ANÁLISE E CLASSIFICAÇÃO DA EXPRESSÃO GÊNICA DURANTE O ERITEMA NODOSO HANSÊNICO ATRAVÉS DE DADOS DE MICROARRANJO

Karla Lopes de Almeida

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
Euzenir Nunes Sarno

Rio de Janeiro
Março de 2017

ANÁLISE E CLASSIFICAÇÃO DA EXPRESSÃO GÊNICA DURANTE O ERITEMA NODOSO HANSÊNICO ATRAVÉS DE DADOS DE MICROARRANJO

Karla Lopes de Almeida

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.

Examinada por:

Prof. Flávio Fonseca Nobre, Ph.D.

Prof. Euzenir Nunes Sarno, Livre Docência

Prof. Rosimary Terezinha de Almeida, Ph.D.

Prof. Milton Ozório Moraes, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2017

Almeida, Karla Lopes de

Análise e classificação da expressão gênica durante o Eritema Nodoso Hansênico através de dados de microarranjo/Karla Lopes de Almeida. – Rio de Janeiro: UFRJ/COPPE, 2017.

XII, 90 p.: il.; 29, 7cm.

Orientadores: Flávio Fonseca Nobre

Euzenir Nunes Sarno

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Biomédica, 2017.

Referências Bibliográficas: p. 69 – 79.

1. Microarranjo. 2. Classificação gênica. 3. Eritema Nodoso Hansênico. 4. Hanseníase. I. Nobre, Flávio Fonseca *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Título.

Aos meus pais Isabel e Francisco,
À minha irmã Cristiane e minha
sobrinha Manuela

Agradecimentos

À Deus pelo suporte espiritual.

Ao meu orientador Professor Dr. Flávio Fonseca Nobre pela confiança, troca de conhecimentos, amizade e carinho. Muito Obrigada!

À minha orientadora Professora Euzenir Nunes Sarno pelas saudáveis discussões sobre este trabalho. Muito Obrigada!

À Professora Dra. Rosimary Terezinha de Almeida, pelas discussões, ideias e pelo tempo disponibilizado. Muito Obrigada!

Ao meu namorado Jassiel Vladimir por seu apoio, companheirismo e me manter feliz em todos os momentos. Te amo!

À minha família por todo apoio e incentivo. Amo vocês!

Aos amigos do Laboratório de Engenharia de Sistemas da Saúde pelas ideias, boa convivência, incentivo e companheirismo. Muito Obrigada!

À coordenação do Programa de Pós-Graduação em Engenharia Biomédica. Muito Obrigada!

Agradeço ao CNPq pelo apoio financeiro a mim e ao meu programa, sem o qual este trabalho não poderia ter sido realizado.

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

ANÁLISE E CLASSIFICAÇÃO DA EXPRESSÃO GÊNICA DURANTE O ERITEMA NODOSO HANSÊNICO ATRAVÉS DE DADOS DE MICROARRANJO

Karla Lopes de Almeida

Março/2017

Orientadores: Flávio Fonseca Nobre
Euzenir Nunes Sarno

Programa: Engenharia Biomédica

A hanseníase é uma doença infecciosa crônica que afeta milhões de pessoas em todo o mundo. Atualmente, a identificação de episódios reacionais da hanseníase é de importância na pesquisa, por se tratar de uma doença negligenciada. Esses episódios são os que causam as maiores sequelas nos portadores da hanseníase. Entretanto, ainda não se obteve um método de diagnóstico que prediz o episódio, previamente ao aparecimento dos sinais e sintomas. O objetivo deste trabalho é propor uma metodologia para analisar e classificar a expressão gênica presente durante o episódio reacional Eritema Nodoso Hansênico (ENH) através de dados de microarranjo. Para a aplicação dessa metodologia foram utilizados dados dos níveis de expressão de experimentos de microarranjo de induvíduos com ENH e Lepra Lepromatosa (LL). Com a utilização de três técnicas estatísticas (teste t de *Student* com *fold-change*, Método Diferencial - Análise de Componentes Principais (MD-ACP) e modelo linear para análise de dados de microarranjo-limma), foi possível identificar 53 genes diferencialmente expressos. Estes genes foram considerados os mais significantes para a classificação entre os grupos ENH e LL; e utilizados para criar um modelo matemáticos preditivo, utilizando a modelagem regressão logística. Nossa metodologia identificou genes que foram previamente descritos na literatura. O modelo de regressão logística proposto usando esses genes mostrou-se eficiente quando utilizado em um novo conjunto de dados. O modelo apresentou uma acurácia de 92% e uma especificidade de 83,3%.

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

GENE EXPRESSION ANALYSIS AND CLASSIFICATION DURING ERYTHEMA NODOSUM LEPROSUM BY MICROARRAY DATA

Karla Lopes de Almeida

March/2017

Advisors: Flávio Fonseca Nobre
Euzenir Nunes Sarno

Department: Biomedical Engineering

Leprosy is a chronic infectious disease that affects millions of people around the world. Currently, the identification of leprosy reactions is of importance in researching because it is a neglected disease. These reactions are the ones that cause the greatest sequels in people with leprosy. However, a diagnostic method that predicts the reaction has not yet been obtained, prior to the appearance of the signs and symptoms. The objective of this work is to propose a methodology to analyze and classify the gene expression present during the reaction stage Erythema Nodosum Leprosum (ENL) through microarray data. For the application of this methodology, data of the expression levels of microarray experiments with ENL and Lepromatous Leprosy (LL) patients were used. We used three statistical techniques (Student's t test with *fold-change*, Differential Method - Principal Components Analisys (DM-PCA) and linear model microarray analisys-limma), to identify 53 differentially expressed genes. These genes were considered the most significant for the classification between the ENH and LL groups and they were used to create a predictive mathematical model using logistic regression modeling. Our methodology identified genes that were previously described in the literature. The logistic regression model proposed using these genes proved to be efficient when used in a new dataset. The model presented an accuracy of 92% and a specificity of 83.3%.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A hanseníase é uma doença infecciosa crônica que afeta milhões de pessoas em todo o mundo. Estima-se que centenas de milhares de novos casos são diagnosticados a cada ano, principalmente na Índia e no Brasil [1].

A hanseníase é causada pelo Mycobacterium leprae (M. Leprae), ou pelo recém-descoberto Mycobacterium lepromatosis [2]. Estes organismos são conhecidos por terem uma elevada afinidade pelas células de Schwann (neurotropismo), mas podem afetar a maioria dos órgãos humanos, com exceção do sistema nervoso central.

A evolução da hanseníase depende da resposta imunológica do hospedeiro, pois as manifestações clínicas podem variar a partir de lesões menores de pele e danos nos nervos até deformidades e comprometimentos sistêmicos.

A compreensão da hanseníase torna-se difícil sem o auxílio de um sistema de classificação, devido ao seu amplo espectro de manifestações clínicas e imunológicas. Para realizar a classificação clínica da hanseníase, dois sistemas têm sido mais utilizados; o sistema de classificação de Ridley e Jopling [3] e a de classificação adotada pela Organização Mundial da Saúde (OMS) [4].

A partir da classificação adotada pela OMS, foi possível implementar o tratamento com poliquimioterapia, sendo baseada no número de lesões cutâneas. Sendo assim, com o sucesso da poliquimioterapia (PQT) proposta pela OMS em 1982, a atenção passou a se concentrar nas reações da hanseníase, que atualmente é um dos problemas mais significativos no cuidado aos indivíduos. Os episódios reacionais ou reações hansênicas são processos inflamatórios agudos que ocorrem em 30% a 50% dos indivíduos no decurso da doença [5].

Essas reações são consideradas de alta gravidade, pois além de afetar a pele e os nervos, podem resultar em deficiência física dos indivíduos. Existem dois tipos de reação hansênica: a Reação Hansênica Tipo 1 (reação reversa, RR) e a Reação Hansênica Tipo 2 (eritema nodoso hansênico, ENH). A RR consiste no desenvolvi-

mento de novas lesões dermatologias (manchas ou placas), edema nas lesões antigas (com ou sem espessamento) e dor nos nervos periféricos (neurite). O EHN consiste no aparecimento de nódulos subcutâneos dolorosos, devido à formação de complexos imune, acompanhados ou não de febre, dores articulares e mal-estar generalizado[6].

O ENH é mais frequentemente encontrado na forma multibacilar da hanseníase e muitas vezes é associada a infecções bacterianas e parasitárias [7]. A recorrência do ENH pode persistir por anos. Estas podem ser caracterizadas pelo surgimento abrupto de nódulos que podem variar de poucos a inúmeros. Esses nódulos podem evoluir para necrose, nas formas mais graves do ENH.

As reações hansênicas constituem intercorrências na doença, com sinais e sintomas que levam o indivíduo ao sofrimento e a sequelas neurológicas; muitas vezes mais expressivas que as esperadas na hanseníase sem quadro reacional. Até o presente momento, ao nosso conhecimento, não existe nenhum teste para o diagnóstico precoce dessas reações hansênicas. O diagnóstico é comumente realizado principalmente após o aparecimento dos sinais e sintomas clínicos. A falta de um método eficiente para o diagnóstico precoce da hanseníase tem levado ao estudo de alterações gênicas que possam interligar à doença.

Estudos por meio do sequenciamento genômico de cepas de M. leprae no mundo, têm demonstrado baixa diversidade genética, o que sugere que as diferenças na susceptibilidade à doença e às manifestações clínicas (incluindo reações hansênicas), são influenciados pelos genes de defesa do hospedeiro [8]. Só recentemente, estudos de epidemiologia genética identificaram fatores de susceptibilidade para reações hansênicas [9]. Tais estudos requerem um entendimento das informações gênicas dos portadores da hanseníase.

As informações genéticas de um ser vivo são guardadas em uma biomolécula conhecida como ácido desoxirribonucleico ou DNA . Essa molécula abriga algumas das principais informações necessárias para a manutenção da vida deste organismo. Essas informações são codicadas em nível molecular onde pequenas porções do DNA, conhecidas como genes, são transcritas em moléculas de ácido ribonucleico (RNA), em um processo conhecido como expressão gênica. Os RNAs são, então, traduzidos em proteínas, que são as biomoléculas funcionais das células e que trabalham na manutenção de um ambiente favorável à sustentação do ciclo de vida celular.

Um dos principais desafios da biologia molecular atual é medir e avaliar os perfis de expressão gênica em diferentes tipos de tecidos biológicos. Uma das suas finalidades é o entendimento dos mecanismos de transformação molecular entre diferentes condições biológicas, como por exemplo: tecidos sadios e patológicos. Com raras exceções, todas as células de um organismo contêm a mesma carga genética, ou seja, o mesmo DNA genômico. Os níveis de expressão dos genes de grupos celulares morfologicamente diferentes constituem parcelas importantes na diferenciação entre

tipos celulares. Assim, a comparação dos perfis de expressão gênicas entre diferentes tecidos, pode levar ao entendimento de diversos fenômenos biológicos encontrados nos organismos.

A abordagem voltada para os sistemas biológicos não é nova na literatura científica e já tem sido discutida desde a década de 40. A principal discussão da literatura foi quando os modelos teóricos esbarravam na ausência de técnicas experimentais, para medir os níveis de atividade das moléculas envolvidas com o sistema. O que se tornou possível, nos dias atuais, com as técnicas de análise de expressão gênica em larga escala. Tais técnicas medem os níveis de RNAs produzidos pelos genes em condições específicas, logo, é necessário considerar que esses dados fornecem um esboço do que acontece realmente em nível proteico, para os sistemas biológicos.

