experimento gerado se tem 22 miRNAs e para cada miRNA se tem 90 mRNAs (colunas)

A Rede gerada, mostrada na Figura 25, mostra os miRNAs como círculos brancos e os mRNAs como quadrados pretos, e os traços ligando cada um deles como possíveis interações miRNA-mRNA.

Com estas possíveis interações expostas pelo GenMir++, o grupo de Imunogenética Molecular seguiu o estudo confirmando em bancada as relações encontradas por este *pipeline*.

O resumo intitulado “MicroRNA-mRNA Network Controlling the Promiscuous Gene Expression in the Thymus of NOD (Non Obese Diabetic) Mice: Implications in the Emergence of Type 1 Diabetes Mellitus” apresentado no 55º Congresso Brasileiro de Genética utilizou o *pipeline* do presente trabalho.

4.6. Considerações Finais

O *pipeline* desenvolvido trouxe uma solução padrão para a análise em conjunto de dados de *microarray* de mRNA e miRNA. Antes do desenvolvimento do mesmo, as análises realizadas passavam por etapas realizadas manualmente, como a normalização de dados de miRNA.

Com o uso deste *pipeline* a análise destes experimentos passou a ser quase que inteiramente automatizada, tendo como pontos de parada a troca entre programas, porém sempre tendo como saída de uma etapa os dados para a entrada da etapa seguinte.

Durante o desenvolvimento deste *pipeline* foram utilizados dados de diversos experimentos, para a utilização de mais de um conjunto de testes em cada etapa. Durante os testes em cada etapa, os programas mais utilizados e

conhecidos foram testados, além do desenvolvimento de vários *scripts* de normalização até ser encontrado o que melhor se adaptou ao *pipeline*.

Na fase de desenvolvimento do *script* de normalização, além da busca por uma metodologia mais adequada para a normalização de dados de miRNA, diversas funções foram testadas para a criação dos arquivos de saída e adequação das etapas seguintes.

A etapa de Redes de Interação foi utilizada, com sucesso, como uma ferramenta de busca de novos alvos, sendo criado com a intenção de confirmar se mRNAs são alvos de miRNAs. O GenMir++, também se mostrou eficiente na busca de novos alvos, sendo que estas possíveis interações foram confirmadas em bancada pelo Laboratório de Imunogenética Molecular.

Um ponto a ser desenvolvido no futuro seria a automatização total do *pipeline*, pois devido a utilização de *scripts* dentro de ambientes matemáticos e estatísticos – plataforma R e MatLab – não foi possível esta automatização durante o desenvolvimento deste projeto. Além da automatização, um ambiente mais amigável ao usuário poderá ser desenvolvido, facilitando a utilização do *pipeline* por usuários leigos em informática.

Conclusão

5. Conclusão

Com base no *pipeline* desenvolvido e nos testes realizados com o mesmo, pode-se concluir que os objetivos deste trabalho foram alcançados. O *pipeline* desenvolvido é capaz de realizar a análise *in silico* de dados de expressão gênica provenientes de experimentos de *microarray* de miRNA e mRNA, desde a normalização até a criação de uma rede gênica de interação de miRNA e seus alvos.

Vale ressaltar que o *script* de normalização criado é de extrema utilidade, visto que a normalização pode ser uma etapa muito dispendiosa.

Uma característica importante observada durante o desenvolvimento deste projeto foi a utilização do *software* GenMir++ como uma ferramenta para busca de possíveis novos alvos com base em dados de *microarray*. Utilizando-se dessa tecnologia, o tempo gasto em bancada para a validação de possíveis alvos é reduzido drasticamente, sendo que o sistema traz ao usuário um número de possíveis alvos com uma probabilidade maior de serem confirmados.

Por tanto, conclui-se que os objetivos do projeto foram alcançados, sendo que o *pipeline* resultante já foi utilizado por pesquisadores do Laboratório de Imunogenética Molecular e os resultados apresentados no 55º Congresso Brasileiro de Genética no resumo intitulado “MicroRNA-mRNA Network Controlling the Promiscuous Gene Expression in the Thymus of NOD (Non Obese Diabetic) Mice: Implications in the Emergence of Type 1 Diabetes Mellitus”.

Referências Bibliográficas

REFERÊNCIAS BIBLIOGRÁFICAS

AGBAVWE, C.; SOMOZA, M. M. Sequence-dependent fluorescence of cyanine dyes on microarrays. **PLoS One**, v. 6, n. 7, 2011.