Entre essas técnicas de medição, ressaltam-se os experimentos por microarranjo. Essa técnica tem sido amplamente utilizada em pesquisas de biologia molecular para medir os níveis de expressão de milhares de genes em condições biológicas específicas. Os dados obtidos pela técnica são importantes para justificar ou levantar novas hipóteses biológicas acerca dos mecanismos de transformação celular entre os diversos tipos de tecidos estudados. O processo de obtenção dos dados de expressão gênica por meio da técnica de microarranjo é muito complexo e repleto de etapas experimentais de natureza distinta

Vários métodos para análise de microarranjo já foram propostos na literatura, dentre eles é possível citar alguns que têm sido tradicionalmente utilizados, como: a busca de genes diferencialmente expressos (DE), onde são realizados testes de hipóteses para identificar diferenças significativas nos níveis de expressão entre diferentes tipos biológicos estudados; e a identificação de grupos de genes com perfis de expressões similares. A seleção de variáveis tem como objetivo a eliminação de variáveis (genes, no nosso caso) que contenham pouca ou nenhuma informação preditiva relevante. Esta seleção de genes pode melhorar a compreensão dos modelos classificadores resultantes e viabilizar a construção de um modelo que generalize de forma mais eficiente a classificação das amostras dentro do conjunto de dados.

Ao nosso conhecimento, não existe um procedimento de análise genérico que possa ser empregado indiscriminadamente para qualquer grupo de medidas de expressão gênica obtidas através de experimentos de microarranjo. O ideal é que cada conjunto de dados a ser analisado seja cuidadosamente avaliado e os métodos de análise mais adequados sejam escolhidos para sua correta utilização. Essa avaliação deve ser feita através de uma análise exploratória visando duas finalidades específicas. A primeira delas é a avaliação da qualidade dos dados e a busca por eventuais fatores experimentais sistemáticos que devem ser removidos pelos métodos de normalização. A segunda é a busca de características dos dados que justifiquem suposições necessárias para a utilização adequada de métodos específicos de análise.

Atualmente, a identificação de episódios reacionais da hanseníase é importante visto as características da doença. Esses episódios são os que causam as maiores sequelas nos portadores da hanseníase. Entretanto, ainda não se obteve um método de diagnóstico que prediz o episódio, previamente ao aparecimento dos sinais e sintomas. Neste trabalho se propõe uma metodologia para analisar e classificar a expressão gênica presente durante o episódio reacional Eritema Nodoso Hansênico, utilizando dados de microarranjo. São utilizadas três técnicas para a identificação de genes diferencialmente expressos; o teste t de Student com *fold-change*, o modelo derivado a partir das componentes principais (MD-ACP) e o modelo linear para análise de microarranjo (limma). A partir dos genes obtidos em comum com as três técnicas, é desenvolvido um modelo de predição, por meio da regressão logística, para analisar e classificar a expressão gênica presente durante o episódio reacional Eritema Nodoso Hansênico usando-se dados de microarranjo.

A presente dissertação está estruturada sete capítulos, o Capítulo 2 descreve a revisão da literatura sobre os métodos de análise e classificação dos dados obtidos por microarranjo. Posteriormente, o Capítulo 3 apresenta a fundamentação teórica. No Capítulo 4 são descritos os dados e métodos utilizados na metodologia proposta. Os resultados obtidos são apresentados e discutidos nos Capítulos 5 e 6, respectivamente. Finalmente, o Capítulo 7 contém as principais conclusões deste trabalho.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo deste trabalho é propor uma metodologia para analisar e classificar a expressão gênica presente durante o episódio reacional Eritema Nodoso Hansênico utilizando dados de microarranjo.

1.1.2 Objetivos Específicos

- Identificar os genes mais diferencialmente expressos entre a classificação Lepra Lepromatosa e o episódio reacional Eritema Nodoso Hansênico, utilizando três técnicas diferentes (teste t de Student com *fold-change*, modelo derivado a partir das componentes principaise modelo linear para análise de microarranjo).
- Identificar os genes diferencialmente expressos encontrados em comum nas três técnicas utilizadas e criar um subconjunto de dados utilizando esses genes.

- Obter um modelo matemático para classificação de tipo de reação usando a regressão logística (RL) com base no subconjunto de genes selecionados.

Capítulo 2

Revisão de Literatura

2.1 Análise de expressão gênica

Existem diversas técnicas que são aplicadas no campo da biologia molecular para a compreensão das respostas gênicas de muitos organismos submetidos a várias situações ambientais [10]. Um dos procedimentos mais conhecidos para obter a expressão de milhares de genes consiste na tecnologia de microarranjo.

Um dos primeiros trabalhos a utilizar microarranjos é o estudo de Schena *et al* [11]. Após fazer experimentos com Arabidopsis thaliana, os autores sugerem que em decorrência do grande crescimento no sequenciamento do genoma humano, o microarranjo poderia ser utilizado para a medição dos níveis de padrões da expressão, podendo monitorar uma ampla variedade de condições patológicas e fisiológicas através das mudanças dos padrões da expressão gênica; fornecendo com isto uma ligação entre a medicina clínica e a expressão gênica.

Posteriormente, Golub *et al* [12] utilizaram microarranjo de DNA para monitorar a expressão de genes em um método sistemático de classificação do câncer. Eles propuseram uma classificação molecular do câncer, dividindo-a em duas etapas: descoberta de classe e predição de classe. As etapas referem-se à definição de subtipos de tumor previamente desconhecidos e à atribuição de classes já definidas nas amostras de tumores em futuras amostras.

A capacidade do microarranjo em medir a expressão de milhares de genes simultaneamente é de grande importância e, com novas tecnologias sendo disponibilizadas para esse objetivo, novas estratégias para o estudo desses genes vêm se aprimorando [13]. Cabe salientar que algumas dessas tecnologias podem obter resultados diferentes e/ou complementares aos dados de microarranjo [14].

Nenhuma das tecnologias existentes atende a todos os objetivos que se espera de um experimento, contudo, há vantagens e desvantagens para cada uma delas [15].

Dentre essas tecnologias podemos citar: ESTs (*Expression Sequence Target*) [16], SAGE (*Serial Analysis for Gene Expression*) [17], RT-qPCR (*Quantitative reverse transcription PCR*) [18] e RNA-seq (*Next Generation Sequencing RNA*) [19, 20].

2.2 Identificação de genes diferencialmente expressos

Um dos primeiros métodos utilizados para identificar genes diferencialmente expressos foi o fold-change. Esse método é bastante popular devido a sua simplicidade para um método de seleção de genes diferencialmente expressos [21]. Contudo, esse método não leva em conta a variância dos valores de expressão medidos. Então, é recomendado que se use em combinação com outro método estatístico, como por exemplo o teste t de Student [22]. O teste t envolve uma formulação de uma hipótese nula e uma pesquisa para cada gene [23].

No trabalho de Rollins e Teh [24] foi proposto um novo método baseado em Análise de Componentes Principais (ACP) para analisar dados de microarranjos e identificar assinaturas de genes específicos de amostras em estudos genômicos funcionais. O método proposto, o qual iremos denominar de Método Diferencial baseado na ACP (MD-ACP), utiliza duas abordagens para explorar os dados: Eigengene (EG), onde os genes são tratados como variáveis; e Eigenassay (EA), onde os ensaios (amostras) são tratados como variáveis. O método usa a contribuição do gene nas componentes da ACP para criar um ordenamento das assinaturas específicas. A contribuição dos genes é obtida pela combinação linear da matriz de pesos, obtido na ACP, e a expressão gênica das amostras.

Outro método que vem sendo utilizado para identificar genes diferencialmente expressos é o modelo linear para dados de microarranjo (limma). Diboun *et al* [25] aplicaram dois testes estatísticos diferentes (Z-score e limma) para identificar transcritos diferencialmente expressos nas amostras de tecido; e concluíram que a análise de microarranjo em amostras amplificadas é melhor na detecção de genes diferencialmente expressos, quando o método limma é utilizado. Trabalhos mais recentes têm utilizado esse método para encontrar genes, como por exemplo Yu *et al* [26] utilizaram o limma para encontrar genes de mecanismos e suas potenciais relações regulatórias em câncer de pulmão.

Além dos métodos que foram descritos, outros métodos estatísticos estão sendo utilizados para analisar dados de microarranjos. Podem ser mencionados alguns modelos como: o método bayesiano [27, 28]; os testes não paramétricos como o de Wilcoxon Mann-Whitney [29] e o de Wilcoxon Signed Rank [30].

2.3 Métodos de Classificação

Uma vez que os genes diferencialmente expressos são identificados, o próximo passo é identificar as classes conhecidas das amostras usando esses genes como preditivos, para caracterizar as propriedades dessas amostras.

Os métodos de aprendizado de máquina (AM) que comumente são usados neste campo de estudo incluem regressão logística (RL) [31], máquina de vetores de suporte (SVM, do inglês: *support vector machine*) [32], k-vizinhos mais próximos (KNN, do inglês *k-nearest neighbors algorithm*) [33] e análise discriminante linear (LDA, do inglês *linear discriminant analysis*) [34].

Muitos estudos têm usado uma grande variedade de conjuntos de dados de expressão gênica. A classificação dos dados de expressão gênica pode ser categorizada em duas abordagens amplas: a previsão de classe (*class prediction*) e a descoberta de classe (*class discovery*). Dentre eles, a primeira abordagem tem relevância devido ao enfoque nos métodos supervisionados (árvores de decisão, redes neurais e SVMs). Os métodos supervisionados aprendem com dados rotulados indicando a classe a que as amostras pertencem no conjunto de dados, o qual é composto por vários valores de expressão gênica. Os métodos supervisionados são de importância podendo ser usados para fins de diagnóstico na medicina e podem ajudar no entendimento das dependências entre classes (doenças) e características (valores de expressão gênica) [35].

O trabalho de Ramaswamy [36] utilizou o SVM para determinar um classificador para determinar o diagnóstico de câncer, através da análise da expressão gênica por microarranjo.Guyon *et al* [37] propuseram um novo método utilizando o SVM baseado em eliminação de característica recursiva (*Recursive Feaure Elimination* - RFE), que é capaz de eliminar a redundância dos genes e ao mesmo tempo obter um conjunto de genes mais compacto e razoável. Porém, Zhang *et al* [38] desenvolveram um algoritmo recursivo adaptado ao SVM (*Recursive Support Vector Machine* - R-SVM) para seleção de genes, que demonstrou uma melhor performance quando comparado ao SVM-RFE.

Khan *et al* [39] utilizaram redes neurais para a predição do tipo de câncer. O método consiste em três etapas principais: análise de componentes principais, seleção de genes relevantes e previsão de redes neurais artificiais.

Diversos estudos apontam para o uso da técnica de regressão logística, para modelagem, previsão e classificação [40] e [41].

Na obtenção de modelos de regressão, é importante destacar os testes para adequação do modelo. Para avaliar a adequabilidade e o ajuste de modelos de regressão logística, Hosemer *et al* [42] fizeram uma revisão trazendo as principais

características, vantagens e desvantagens de uma série de testes utilizados para esse fim. Três dessas medidas são utilizadas para avaliar o ajuste nesse trabalho: o teste Chi-Quadrado de Pearson, Deviance e de Hosmer-Lemeshow. El Sanharawi e Naudet[43] salientam que a regressão logística é um método de análise multivariada que permite obter uma quantificação da associação existente entre uma saída de interesse e cada um dos fatores de influência, controlando o efeito simultâneo dos outros fatores.

Segundo Montgomery e Runger [44], a coleção de ferramentas estatísticas que são usadas para modelar e explorar relações entre variáveis que estão relacionadas de maneira não determinística é chamada de análise de regressão, uma das técnicas estatísticas mais utilizadas. Muitas aplicações da análise de regressão envolvem situações em que há mais de um regressor, para as quais o modelo de regressão múltipla deve ser desenvolvido. Para os autores, a regressão linear, técnica vastamente estudada e aplicada, funciona muito bem quando a variável de resposta é quantitativa. Já na situação em que a variável de resposta é qualitativa, como no caso de que um dispositivo obtenha "sucesso" ou "falha", o modelo de regressão utilizado é a regressão logística, cujos parâmetros são geralmente estimados pelo método de máxima verossimilhança. Segundo Pardoe [45]), a análise de regressão linear múltipla é imprópria para saídas dicotômicas. Isso porque os resíduos de tal modelo não satisfariam quase nunca as quatro suposições usuais para regressão linear (média zero, variância constante, normalidade, independência). Antes de começar um estudo de regressão logística, é importante entender que o objetivo de uma análise que utiliza esse método é o mesmo de todas as outras técnicas de construção de modelos utilizada em estatística: encontrar o modelo que melhor se adere e que seja o mais parcimonioso, e ainda, razoável para descrever a relação entre uma variável de saída (dependente ou de resposta) e um conjunto de variáveis independentes (preditoras ou explicativas) [46]. Segundo Figueira [47], o modelo de regressão logística binária é advindo dos modelos lineares generalizados. Estes últimos são especificados por três componentes: uma componente aleatória, uma componente sistemática e uma função de ligação, que conecta os valores esperados das observações às variáveis explanatórias. De acordo com Hosmer, Lemeshow e Sturdivant [46], muitas funções de distribuição foram propostas para a análise de variáveis de saída dicotômicas. As razões principais para se escolher a função logística é que a mesma é extremamente flexível e facilmente utilizável e também, permite uma interpretação criteriosa.

Capítulo 3

Fundamentos Teóricos

3.1 Hanseníase

A hanseníase é uma doença infectocontagiosa de evolução crônica e sistêmica; considerada uma das doenças mais antigas que afligem o homem. A doença é amplamente conhecida pelo nome de lepra, porém no Brasil, a substituição do termo lepra por hanseníase [1] foi realizada principalmente para minimizar o forte estigma social que se faz presente até hoje. A hanseníase ainda é considerada um problema de saúde pública em alguns países, dentre os quais se destacam a Índia e o Brasil. Este último apresenta um dos maiores números absolutos de casos e prevalência acima da meta estabelecida pela Organização Mundial da Saúde (OMS), tal como se representa na Figura 3.1. Nesta figura pode ser observada a taxa de prevalência da hanseníase no mundo, segundo dados repostados pela OMS com relação ao ano de 2015. Observa-se que no Brasil, a taxa de prevalência encontra-se entre 1,0 - 2,0 por 10.000 habitantes[48].

3.1.1 Epidemiologia

O controle de uma doença infecciosa envolve redução na sua incidência e prevalência, e por consequência da mortalidade e morbidade ocasionadas pela mesma. A OMS considera a hanseníase eliminada como problema de saúde pública ao se atingir uma prevalência inferior a 1 (um) caso por 10.000 habitantes de determinada região [49]. Em 1985, o número estimado de casos de hanseníase no mundo era de 12 milhões de pessoas, correspondendo a uma prevalência de 19 casos para

[1] O termo hanseníase foi considerado em homenagem ao médico Gerhard Henrik Armauer Hansen, que identificou a bactéria Mycobacterium leprae.

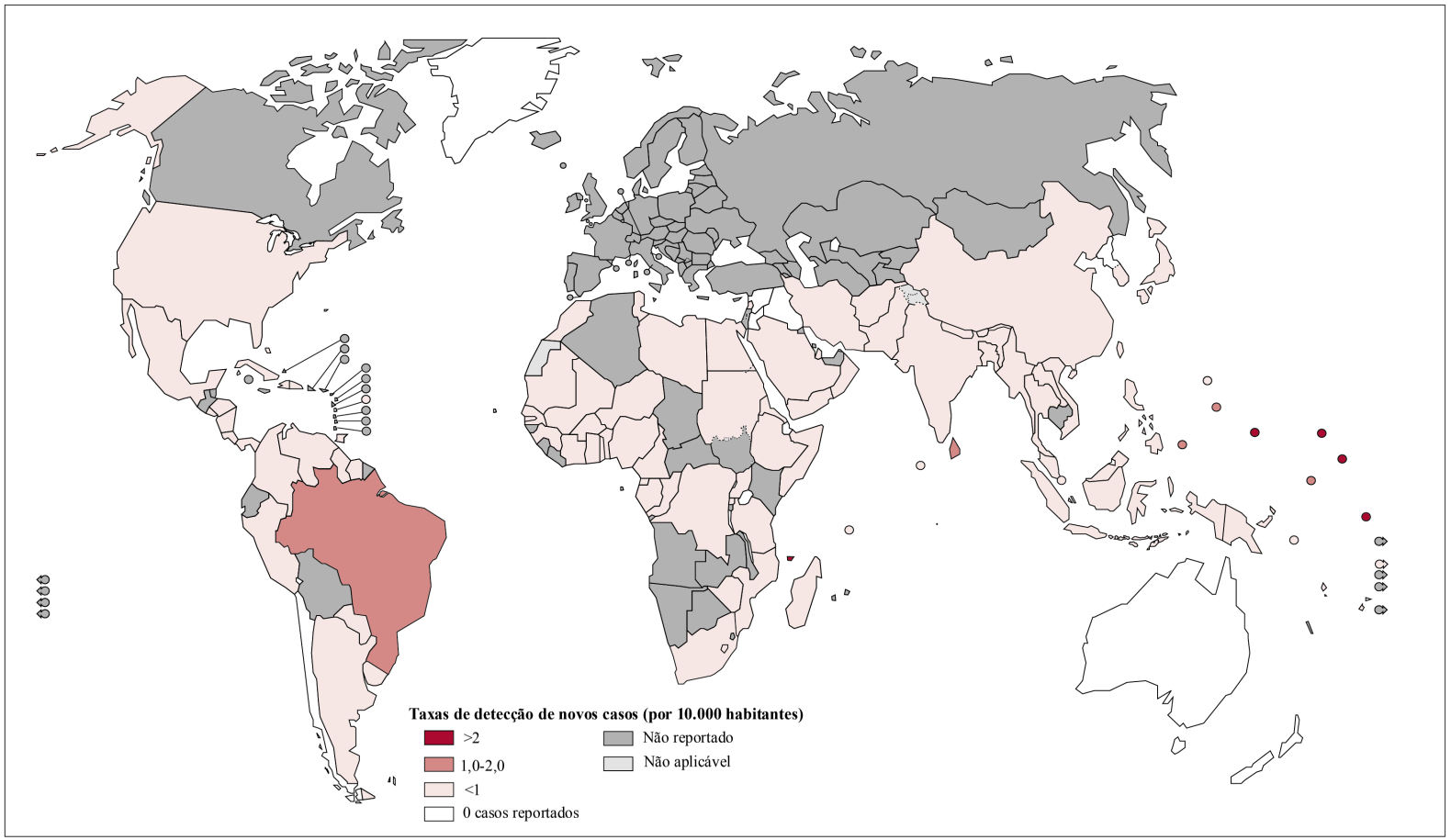


Figura 3.1: Taxa de prevalência da hanseníase no mundo. Os dados foram reportados pela Organização Mundial da Saúde e correspondem ao ano de 2015. As taxas referem-se a cada 10.000 habitantes da população.

10.000 habitantes [50]. Essa prevalência atingiu taxa de menos de um caso para cada 10.000 habitantes, com 597.000 casos novos registrados no mundo ao final do ano 2000 [51].

De acordo com relatórios enviados de 136 países e territórios à OMS, a detecção de novos casos registradas em 2015 foi de 210.758, e a prevalência global registrada no começo de 2016 foi de 174.608 casos [48]. Em todo o mundo, o Brasil se destaca, apresentando aproximadamente 13% dos casos mundiais, com 26.395 novos casos registrados em 2015, o que corresponde a 12,70 casos/100.000 habitantes, com uma prevalência de 1,15/10.000 habitantes.

Em muitos países, a redução na taxa de prevalência não foi acompanhada pela redução no número de novos casos. O diagnóstico tardio e o longo período de incubação da doença são fatores que contribuem para a transmissão ativa da hanseníase, dificultando a redução significativa no número de novos casos; isto ocorre, por exemplo, no Brasil. A Figura 3.2 mostra as taxas de detecção de novos casos de hanseníase em 2015. Pode ser observado que o Brasil se destaca entre os países com mais de 10 casos por 100.000 habitantes.

Apesar da drástica redução no número de casos registrados nas últimas duas décadas, a incidência da doença tem caído lentamente, como demonstrado na Figura 3.3. Nesta figura podemos observar os números de casos reportados no Brasil, na última década [48].

3.1.2 Transmissão

Atualmente, considera-se que a transmissão de M. leprae se dê pelo contato de pessoas suscetíveis à doença com os doentes não tratados das formas multibacilares. No entanto, acredita-se que a transmissão possa ocorrer também por pessoas sem nenhum sintoma de hanseníase, porém com a presença de M. leprae na mucosa nasal. A mucosa nasal é o sítio preferencial para entrada e saída de M. leprae [52]. Portanto, a excreção nasal de M. leprae por indivíduos infectados subclinicamente pode estar implicada na disseminação e transmissão do bacilo, em contatos domiciliares de indivíduos portadores de hanseníase multibacilares [53].

As vias aéreas superiores constituem a principal porta de entrada e via de eliminação do bacilo da M. leprae, constituindo a forma mais provável de transmissão da hanseníase entre seres humanos. Desta forma, a proximidade com os indivíduos portadores de hanseníase é um importante determinante da transmissão da doença. Os indivíduos com maior probabilidade de contrair a doença são aqueles que resi-