AMBROS, V. The functions of animal microRNAs. **Nature**, v. 431, n. 7006, p. 350-355, 2004.

BARTEL, DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. **Cell**, v. 116, 281-297, 2004.

BENGTSSON, H.; JÖNSSON, G.; VALLON-CHRISTERSSON, J. Calibration and assessment of channel-specific biases in microarray data with extended dynamical range. **BMC Bioinformatics**, v. 5, p. 177, 2004.

BILLOUD, B.; DE PAEPE, R.; BAULCOMBE, D.; BOCCARA, M. Identification of new small non-coding RNAs from tobacco and Arabidopsis. **Biochimie**, v. 87, p. 905-910, 2005.

BOLSTAD, B. M.; IRIZARRY, R. A.; ASTRAND, M.; SPEED, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. **Bioinformatics**, v. 19, p. 185-193, 2003.

CHIROMATZO, A. O.; OLIVEIRA, T. Y. K.; PEREIRA G.; COSTA A. Y.; MONTESCO C. A. E.; GRAS D. E.; YOSETAKE, F.; VILAR, J. B.; CERVATO, M.; PRADO, P. R. R.; CARDENAS, R. G. C. C. L.; CERRI, R.; BORGES, R. L.; LEMOS, R. N.; ALVARENGA, S. M.; PERALLIS, V. R. C.; PINHEIRO, D. G.; SILVA, I. T.; BRANDÃO, R. M.; CUNHA, M. A. V.; GIULIATTI, S.; SILVA JR, W. A. MiRNApath: a database of miRNAs, target genes and metabolic pathways. **Genetics and Molecular Research**, v. 6, n. 4, p. 859-865, 2007.

CHURCHILL, G. A. Discussion to statistical challenges in functional genomics-comment. **Statistical science**, v. 18, p. 64-69, 2003.

CROUSER, E. D.; JULIAN, M. W.; CRAWFORD, M.; SHAO, G.; YU, L.; PLANCK, S. R.; ROSENBAUM, J. T.; PATRICK NANA-SINKAM, S. Differential expression of microRNA and predicted targets in pulmonary sarcoidosis. **Biochemical and Biophysical Research Communications**, 2011.

DAVISON, T.; JOHNSON, C.; ANDRUSS, B. Analyzing micro-RNA expression using microarray. **Methods in Enzymology**, v. 411, p. 14-34, 2006.

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

ENRIGHT, A. J.; BINO, J.; GAUL, U.; TUSCHL, T.; SANDER, C.; MARKS, D. S. MicroRNA targets in Drosophila. **Genome Biology**, v. 5, n. 1, p. R1 1-14, 2003.

ESTEVEZ, G.H. Métodos estatísticos para a análise de dados de cDNA microarray em um ambiente computacional integrado. São Paulo; 2007. Tese de Doutorado – Universidade de São Paulo.

FERREIRA FILHO, D.; LEANDRO, R. A. Tutorial de Microarray. **54º RBRAS e 13º SEAGRO**, 2009.

GENTLEMAN, R. C.; CAREY, V. J.; BATES, D. M.; BOLSTAD, B.; DETTLING, M.; DUDOIT, S.; ELLIS, B.; GAUTIER, L.; GE, Y.; GENTRY, J.; HORNIK, K.; HOTHORN, T.; HUBER, W.; IACUS, S.; IRIZARRY, R.; LEISCH, F.; LI, C.; MAECHLER, M.; ROSSINI, A. J.; SAWITZKI, G.; SMITH, C.; SMYTH, G.; TIERNEY, L.; YANG, J. Y.; ZHANG, J. Bioconductor: open software development for computational biology and bioinformatics. **Genome Biology**, v. 5, n. 10, p. 80, 2004.

HARRINGTON, C. A.; ROSENOW, C.; RETIEF, J. Monitoring gene expression using DNA microarrays. **Current Opinion on Microbiology**, v. 3, n. 3, p. 285-291, 2000

HE, L.; HANNON, G. J. MicroRNAs: small RNAs with a big role in gene regulation. **Nature Reviews Genetics**, v. 5, p. 522-531, 2004.

HE, L.; THOMSON, J. M.; HEMANN, M. T.; HERNANDO-MONGE, E.; MU, D.; GOODSON, S.; POWERS, S.; CORDON-CARDO, C.; LOWE, S. W.; HANNON, G. J.; HAMMOND, S. M. A microRNA polycistron as a potential human oncogene. **Nature**, v. 435, p. 828-833, 2005.