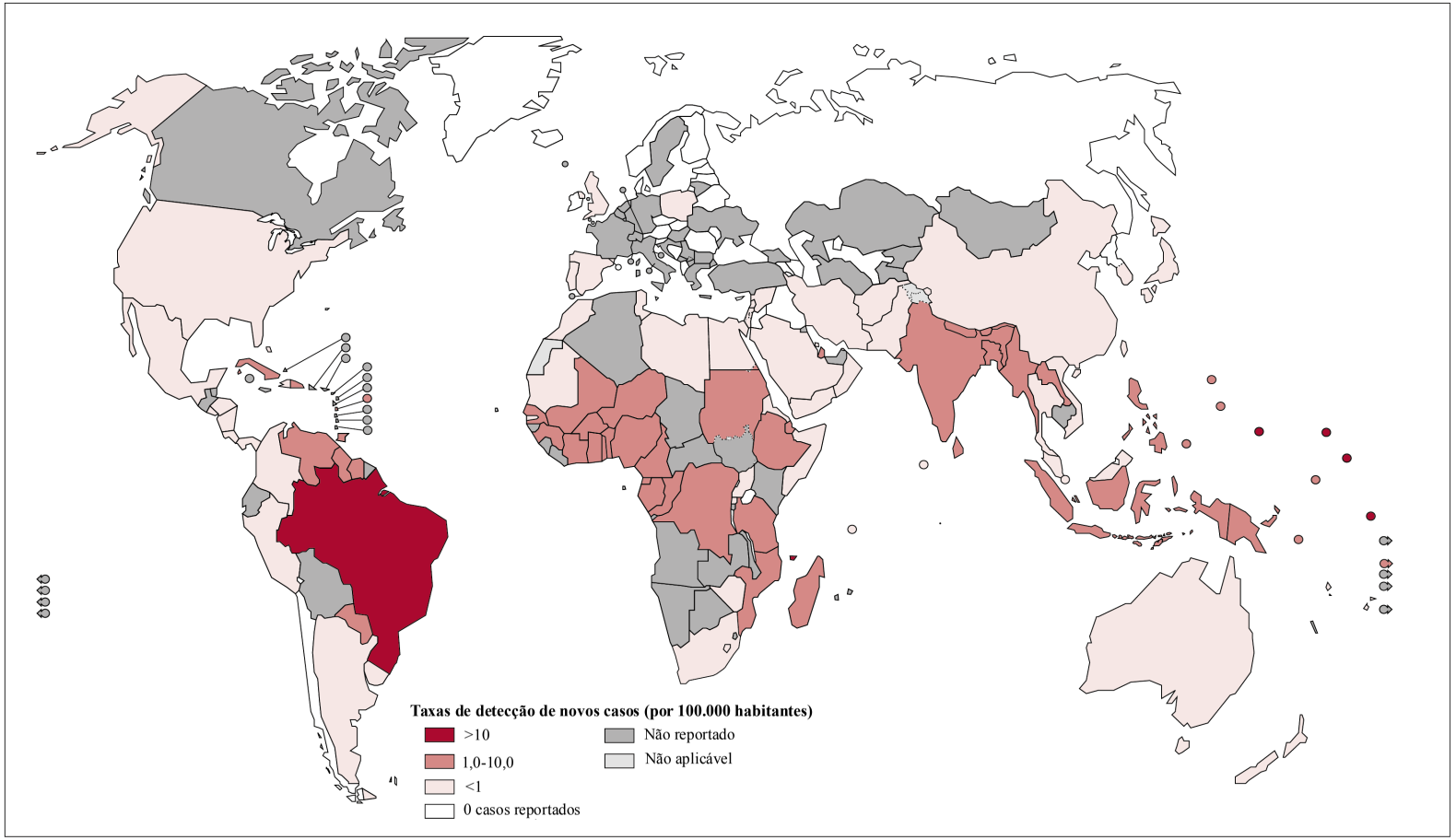
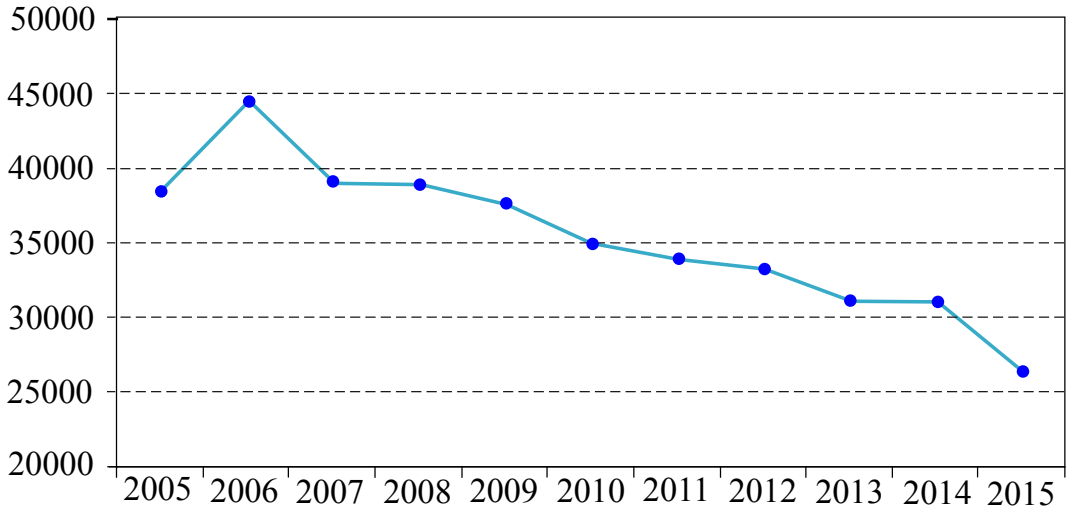


Figura 3.2: Taxa de incidência da hanseníase no mundo. Os dados são reportados pela Organização mundial da Saúde e correspondem ao ano de 2015. As taxas referem-se a cada 100.000 habitantes da população.



* Fonte: WHO (2016)

Figura 3.3: Número de casos de hanseníase na última década no Brasil.

dem na mesma casa, que vivem no peri-domicílio ou que tenham contato íntimo e frequente com o indivíduo.

Um estudo realizado recentemente por Sales *et al* [54], confirmou que o principal fator de risco entre os contatos de casos de hanseníase foi a proximidade com os doentes. Verificou-se que os contatos intradomiciliares apresentam um risco praticamente duas vezes maior de adoecer quando comparados àqueles extradomiciliares; apesar da força desta associação ser diferente para casos incidentes e co-prevalentes. Os casos incidentes são entendidos como os que adoecem no acompanhamento após o início do tratamento do caso índice. Os casos co-prevalentes são casos secundários diagnosticados no momento do exame de todos os familiares reportados pelo indivíduo índice.

Não existe um conhecimento solidificado em relação aos prováveis fatores de risco implicados na transmissão da hanseníase. No entanto, o trabalho de Schaible *et al* [55] descreve que alguns fatores são apontados para novos casos de aquisição de hanseníase, tais como: idade, relação genética, resposta imune celular e humoral, entre outros. Por outro lado, o risco de adoecimento é também aumentado pelo parentesco, onde a relação de consanguinidade de primeiro grau é um fator muito importante, indicando um papel fundamental de fatores genéticos no estabelecimento da infecção por M. leprae, que por sua vez é modulada por fatores ambientais [56–58]. Sendo assim, a hanseníase apresenta um caráter multifatorial, onde o elevado risco de desenvolver hanseníase possui uma contribuição tanto genética quanto de fatores ambientais [59], e embora a infecção pelo M. leprae seja necessária, ela não é suficiente para o estabelecimento da doença.

A infecção causada pelo M. leprae apresenta um longo período de latência, onde o tempo médio de incubação nos indivíduos pode variar de quatro até mais de dez anos

[60]. Acredita-se que este longo período de incubação contribua para a transmissão da hanseníase, pois o indivíduo já poderia estar transmitindo a doença antes mesmo de apresentar manifestações clínicas. Após o contato do M. leprae com o hospedeiro, o bacilo pode penetrar no corpo através das vias aéreas superiores. Posteriormente, o bacilo pode se espalhar em áreas de baixas temperaturas como as extremidades do corpo (pés, mãos e face).

3.1.3 Diagnóstico

O diagnóstico de hanseníase é clínico e baseia-se no achado de um ou mais sinais, como: presença de lesões cutâneas com perda de sensibilidade, nervos periféricos espessados e a detecção de bacilos álcool-ácido resistentes no raspado intradérmico. As lesões de pele na hanseníase podem se manifestar como manchas, placas, infiltração cutânea, nódulos, tubérculos, alopecia, madarose, triquíase e lesões de mucosa [3]. Os distúrbios sensitivos nas lesões podem ser causados tanto pela ação do bacilo nos nervos como pela reação do sistema imune ao bacilo. A neurite pode levar a perda de força muscular, com posterior paralisia. Esta pode originar incapacidades e deformidades, tais como: articulações anquilosadas (sem movimento) e em garras, ectrópio (reviramento da pálpebra) e lagoftalmia (eversão e desabamento da pálpebra inferior) [61]. Outras deformidades também podem se originar de uma infecção secundaria grave, devido a traumas em regiões sem sensibilidade. Isto pode causar necrose, levando muitas vezes a amputação da região ou a osteomielite, que é um processo de reabsorção das extremidades ósseas, particularmente os dedos dos pés e mãos.

Cerca de 30% dos indivíduos não manifestam nenhum dos sintomas acima, dificultando o diagnóstico clínico [62]. Aliado a isto, a resistência dos indivíduos em procurar atendimento médico torna o diagnóstico clínico frequentemente tardio, com aparecimento de sintomas irreversíveis como a perda de sensibilidade ou incapacidade motora.

A hanseníase pode ser confundida com outras doenças de pele (Ptiríase Versicolor, Leishmaniose cutânea, Eczemátide, Vitiligo, micoses cutâneas etc) e com outras doenças neurológicas (Síndrome do túnel do carpo, neuralgia parestésica, neuropatia diabética, lesões por esforço repetitivo, etc) que apresentam sinais e sintomas semelhantes. Portanto, deve ser feito diagnóstico diferencial em relação a essas doenças [63]. Existem poucos exames laboratoriais utilizados na rotina para auxiliar no diagnóstico da hanseníase, tais como: baciloscopia, histopatologia, teste de Mitsuda, sorologia e reação em cadeia da polimerase (PCR), porém estes testes não fazem parte da rotina das unidades de saúde. Embora a sorologia e o PCR sejam rei-

vindicadas por profissionais de saúde como ferramentas auxiliares, estes testes são utilizados exclusivamente em pesquisa.

Em áreas não endêmicas, é muito comum não considerar a hanseníase no diagnóstico diferencial, podendo levar à detecção tardia de casos [64]. Este atraso no diagnóstico pode trazer consequências para os indivíduos, como dano neural irreversível e sequelas em vários graus até a deficiência física [51].

Ainda não está disponível nenhum teste sorológico que pudesse ser implementado rotineiramente no diagnóstico da hanseníase. Imunoensaios tem sido desenvolvido para detectar anticorpos contra o antígeno PGL-1 do M. leprae [65]. Entretanto, esses ensaios ainda não possuem satisfatório grau de sensibilidade e especificidade para a aplicação de diagnóstico.

Contudo, o diagnóstico laboratorial teve grandes avanços nos últimos anos e foram desenvolvidos métodos para a extração, amplificação e identificação de DNA do M. leprae em amostras clínicas. Esse diagnostico utiliza a técnica de reação em cadeia da polimerase (PCR) e outras técnicas moleculares. Alguns genes foram utilizados no desenvolvimento do ensaio de PCR para a detecção do bacilo do M. leprae em amostras clínicas [66–68]. Esses ensaios têm sido muito úteis em casos de difícil diagnóstico da doença, devido a sua alta especificidade e sensibilidade, possibilitando a sua utilização em quase todos os tipos de amostras clínicas, incluindo linfa, sangue, secreção nasal e biopsias de tecido. Embora a técnica ainda seja cara e não esteja disponível como um teste de rotina, a PCR pode fornecer um complemento excelente para diagnóstico clínico e histopatológico da hanseníase [68].

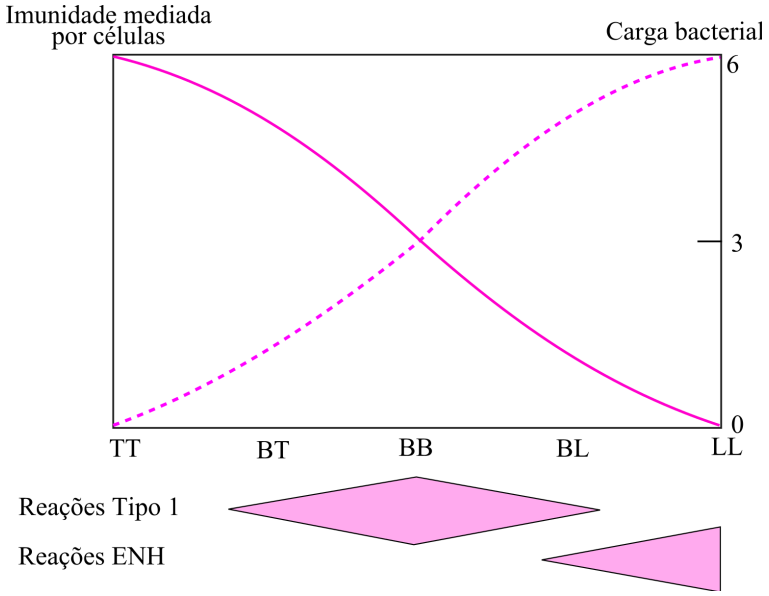
3.1.4 Classificação

As principais manifestações clínicas da hanseníase dependem da resposta imune do hospedeiro, que influencia não só na suscetibilidade, mas também no curso da infecção [3]. Elas incluem lesões na pele, nas mucosas e nos nervos. Devido ao amplo espectro de manifestações clínicas em que se apresenta, a hanseníase é classificada segundo suas características clínicas, histopatológicas e imunológicas por dois sistemas principais: classificação de Madri (definida no VI Congresso Internacional de Hanseníase em 1953) [4] e Classificação de Ridley & Jopling [3].