HUANG, J. C.; BABAK, T.; CORSON, T. W.; CHUA, G.; KHAN, S.; GALLIE, B. L.; HUGHES, T. R.; BLENCOWE, B. J.; FREY, B. J.; MORRIS, Q. D. Using expression profiling data to identify human microRNA targets. **Nature Methods**, v. 4, p. 1045-1049, 2007.

KRICHEVSKY, A. M.; KING, K. S.; DONAHUE, C. P.; KHRAPKO, K.; KOSIK, K. S. A microRNA array reveals extensive regulation of microRNAs during brain development. **RNA**, v. 9, p. 1274-1281, 2003.

LEE, R. C.; FEINBAUM, R. L.; AMBROS, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. **Cell**, v. 75, n. 5, p. 843-854, 1993.

LIM, L. P.; GLASNER, M. E.; YEKTA, S.; BURGE, C. B.; BARTEL, D. P. Vertebrate MicroRNA Genes. **Science**, v. 299, n. 5612, p. 1540, 2003.

LIM, L. P.; LAU, N.C.; GARRETT-ENGELE, P.; GRIMSON, A.; SCHELTER, J. M.; CASTLE, J.; BARTEL, D. P.; LINSLEY, P. S.; JOHNSON, J. M. Microarray analysis shows that some microRNAs downregulate large amounts of target mRNAs. **Nature**, v. 433, n. 7027, p. 769-773, 2005.

LIN, J. HORIKAWA, Y.; TAMBOLI, P.; CLAGUE, J.; WOOD, C. G.; WU, X. Genetic variations in microRNA-related genes are associates with survival and recurrence in patients with renal cell carcinoma. **Carcinogenesis**, v. 31, n. 10, p. 1805, 1812, 2010.

LOPEZ, F.; ROUGEMONT, J.; LORIOD, B.; BOURGEOIS, A.; LOI, L.; BERTUCCI, F.; HINGAMP, P.; HOULGATTE, R.; GRANJEAUD, S.; Feature extraction and signal processing for nylon DNA microarrays. **BMC Genomics**, v. 5, n. 38, p. X, 2004.

RAJEWSKI, N. MicroRNAs target predictions in mammals. **Nature Genetics**, v. 38, p. S8-S13, 2006.

RAO, Y.; LEE, Y.; JARJOURA, D.; RUPPERT, A. S.; LIU, C.; HSU, J. C.; HAGAN, J. P. A comparison of normalization techniques form microRNA microarray data. **Statistical Applications in Genetics and Molecular Biology**, v. 7, n. 1, 2008.

ROMERO, D. G.; PLONCZYNSKI, M. W.; CARVAJAL, C. A.; GOMEZ-SANCHEZ, E. P.; GOMEZ SANCHEZ, C. E. Microribonucleic acid-21 increases aldosterone secretion and proliferation in H295R human adrenocortical cells. **Endocrinology**, v. 149, n. 5, p. 2477-2483, 2008.

SAEED, A. I.; SHAROV, V.; WHITE, J.; LI, J.; LIANG, W.; BHAGABATI, N.; BRAISTED, J.; KLAPA, M.; CURRIER, T.; THIAGARAJAN, M.; STURN, A.; SNUFFIN, M.; REZANTSEV, A.; POPOV, D.; RYLTSOV, A.; KOSTUKOVICH, E.; BORISOVSKY, I.; LIU, Z.; VINSAVICH, A.; TRUSH, V.; QUACKENBUSH, J. TM4: a free, open-source system for microarray data management and analysis. **Biotechniques**, v. 34, n. 2, p. 374-378, 2003.

SEVIGNANI, C.; CALIN, G. A.; SIRACUSA, L. D.; CROCE, C. M. Mammalian microRNAs: a small world for fine-tuning gene Expression. **Springer Science Business Media**, v. 17, p. 189-202, 2006.

SMITH, G. K.; SPEED, T. P. Normalization of cDNA microarray data. **Methods**, v. 31, p. 265-273, 2003.

TORRES, A.; TORRES, K.; PASZKOWSKI, T.; JODLOWSKA-JEDRYCH, B.; RADOMANSKI, T.; KSIAZEK, A.; MACIEJEWSKI, R. Major regulators of microRNA biogenesis Dicer and Drosha are down-regulated in endometrial cancer. **Tumor Biology**, v. 32, n. 4, p. 769-776, 2011.

WINTER, J.; JUNG, S.; KELLER, S.; GREGORY, R. I.; DIEDERICH, S. Many roads to maturity: microRNA biogenesis pathways and their regulation. **Nature Cell Biology**, v. 11, n. 3, p. 228-234, 2009.