Na classificação de Madri, consideram-se dois polos estáveis e opostos (virchowiano e tuberculóide) e dois grupos instáveis (indeterminado e dimorfo). A classificação proposta por Ridley & Jopling [3] é amplamente utilizada em pesquisas e leva em consideração a imunidade dentro de um espectro de resistência do hospedeiro, como mostrado na Figura 3.4. Considerando a Figura 3.4, observa-se que na forma polar tuberculóide (TT), a carga bacilar é baixa, com poucas lesões na pele,

predomina a resposta imune celular com baixa ou nenhuma produção de anticorpos. No outro extremo, observa-se a forma lepromatosa (LL), com alta carga bacilar, numerosas lesões na pele, alta produção de anticorpos e baixa ou ausente resposta imune celular. Além das formas polares descritas, existem ainda três formas intermediárias e clinicamente instáveis chamadas de: borderline-tuberculóide (BT), borderline-borderline (BB) e borderline-lepromatosa (BL) [3].

Os primeiros sinais da hanseníase são dificilmente notados pelos indivíduos ou até mesmo por profissionais da saúde. Nesses casos, a classificação é denominada como forma indeterminada [69]. A doença ainda pode se apresentar sob a forma neural pura, sem manifestações cutâneas e bacilos escassos, mesmo em biópsias de nervos [70].



* Adaptado de Ridley e Jopling [3]

Figura 3.4: Formas clínicas da hanseníase. Esquema demonstra o perfil espectral da doença. Representação baseada na classificação de Ridley e Jopling. Estão incluídos aspectos da resposta imune do indivíduo e os episódios reacionais.

Além desta classificação espectral proposta por Ridley e Jopling [3], a OMS propõe uma subdivisão simplificada dos indivíduos para fins de tratamento. Neste caso, os indivíduos são classificados em: multibacilares e paucibacilares. Os multibacilares são caracterizados pela presença de carga bacilar positiva na biopsia, de seis ou mais lesões cutâneas e esfregaço positivo para a presença da bactéria. Estes incluem as formas LL, BB e BL. Os paucibacilares são caracterizados pelo aparecimento de raros bacilos, apresentando cinco ou menos lesões cutâneas e esfregaço negativo. Estes incluem as formas TT e BT.

3.1.5 Reações Hansênicas ou Episódios Reacionais

Ao longo do curso natural da doença, os indivíduos podem desenvolver reações, as quais representam episódios inflamatórios que se intercalam no curso crônico da hanseníase. As reações são divididas em reação do tipo I (ou Reação Reversa - RR), que preferencialmente ocorre nas formas BL, BB e BT; e reação do tipo II (ou Eritema Nodoso Hansênico - ENH) que ocorre mais frequentemente nas formas BL e LL. Ambas as reações são caracterizadas por uma reativação do processo inflamatório. Esses processos inflamatórios podem culminar em lesões na pele e neurite aguda (no caso da reação do tipo I) e até mesmo complicações sistêmicas (no caso da reação do tipo II) [71]. Os episódios reacionais são as maiores complicações em hanseníase. Estes podem ocorrer em indivíduos virgens de tratamento, durante e/ou após o tratamento [72], deixando geralmente sequelas.

Reação tipo I ou reação reversa (RR) é um episódio de inflamação aguda na pele e nervos periféricos, resultado de uma hipersensibilidade tardia contra antígenos do bacilo, que ocorre em aproximadamente 30% dos indivíduos [73].

A reação tipo II geralmente se manifesta como eritema nodoso (ENH) ou multiforme e pode ocorrer em indivíduos do polo lepromatoso. Esta reação se apresenta como um quadro febril debilitante com o aparecimento de nódulos papulares cutâneos, acompanhado de inflamação nos nervos, olhos e testículos [51].

Os diferentes tipos de reação parecem ter mecanismos imunológicos distintos que ainda são pouco compreendidos.

3.1.5.1 Eritema Nodoso Hansênico

O Eritema Nodoso Hansênico (ENH) é uma das principais causas de danos permanentes nos nervos periféricos, podendo causar deformidades e incapacidades observadas na hanseníase. Acredita-se que o ENH seja um episódio relacionado à formação de complexos imunes com consequente ativação do sistema complemento (que é o principal mediador humoral do processo inflamatório junto aos anticorpos) [74]. Dados recentes sugerem que a resposta imune mediada por células desenvolve um importante papel na evolução do ENH [75].

A maioria dos indivíduos apresenta múltiplos episódios agudos de curta duração ou ENH crônico com durabilidade maior do que 6 meses [76], podendo se prolongar durante vários anos [77]. A forma LL e o índice de bacilos 4 representam fatores de risco significativos para o desenvolvimento do ENH [76, 78, 79]. Em 1999, o trabalho de Manandhar *et al* [78] demonstrou que o infiltrado dérmico também representa um fator de risco. Outros fatores de risco como gravidez, lactação, puberdade, infecção

intercorrente, vacinação e estresse psicológico têm sido associados [80], embora não tenham sido confirmados em estudos prospectivos.

Uma das diferenças histológicas entre o ENH e forma LL é a infiltração de neutrófilos, característica nas lesões do ENH. Em 2010, Lee *et al* [81] investigaram os mecani*smos* de recrutamento de neutrófilos no local da doença. O perfil de expressão gênica das lesões do ENH compreende a ativação da via integrada de TLR2 e do receptor Fc, migração de neutrófilos e inflamação. Os principais aspectos desta via incluem: - FcR ou TLR2 que induzem a liberação de IL-1; - ativação do endotélio, incluindo a regulação positiva de E-selectina e subsequente ligação de neutrófilos; e - regulação positiva dos mediadores inflamatórios associados tanto com neutrófilos e monócitos / macrófagos [81].

Segundo a OMS [82], o ENH deverá ser tratado como grave, diante da presença de qualquer um dos seguintes sinais: dor ou incômodo em um ou mais nervos com ou sem perda de função; ulceração dos nódulos de ENH; dor ou vermelhidão nos olhos com ou sem diminuição da acuidade visual; edema dolorido nos testículos (orquite) ou nos dedos (dactilite).

3.2 Microarranjo

Os microarranjos de DNA podem prover informação importante para o diagnóstico de enfermidades através dos níveis de expressão gênica, devido à habilidade de medir a quantidade de RNA transcrito para milhares de genes simultaneamente [83]. Os microarranjos visam compreender as variações moleculares nos processos relacionados às doenças celulares.

Os microarranjos surgiram como uma alternativa para detectar e quantificar transcritos em larga escala, sendo uma ferramenta poderosa para identificação e quantificação de ácidos nucleicos. A habilidade de identificar e extrair informações funcionais de genes em nível de DNA têm concentrado esforços em todo o mundo [84, 85]. De modo geral, a análise de expressão gênica fornece informações importantes sobre o funcionamento de uma célula, já que alterações genéticas modificam diretamente a produção de proteínas necessárias para o seu funcionamento.

As tecnologias baseadas em microarranjos de DNA fornecem um meio relativamente simples para analisar simultaneamente a expressão de milhares de genes [85]. Esse paralelismo aumenta consideravelmente a velocidade dos experimentos e permite comparações significativas entre os produtos gênicos representados no microarranjo [86]. Entre outros usos, a análise dos dados produzidos por microarranjos se tornou a técnica padrão para monitorar níveis de expressão gênica em laboratórios de biologia molecular [87].

Os microarranjos de DNA complementar (cDNA), também chamado de Chip de DNA, operam através do princípio de hibridização de moléculas com sequências homólogas ou complementares [11]. Nessa técnica, milhares de fragmentos de DNA são impressos ordenadamente em lâminas, sejam de oligonucleotídios, cDNA ou ESTs (Expressed Sequence Tags) representativos de genes. A célula é estudada em diferentes condições (por exemplo, amostra e controle) e seu mRNA é extraído. A detecção da hibridização é possível, pois durante a síntese, os ácidos nucleicos das amostras são marcados com moléculas fluorescentes que geram um sinal. O sinal gerado pode ser detectado por meio de leitores (scanners) de fluorescência que é utilizado para analisar os resultados.

Os dados de microarranjos são geralmente organizados como uma matriz M $(n\,x\,m)$, onde as linhas (n) representam os genes e as colunas (m) representam as amostras biológicas cujo nível de expressão foi monitorado. Matematicamente, o nível de expressão gênica é dado pela razão entre as intensidades observadas para a amostra e o controle [88] e para a normalização dos dados. Isto é, para que estes se aproximem de uma distribuição normal, é comum transformar-se o valor do nível de expressão pelo logaritmo. O objetivo de muitos experimentos de microarranjos é identificar genes que são diferencialmente transcritos, no que diz respeito a diferentes condições biológicas de culturas de células ou amostras de tecido.

O perfil de expressão pode ser usado para comparar o nível de transcrição do gene em condições clínicas, a fim de: 1) identificar biomarcadores de diagnóstico ou de prognóstico; 2) classificar doenças; 3) monitorar a resposta à terapia; e 4) compreender os mecanismos envolvidos na gênese dos processos da doença [21]. Por estas razões, microarranjos de DNA são considerados instrumentos importantes para descobertas na medicina clínica.

Os microarranjos de DNA são coleções de sondas (segmentos de fitas simples de DNA) distribuídas ordenadamente sobre uma superfície rígida feita de vidro ou silicone, chamada lâmina ou slide. Cada lâmina comercialmente disponível pode conter milhares de sondas distribuídas em um ou até oito arranjos, dependendo do fabricante. Cada sonda representa um único gene, porém coletivamente elas podem representar o genoma inteiro de determinado organismo. Sobre a superfície do slide, as sondas são posicionadas em locais específicos, os spots.

3.2.1 Tipos de microarranjos

Os microarranjos podem ser classificados de acordo com três critérios:

1. Comprimento das sondas (segmentos de fitas simples de DNA);
   (a) Matrizes de DNA complementar (cDNA): usam longas sondas de centenas ou milhares de pares de bases (bps);
   (b) Matrizes de oligonucleotídeo: utilizam sondas curtas de 50bps ou menores.
2. Método de fabricação;
   (a) Deposição: sequência de matrizes anteriormente sintetizadas;
   (b) Síntese in situ: fabricação de matrizes de cDNA usando a tecnologia in situ. Existem os seguintes tipos de síntese:
      i. fotolitografia;
      ii. impressão por jato de tinta;
      iii. síntese eletroquímica ou impressão de contato.
3. Número de amostras que podem ser analisadas simultaneamente em uma matriz.
   (a) Matrizes de canal único: analisa uma única amostra por vez;
   (b) Matrizes de canais múltiplos: pode analisar duas ou mais amostras simultaneamente.

Na fotolitografia, cada sonda (oligonucleotídeos de 25 bases) é sintetizada diretamente na superfície do slide, neste caso chamado de chip (fixação *in situ*). Estes chips de DNA são conhecidos como arranjos de DNA de alta densidade. É uma metodologia robusta e capaz de comportar até 500.000 sondas por chip. Possui alta especificidade e reprodutibilidade, no entanto é uma técnica bastante onerosa e com flexibilidade limitada, uma vez que os equipamentos de fixação e detecção são geralmente de uso restrito dos fabricantes. Já nas tecnologias que utilizam a impressão a jato de tinta ou de contato, as sondas são pre-sintetizadas para somente depois serem ligadas à superfície da lâmina de vidro (hibridização por spotting).

Existem três principais tipos de uso para os microarranjos de DNA em medicina:

1. Encontrar diferenças nos níveis de expressão entre grupos pré-definidos de amostras (comparação de classe);

2. Identificar associações de classes de uma amostra com base no perfil de expressão do gene (predição de classe);

3. Analisar um conjunto de expressões gênicas com o objetivo de encontrar subgrupos que tenham características em comum (descoberta de classe).

3.3 Método de análise da expressão gênica

A implementação de uma análise de expressão gênica bem-sucedida requer o desenvolvimento de vários protocolos laboratoriais, bem como o desenvolvimento de bases de dados e ferramentas de software para uma coleta e análise dos dados eficientes. Apesar de vários protocolos terem sido publicados de forma detalhada, as ferramentas computacionais necessárias para tal análise estão evoluindo a cada dia, embora, sem um consenso de qual seria a melhor técnica para encontrar padrões na expressão gênica. De fato, está cada vez mais claro que pode nunca existir uma abordagem que seja considerada a melhor de todas e, portanto, a aplicação de várias técnicas pode fornecer diferentes perspectivas dos dados explorados. Assim, a escolha de quais técnicas/algoritmos utilizar para fazer a análise dos dados é uma parte crucial no projeto dos experimentos envolvendo expressão gênica [85].

3.3.1 Teste t de *Student*

Os métodos de seleção de genes diferencialmente expressos univariados baseiam-se na utilidade marginal de cada variável (gene) na discriminação dos grupos, sendo as variáveis ordenadas de acordo com um determinado critério que reflita essa discriminação. As primeiras variáveis do conjunto ordenado, em geral, são posteriormente utilizadas como variáveis independentes em métodos de classificação. Entre os critérios mais usuais, encontram-se os valores p obtidos para testes t.

O teste t de Student é utilizado por alguns investigadores para testar as diferenças significativas expressas por um gene em duas condições diferentes. A principal ideia por trás deste método é a obtenção de um valor de p para cada um dos genes envolvidos e, em seguida, tomar uma decisão com a ajuda de testes estatísticos e o valor de p.

O teste t é um teste de hipóteses paramétrico tradicional para a seleção de genes diferencialmente expressos [89]. A estatística t segue uma distribuição t de student, sob o pressuposto de normalidade dos níveis de expressão. O valor de probabilidade (valor de p) para cada expressão do gene é a probabilidade de conseguir que o teste t seja tão ou mais extremo do que o observado sob a hipótese de ausência de expressão

diferencial (hipótese nula). Neste teste, o valor de p normalmente corresponde a uma hipótese nula, indicando que os genes não são expressos diferencialmente em dois ou mais grupos do experimento. Um valor de p pequeno indica que a hipótese nula de não haver expressão diferencial não é verdadeira e o gene é diferencialmente expresso.

Podemos calcular o valor de t para cada gene usando seguinte equação:

$$t_i = \frac{\mid m_{i1} - m_{i2} \mid}{\sqrt{\frac{(n_1-1)s_{i1}^2+(n_2-1)s_{i2}^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{3.1}$$

onde n_1e n_2 é o número de amostras em cada grupo e i é o i-gésimo gene; m_{i1} e m_{i2} é a média dos grupos e s_{i1}^2 e s_{i2}^2 é o desvio-padrão de cada grupo, respectivamente, para o i-gésimo gene. O valor de p é obtido a partir da distribuição t de Student para a estatística t_i calculada. Um gene é considerado como significativo (diferencialmente expresso), quando o seu valor de p é menor que algum limiar, geralmente escolhido como 0,01 ou 0,05.

Quando muitas hipóteses são testadas, a probabilidade de que o erro tipo I seja cometido aumenta significativamente com o número de hipóteses. O erro tipo I, também chamado de falsos positivos, é o erro que se comete ao rejeitar a hipótese nula quando ela é verdadeira [90]. Comete-se o erro do tipo I ao declarar que um gene apresentava expressão gênica diferencial quando na verdade, isso não ocorre.

Benjamini e Hochberg [90] propuseram controlar a FDR (False Discovery Rate), definida como a proporção de hipóteses nulas H_0 verdadeiras, entre as hipóteses nulas rejeitadas, ou seja, a proporção de erros devido à rejeição errônea de H_0 .

Vamos considerar que tenhamos m hipóteses, H_1, H_2, ... , H_m, e tenhamos os respectivos valores p, p_1, p_2, ... , p_m. Assumindo-se que toleramos uma proporção esperada de falsos positivos igual a q, podemos ordenar em ordem crescente esses, $p_{(1)}$, $p_{(2)}$, ... , $p_{(m)}$, sendo $H_{(i)}$ a hipótese correspondente à $p_{(i)}$. Dado k, o maior valor de i para onde tenhamos $p_i \leq \frac{iq}{m}$, rejeitamos todas as hipóteses $H_{(i)}$, $i = 1, 2, \ldots, k$. Ao invés de trabalharmos com os valores p, podemos obter valores $q_i = \frac{mp_{(i)}}{i}$ e definir o menor valor de q para o qual rejeitamos as hipóteses.

Por exemplo, se definimos 5% FDR, iremos identificar como diferencialmente expressos os genes com valor q menor do que 0.05. Para uma melhor compreensão deste procedimento, suponhamos que 200 genes foram declarados como diferencialmente expressos. Ao nível de 5% FDR teríamos que desses 200 genes, 10 seriam falsos positivos.

A técnica da FDR vem sendo empregada com sucesso em microarranjos, não porque haja muitos tratamentos, mas porque há muitas variáveis respostas, o que

leva à necessidade de reduzir o número de conclusão de falsos positivos para poder manejar, em experimentos posteriores, apenas as variáveis mais promissoras.

3.3.2 Modelagem Linear para Dados de Microarranjo (LIMMA)

Smith [91] propôs uma estatística denominada de estatística t-moderada buscando contornar o problema comum em experimentos com microarranjos, número de ensaios n pequeno e número muito grande de variáveis (genes). A interpretação da estatística t-moderada é a mesma da estatística t ordinária sendo que os erros padrões são estimados para os genes utilizando métodos Bayesianos empíricos. A diferença para a t ordinária está nos graus de liberdade que são aumentados no caso da t-moderada.

O método Bayesiano empírico assume uma distribuição à priori para os dados e os parâmetros dessa distribuição são estimados a partir dos dados. O modelo proposto por Smith ajusta um modelo linear para determinar a expressão diferencial. O modelo linear é dado por

$$E\left[y(i)\right] = X\beta(i) \tag{3.2}$$

onde $y(i)$ contém os dados de expressão para o gene i, X é a matriz de desenho e $\beta(i)$ é um vetor de coeficientes. Tendo em conta que estamos interessados em comparar dois grupos, no nosso caso expressão gênica entre indivíduos LL e ENL, podemos definir uma matriz de contraste $C^T = [-1\ 1]$.

O modelo linear é ajustado aos dados para obter estimativas $\hat{\beta}(i)$ para os coeficientes, e derivar a variância residual $s^2(i)$ do modelo linear. O teste t-moderado é utilizado para obter a estatística para a expressão diferencial entre duas condições, e esse é dado por:

$$\tilde{t}(i) = \frac{\hat{\beta}(i)}{u(i)\tilde{s}(i)} \tag{3.3}$$

sendo $u(i)$ o desvio padrão dos dados não padronizados e $\tilde{s}^2(i)$ é a variância à posteriori residual do modelo linear definido por

$$\tilde{s}^2(i) = \frac{f_0 s_i^2 + f_i s_i^2}{f_0 + f_i} \tag{3.4}$$

onde, f_0 e f_i são os valores à priori e número de graus de liberdade, e s_0^2 e s_i^2 são os valores à priori e a variância residual para o modelo linear, respectivamente. A estatística segue uma distribuição t com graus de liberdade $f_0 + f_i$ se $\beta(i) = 0$. Os

graus de liberdade extra f_0 representa a informação extra que é obtida do conjunto de genes para fazer inferência sobre cada gene.

3.3.3 Método Diferencial baseado na Análise de Componentes Principais (MD-ACP)

O MD-ACP proposto por Rollins e Teh [24] busca um subconjunto ordenado de genes relevantes para distinguir duas condições em amostras biológicas, no nosso caso diferenciar as duas condições - ENH do LL. O método inicia transformando o conjunto de variáveis originais (genes), possivelmente correlacionadas, num conjunto de variáveis não correlacionadas, usando ACP.

A análise de componentes principais é um dos métodos mais populares de análise multivariada, devendo-se a sua origem aos trabalhos de Pearson [92] e Hotelling [93]. A ACP é uma técnica matemática que tem como um de seus objetivos básicos reduzir a dimensionalidade dos dados. É uma técnica que transforma um conjunto de variáveis originais em outro conjunto de variáveis, denominadas de componentes principais (CP), linearmente não correlacionadas e de mesma dimensão. Cada componente é uma combinação linear de todas as variáveis, mantendo a variância e covariância estrutural do conjunto de variáveis originais [94].

A ACP se baseia na obtenção dos autovalores da matriz de covariância ou da matriz de correlação dos dados. Considere-se um conjunto de dados padronizados descritos pela matriz $X(m\,x\,n)$, em que cada elemento dessa matriz é representada por $x_{ij}, i = 1, 2, \ldots, m$ variáveis e $j = 1, 2, \ldots n$ experimentos. A matriz de covariância M é descrita por:

$$M_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} \left(x_{ik} - \bar{x_i}\right) \left(x_{jk} - \bar{x_j}\right), \quad i, j = 1, \cdots, m \tag{3.5}$$

onde os valores de $\bar{x_i}$ e $\bar{x_j}$ são as médias das variáveis i e j, respectivamente.

Usando a matriz M podemos calcular os autovetores ortonormais l_k (ou seja, os coeficientes) e os autovalores λ_k (ou seja, a variância), relativos a cada k-ésima CP, $k = 1, 2, \ldots, n$. Geometricamente, essas CPs representam a seleção de um novo sistema de coordenadas obtido pela rotação do sistema original [92]. A primeira CP explica o máximo de variabilidade possível e cada componente subsequente responde pelo máximo possível da variabilidade remanescente.

Dadas duas variáveis, X e Y, podemos fazer um gráfico de dispersão e observar como os pontos relativos as observações se distribuem no par de eixos originais (Fig. 3.5). Com a ACP definimos dois novos eixos, ortogonais entre si, e projetando os dados originais nesse novo par de eixos obtemos as coordenadas z_1 e z_2 para cada

ponto. Observe na Figura 3.5 que o eixo z_1 está na direção de maior variabilidade dos dados, e o segundo eixo, está na direção de menor variabilidade dos dados. Essa é uma das características que possibilita o uso da ACP para redução de dimensionalidade de um conjunto de dados, retendo as componentes que explicam a maior parcela de variabilidade.