WU, C. C.; HUANG, H. C.; JUAN, H. F.; CHEN, S. T. GeneNetwork: an interactive tool for reconstruction of genetic networks using microarray data. **Bioinformatics**, v. 20, n. 18, 2004.

Apêndice

APÊNDICE

```
library(kth)
library(aroma.light)

#####
##
## Normalização de miRNA e geração do arquivo Z para GenMir, com ##
## dados normalizados ##
## ##
#####

exportMirna<-function()
{
  X=NULL
  Yn=NULL
  Y = matrix(nrow=768,ncol=im)

  ### Alterando o diretório de trabalho para a pasta contendo os
  dados de miRNA
  zm = paste(z,"/mirna",sep="")
  setwd(zm)
  print(paste("O diretorio de trabalho agora e",getwd()))

  ### Lendo lista de arquivos .tav
  mrn<-list.files(pattern=".tav")
  im<-length(mrn)

  ### Lenda dados de expressão dos arquivos TAV
  for (x in 1:im){
    X <- read.csv(file=mrn[x],head=FALSE,sep="\t")
    Y[,x] = X[,14]
  }

  summary(Y)
  is.matrix(Y)

  ### Normalizando por Quantil
  Yn <- normalizeQuantile(Y)

  ### Gerando Gráficos pré e pós normalizados
  layout(matrix(1:1, ncol=12))
  xlim <- range(Y, Yn, na.rm=TRUE);
  plotDensity(Y, lwd=2, xlim=xlim, main="Distribuições de valores
originais")
  X11()
  plotDensity(Yn, lwd=2, xlim=xlim, main="Distribuições de valores
normalizados")

  exportMEV(Yn.obj=ma, X.obj=gpr, 1:a, idxVec=NULL, slidenames =
NULL)

  ### Gerando arquivo Z
  write.table(Yn,"Z.txt",append=FALSE,
quote=FALSE,col.names=FALSE, row.names=FALSE, sep="\t")
}
```

```
## Fim da Função exportMirna ##

#####
##                                     ##
## Função para a geração do arquivo X de entrada do GenMir com      ##
## dados NORMALIZADOS                                               ##
##                                     ##
#####

exportX<-function()
{
  print("Gerando arquivo X.txt")

  gnm<-list.files(pattern=".tav.mod.normalised.txt")
  i<-length(gnm)

  gX = NULL
  names = NULL

  names = matrix(nrow=1,ncol=i)
  gX = matrix(nrow=4608,ncol=i)

  for(j in 1:i){
    names[1,j] = as.matrix(gnm[j])
    file = as.matrix(read.table(gnm[j], sep="\t"))
    for(k in 2:4608){
      gX[k-1,j] = file[k,7]
    }
  }
  write.table(gX,"X.txt",append=TRUE,col.names=FALSE,
row.names=FALSE, quote=FALSE, sep="\t")
}
## Fim da Função exportX ##

#####
##                                     ##
## Função que cria os arquivos .MEV com dados NORMALIZADOS de      ##
## entrada no MEV                                                  ##
##                                     ##
#####

exportMEV<-function(ma.obj, gpr.obj = gpr, slideVec, idxVec =
NULL,slidetype="Stem32", slidenames = NULL,...)
{
  nrow <- length(ma.obj$M[, 1])
  idx <- 1:nrow
  nr <- length(idx)
  UID <- 0
  for (k in 1:nr) {
    UID[k] <- paste(gpr.obj$layout$id[k], sep = "")
  }
  rg <- as.RGData(ma.obj, slide = slideVec)
  IB <- rg$R[,]
  IA <- rg$G[,]
  if (length(slideVec) == 1) {
    IB <- as.character(IB)
    IA <- as.character(IA)
  }
  if (length(slideVec) > 1) {
```

```

        for (k in 1:length(slideVec)) {
            IB[, k] <- as.character(IB[, k])
            IA[, k] <- as.character(IA[, k])
        }
    }
    nbrPerBlock <- ma.obj$layout$nspot.r * ma.obj$layout$nspot.c
    nbrPerRowOfBlocks <- nbrPerBlock * ma.obj$layout$ngrid.c
    MR <- gpr.obj$grid.r
    MC<-gpr.obj$grid.c
    Row<-0
    Column <- 0
    for (z in 1:(gpr$layout$ngrid.r*gpr$layout$nspot.r)){
        Row[(z*gpr.obj$layout$nspot.c)-(gpr.obj$layout$nspot.c-
1)):(z*gpr.obj$layout$nspot.c)]<-z
    }

    SC<-gpr.obj$spot.c
    SR<-gpr.obj$spot.r
    Column<-gpr.obj$spot.c
    pos.data <- cbind(Row, Column, MR, MC, gpr.obj$Block,
        SR, SC)
    for (m in 1:length(pos.data[1, ])) {
        pos.data[, m] <- as.character(pos.data[, m])
    }
    gc()
    if (is.null(idxVec) == FALSE) {
        if (length(slideVec) == 1) {
            IA <- IA[idxVec]
            IB <- IB[idxVec]
        }
        else {
            IA <- IA[idxVec, ]
            IB <- IB[idxVec, ]
        }
        pos.data <- pos.data[idxVec, ]
    }
    print("Os seguintes arquivos foram criados em seu diretorio de
trabalho:")
    slidenams<-getSlideNames(gpr)
    for (numSlideProc in 1:length(slideVec))
    {
        out.obj <- data.frame(I(UID), IA[, numSlideProc],IB[,
numSlideProc], pos.data)
        colnames(out.obj) <- c("UID", "IA", "IB", "R", "C", "MR",
            "MC", "Block", "SR", "SC")
        filename <- print(paste(slidenams[numSlideProc],
            "mev", sep = "."))
        write.table(out.obj,filename,sep="\t", dec=".",row.names =
FALSE, na = "NaN",quote=FALSE)
    }
    Names<- gpr$Name[,2]
    anno<- data.frame(cbind(UID,Row,Column,MR,MC),Names)
    write.table(anno,"anotationmouse.ann",sep="\t", dec=".",row.names
= FALSE, na = "NaN",quote=FALSE)
    print("Função experimental, use por sua conta e risco.")
}
## Fim da Função exportMEV ##