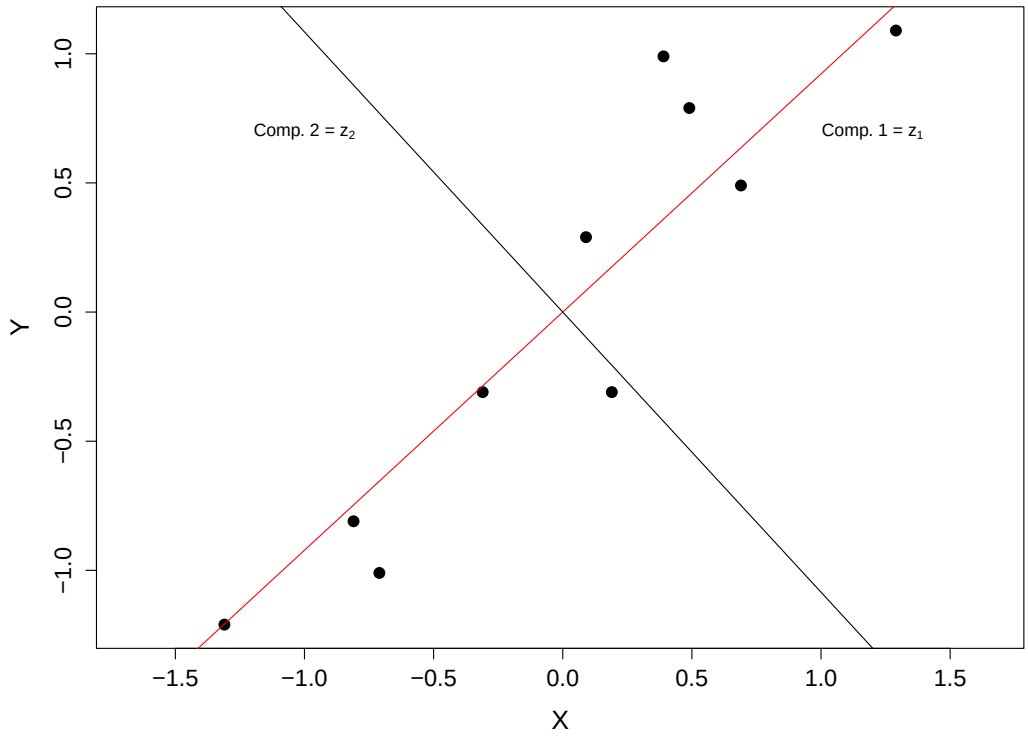


Figura 3.5: Representação gráfica dos dados originais, X - eixo horizontal e Y - eixo vertical, e da rotação obtida com a Análise de Componentes Principais, resultando nos eixos z_1 e z_2.

Um ponto importante na ACP é a escolha do número de CPs a ser retida. Para isto, utilizando como exemplo a área de intersse biológica, podem ser utilizadas duas abordagens para selecionar o número de componentes: podem-se usar componentes que se correlacionam com um fenótipo de interesse [95] ou usar componentes suficientes para incluir a maior parte da variação nos dados [39].

A projeção das amostras sobre qualquer CP, ou seja, o mapeamento dos dados sobre as componentes é denominado escore. Os escores dos dados S_{ij} podem ser obtidos por meio de uma combinação linear das variáveis originais e vetor de coeficientes das CPs, expressas da seguinte forma:

$$S_{ij} = l_{1j}x_{i1} + l_{2j}x_{i2} + \cdots + l_{nj}x_{im} = \sum_{k=1}^{n} l_{kj}x_{ik} \tag{3.6}$$

$$i = 1, \cdots, m;\ j = 1, \cdots, n;\ k = 1, \cdots, n,\ k = 1, \ldots n$$

sendo S_{ij} o valor do ponto e l_{ij} o valor da carga para a i-ésima variável e a j-ésima componente principal; k representa o número do ensaio.

Se projetarmos os dados originais em cada CP, obtendo os respectivos escores, poderemos visualmente identificar qual das CPs possibilita uma melhor visualização de grupos presentes nos dados, assumindo que conhecemos os rótulos ou classes de cada observação. A figura 3.6 mostra um exemplo hipotético, onde temos que a primeira CP sugere visualmente ser melhor para identificarmos os grupos nos dados.

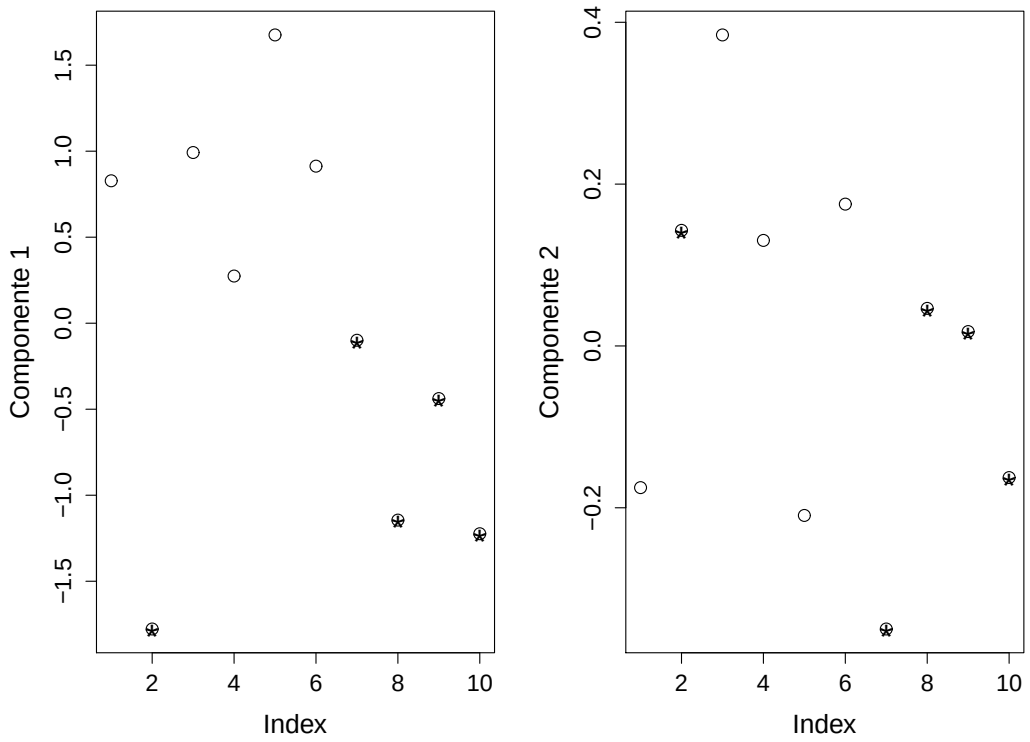
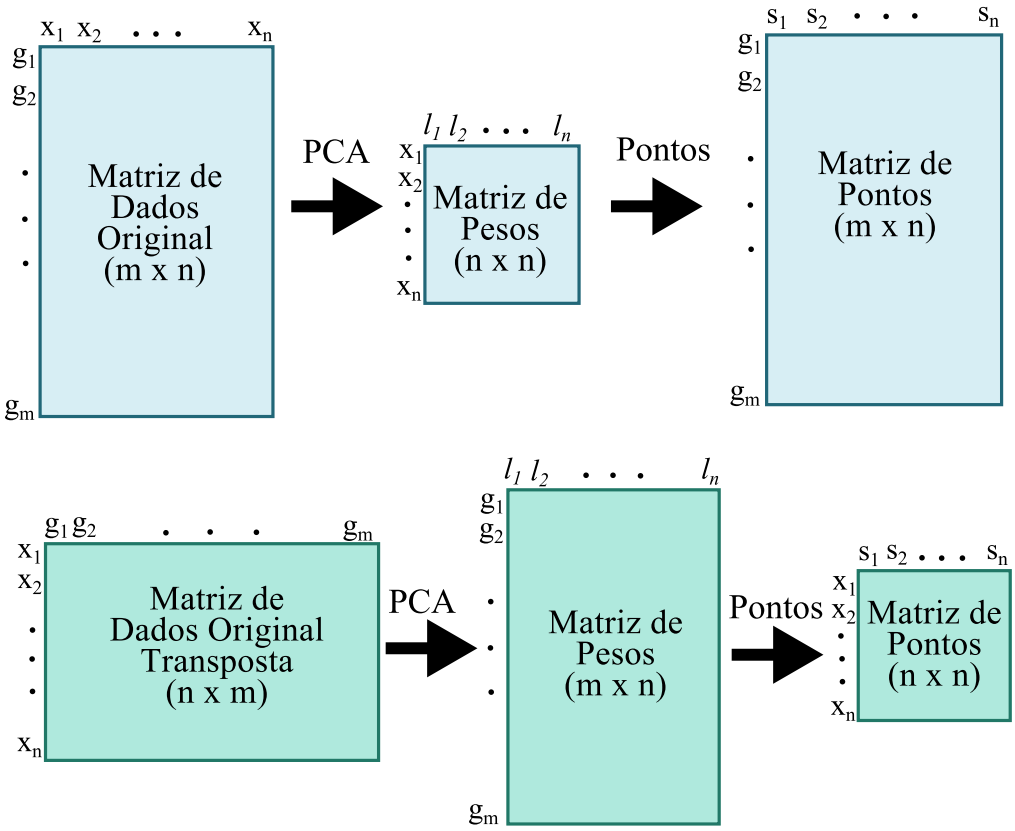


Figura 3.6: Gráfico de escores para as duas primeiras Componentes Principais para conjunto de dados com dois grupos, grupo A representado por "*" e grupo B representado por "o". Observa-se que na primeira Componente Principal visualizamos uma separação boa entre os dois grupos.

Para dados de microarranjo temos uma matriz de dados $X\,(m\,x\,n)$, onde n é o número de ensaios ao longo das colunas e m representa o número de genes ao longo das linhas. As células nesta matriz são dadas como x_{ij} que é o nível de expressão do i-ésimo gene para o j-ésimo ensaio. O método DM-ACP [24] se baseia inicialmente em obter a ACP para a matriz original dos dados, X, e também obter a ACP para a matriz de dados transposta, X^T, isto é, considerando como variáveis

os experimentos. Quando consideramos os dados originais X, as CPs obtidas serão denominadas de eigengenes (EG), e quando obtemos as CPs para a matriz transposta essas serão denominadas de eigenassays (EA). Na parte superior da Figura 3.7 temos esquematicamente o processo de obtenção da matriz de pesos $L\,(n\,x\,n)$ (coeficientes) das CPs (correspondente à contribuição de cada gene nas PCs); e a matriz de escores $S\,(m\,x\,n)$ (correspondente às coordenadas de cada amostra projetadas nas CPs). A parte inferior da Figura 3.7 mostra o procedimento para obtenção dos escores a partir da matriz transposta dos dados originais X^T[24].



* Adaptado de Rollins *et al* [96]

Figura 3.7: Representação visual das matriz de dados obtidas através da Análise de Componentes Principais.

Uma vez que as CPs são estimadas, podemos utiliza-las para visualizar em qual dessas teremos uma melhor distinção entre grupos da amostra de interesse. Uma vez identificada a CP que apresenta maior discriminação visual entre os grupos, a contribuição de cada gene nessa CP pode ser estimada usando as equações 3.11 e 3.14, descritas mais adiante.

Para obter a matriz de escores para o caso EG, o procedimento é padronizar X, resultando em uma matriz Z, onde cada elemento é igual a $z_{ij} = \frac{x_{ij} - \bar{x_j}}{s_j}$, sendo x_{ij} o valor da expressão gênica do gene i na amostra j; e $\bar{x_j}$ e s_j a média e o desvio

padrão da amostra j, respectivamente. Para obtermos os escores de EA, não há necessidade de padronizamos X.