```



```
#####
##
## Especifica o diretório onde os arquivos se encontram
##
#####

print("Coloque na PASTA DE TRABALHO os arquivos .tav a serem
analizados (cDNA), assim como o arquivo names.csv")

print("Coloque na PASTA DE TRABALHO/mirna os arquivos .tav de miRNA")

z<-readline("Qual e o diretorio de trabalho?")
setwd(z)

print(paste("O diretorio de trabalho agora e",getwd()))

#####
##
## Busca dos arquivos .TAV
##
#####

x<-list.files(pattern=".tav")
nomes<-as.vector(read.csv("names.csv"))
a<-length(x)
tavb<-matrix(ncol=17, nrow=4608)
for (b in 1:a){
  vari<-read.table(x[b],sep="\t", dec=".", header=FALSE)
  tavb<-vari[,1:17]
  write.table(tavb, paste(x[b],"mod",sep="."), sep="\t", dec=".",
col.names=FALSE, row.names=FALSE)
}
x<-list.files(pattern="mod")

#####
##
## Criação das variáveis para a realização da normalização e criação##
## dos arquivos de saída
##
#####

adf<-readADF("mouse4608ADF.adf")
gpr<-SpotfinderData$read(x)
gpr$Name<-nomes
gpr$Block<-adf[,2]
gpr$grid.r<-adf[,2]
gpr$grid.c<-adf[,1]
gpr$spot.c<-adf[,3]
gpr$spot.r<-adf[,4]

setName(gpr$layout, nomes[,2])
setId(gpr$layout, nomes[,1])
s<-indexOf(gpr$layout, id="Scorecard")
gpr$integralB[s,]<-NA

raw <- as.RawData(gpr)
ma <- getSignal(raw, bgSubtract=FALSE)
normalizeWithinSlide(ma,"p") ##print-tip group lowess normalization
normalizeAcrossSlides(ma)
```

```
#####
##
## Chamada das funções de geração dos arquivos de saída      ##
##
#####

print("<----- GERANDO ARQUIVOS MEV ----->")

exportMEV(ma.obj=ma, gpr.obj=gpr, l:a, idxVec=NULL, slidenames = NULL)

print("<----- GERANDO ARQUIVO ENTRADA GENMIR - 'X': EXPRESSÃO cDNA ---
-->")
exportX()

print("<---- NORMALIZANDO DADOS DE miRNA ----->")

print("<---- GERANDO ARQUIVO ENTRADA GENMIR - 'Z': EXPRESSÃO miRNA ---
-->")

exportMirna()

print("<---- Verifique nas pastas de trabalho os arquivos gerados. ---
-->")

#####
```