Os elementos da matriz escore EG, S^{EG} , são determinados pela seguinte equação:

$$S_{ij}^{EG} = \sum_{k=1}^{n} l_{kj}^{EG} z_{ik} = \sum_{k=1}^{n} g_{ijk}^{EG} \tag{3.7}$$

$$i = 1, \cdots, m;\ j = 1, \cdots, n;\ k = 1, \cdots, n$$

sendo S_{ij}^{EG}o escore e l_{ij}^{EG} a carga para o i-ésimo gene usando a j-ésima componente principal de EG; e g_{ijk}^{EG} a contribuição do i-ésimo gene na k-ésima amostra da j-ésima componente principal de EG.

Na abordagem eigenassay (EA) utiliza-se a matriz transposta de$X$, tratando os ensaios como variáveis. Seguindo Rollins [96], X^T não é padronizado na abordagem EA, como foi utilizado na abordagem, EG. Os elementos da matriz de scores, S^{EA}, são determinados segundo a equação:

$$S_{ij}^{EA} = \sum_{p=1}^{m} l_{pj}^{EA} x_{pi} = \sum_{p=1}^{m} g_{ijp}^{EA} \tag{3.8}$$

$$i = 1, \cdots, n;\ j = 1, \cdots, n;\ p = 1, \cdots, m$$

sendo S_{ij}^{EA}o escore e l_{ij}^{EA} a carga para a i-ésima amostra usando a j-ésima componente principal de EA; e g_{ijk}^{EA} a contribuição do p-ésimo gene na i-ésima amostra da j-ésima componente principal de EA.

Após a obtenção da CPs para as duas propostas, EA e EG, selecionamos a CP que visualmente apresente uma maior separação entre os grupos das amostras.

Como temos dois grupos nas amostras, A (ENH) e B (LL), com n_a e n_b número de amostras, a média de contribuição do i-ésimo gene na j-ésima componente principal para cada um dos grupos A e B, para o caso EG é obtida por:

$$\bar{g}_{ij}^{EG_A} = \frac{1}{n_a} \sum_{over\ k'} l_{k'j}^{EG} z_{ik'} = \frac{1}{n_a} \sum_{over\ k'} g_{ijk'}^{EG} \tag{3.9}$$

$$\bar{g}_{ij}^{EG_B} = \frac{1}{n_b} \sum_{over\ k''} l_{k'j}^{EG} z_{ik''} = \frac{1}{n_b} \sum_{over\ k''} g_{ijk''}^{EG} \tag{3.10}$$

sendo k' e k'' varia de $k' = 1, \ldots, n_a$ e $k'' = 1, \ldots, n_b$.

Finalmente, a diferença de contribuição gênica ($\bar{dg}_{ij}^{EG}$), da abordagem EG entre os grupos A e B para o i-ésimo gene na j-ésima componente, é dada como:

$$\bar{dg}_{ij}^{EG} = \bar{g}_{ij}^{EG_A} - \bar{g}_{ij}^{EG_B} \tag{3.11}$$

Da mesma forma, as média de contribuição do p-ésimo gene na j-ésima componente principal para os grupos A e B, para o caso EA são obtidas por:

$$\bar{g}_{ij}^{EA_A} = \frac{l_{pj}^{EA}}{n_a} \sum_{over\ i'} x_{pi'} = \frac{1}{n_a} \sum_{over\ i'} g_{i'jp}^{EA} \tag{3.12}$$

$$\bar{g}_{ij}^{EA_B} = \frac{l_{pj}^{EA}}{n_b} \sum_{over\ i''} x_{pi''} = \frac{1}{n_b} \sum_{over\ i''} g_{i''jp}^{EA} \tag{3.13}$$

sendo i' e i'' as amostras que fazem parte dos grupos A e B , respectivamente.

A diferença de contribuição gênica, da abordagem EA entre os grupos A e B para o p-ésimo gene na j-ésima componente principal, é dada por:

$$d\bar{g}_{ij}^{EA} = \bar{g}_{ij}^{EA_A} - \bar{g}_{ij}^{EA_B} \tag{3.14}$$

Uma vez selecionada a CP que visualmente apresenta uma melhor separação dos grupos, isto é, a CP do caso EG ou a CP do caso EA, calculamos as diferenças acima e fazendo-se a ordenação podemos apresentar o gráfico da contribuição diferencial para os grupos A e B. A seleção dos genes a serem inicialmente considerados é feita identificando-se o joelho da curva, e os genes com valores acima desse ponto de corte são selecionados. Uma vez que a contribuição diferencial poder ser sobre expressa, ou seja, gene i tem expressão maior no grupo A, ou subexpressa, onde gene i tenha maior expressão no caso B, temos que considerar dois pontos de corte. O primeiro joelho significativo é o joelho mais ao extremo da curva. A determinação desses pontos de corte se baseia na derivada segunda da curva empírica das contribuições diferencias ordenadas.

3.4 Modelos Classificadores

3.4.1 Regressão Logística

O modelo de regressão logística (RL) é um modelo linear generalizado (MLG), sendo um tipo de análise de regressão muito utilizado para realizar predições ou explicar a ocorrência de um evento específico quando a variável dependente (variável resposta) é de natureza binária. Quanto às variáveis independentes, estas podem ser tanto quantitativas quanto qualitativas.

A regressão logística apresenta três componentes: um componente aleatório, que consiste nas observações da variável aleatória; uma componente sistemática, que relaciona as variáveis independentes com os parâmetros do modelo; e uma função de

ligação, neste caso, a função logit, que conecta os valores esperados das observações às variáveis independentes [97].

A regressão logística foi usada por Zhu e Hastie [98] e Liao e Chin [31] para desenvolver modelos de predição de uma variável resposta binária a partir de dados de microarranjos. Neste modelo de regressão, a variável resposta (Y) é dicotômica, ou seja, são atribuídos a ela dois valores: 1 para o acontecimento de interesse, denominado sucesso e 0 para o acontecimento complementar, o fracasso. A probabilidade do sucesso é dada por π_i e a de fracasso é $1 - \pi_i$.

A variável resposta, indicando a presença ou ausência de uma doença (reação hansênica, por exemplo), é definida de forma adequada do seguinte modo:

$$Y = \begin{cases} 1, & apresenta\, reação\, hansênica \\ 0, & não\, apresenta\, reação\, hansênica \end{cases}$$

Considerando-se uma série de variáveis aleatórias independentes x_1, x_2, ..., x_n, onde x_p representa o nível de expressão do p-ésimo gene e um vetor $\beta = \beta_0, \beta_1, ..., \beta_p$ formado por parâmetros desconhecidos do modelo. A probabilidade de sucesso é dada por:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})} \tag{3.15}$$

e a probabilidade de fracasso é:

$$1 - \pi_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})} \tag{3.16}$$

O logit para o modelo de regressão múltipla é dado por:

$$g\left(x_1\right) = \ln\left[\frac{\pi_i}{1 - \pi_i}\right] = x_i^T \beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{3.17}$$

E o logaritmo da função de verossimilhança pode ser escrito como:

$$l\left(\beta\right) = \sum_{i=1}^{n} \left[y_i x_i^T \beta - \ln\left(1 + \exp\left\{x_i^T \beta\right\}\right)\right] \tag{3.18}$$

A RL prediz a probabilidade de um evento ocorrer, podendo ser um valor entre 0 e 1. Na determinação dos coeficientes da regressão, o método utilizado é o da máxima verossimilhança. O processo iterativo a ser utilizado pode ser o de Newton-Raphson, derivando $l\left(\beta\right)$ em relação a cada parâmetro [97].

Para implementar o modelo estatístico RL é necessário aplicar um conjunto de treino, no qual é considerado como 70% das amostras do conjunto total dos dados simulados. A significância do modelo final obtido pela RL é verificada através do

desvio entre o modelo ajustado e o modelo saturado, onde todos os parâmetros se ajustam perfeitamente a todas as observações. O modelo mais simples é o denominado modelo nulo, formado apenas pelo parâmetro β_0, indicando toda variação ao componente aleatório.

Os coeficientes do modelo foram estimados pelo método de máxima verossimilhança, que estima os valores dos parâmetros permitindo maximizar a probabilidade dos dados observados. Para avaliar quão bom é o ajuste do modelo estatístico, foi utilizado o critério de informação de Akaike (AIC*, Akaike Information Criterion*) [99]. Os valores do AIC fornecem um meio para a seleção do modelo. Este critério penaliza modelos com muitas variáveis sendo que valores menores de AIC são preferíveis. O AIC é dado pela fórmula:

$$AIC = -2log\left(L_p\right) + 2\left(p\right) \tag{3.19}$$

sendo L_p a função de máxima verossimilhança do modelo e p o número de parâmetros a serem estimados no modelo.

A significância dos efeitos do modelo é estimada pela análise de *deviance*, a qual indica a qualidade do ajuste. Quanto melhor for o ajuste do modelo aos dados, menor será o valor da *deviance*. O desempenho do modelo final é avaliado pela análise de resíduos com o intuito de verificar se estes apresentam aleatoriedade e se existem valores discrepantes. Para que o modelo ajustado seja considerado satisfatório, é necessário que as *deviances* residuais se ajustem a uma linha reta.

A estatística *deviance* (D) se baseia na função de verossimilhança e é utilizada para medir a discrepância de um modelo intermediário de p parâmetros em relação ao modelo saturado. Quanto menor a deviance, melhor o ajuste do modelo [100]. Esta estatística pode ser expressa como segue:

$$D = -2\ \ln \frac{(verossimilhança\,do\,modelo\,ajustado)}{(verossimilhança\,do\,modelo\,saturado)} \tag{3.20}$$

Após obter o modelo ajustado, é necessário verificar se o modelo apresenta uma boa descrição dos dados que foram observados, através da análise de resíduos e diagnósticos. Chamamos de Análise dos Resíduos um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos. O resíduo (e_i) é dado pela diferença entre a variável resposta observada (Y_i) e a variável resposta estimada $\left(\hat{Y}_i\right)$, isto é: $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - ... - \hat{\beta}_p x_{ip} \qquad i = 1, ..., n$. A ideia básica da análise dos resíduos é que, se o modelo for apropriado, os resíduos devem refletir as propriedades impostas pelo termo de erro do modelo.

Um problema que pode ser observado em dados binários é a estimativa de parâmetros de um modelo de regressão logística binária ajustadas usando o método de máxima verossimilhança, que às vezes não convergem para valores finitos. Esse fenômeno (também conhecido como probabilidade monótona ou parâmetros infinitos) ocorre por causa de uma condição entre os pontos de amostra conhecidos como separação [101]. Existem duas classes de separação.

Quando a separação completa está presente entre os pontos de amostra, os procedimentos iterativos para maximizar a probabilidade tendem a falhar, quando seria claro que há um problema com o modelo. O problema de detectar a separação entre os pontos de amostra está intimamente relacionado com o problema de encontrar separabilidade linear entre os dois grupos de pontos de amostra. Em particular, no caso de uma resposta binária, a existência de separabilidade linear é equivalente ao caso de separação completa.

O conceito de separabilidade linear tem sido amplamente estudado no contexto do reconhecimento e classificação de padrões. Ripley [102], por exemplo, define dois grupos de pontos de amostra para ser linearmente separável se houver uma combinação linear das covariáveis $x^T\beta$ que seja positiva para os pontos de amostra em um grupo e negativa para os pontos de amostra em o outro. Há uma separação completa entre um conjunto de n pontos de amostra se houver um vetor β tal que:

$$\beta^T x_i < 0 \ \ quando \ \ y_i = 0 \quad e \quad \beta^T x_i < 0 \ \ quando \ \ y_i = 0 \tag{3.21}$$

para $i = 1, \ldots, n$. Ou seja, qualquer vetorβ que dê uma separação completa corretamente prediz a resposta y_i dada a x_i para cada uma das n amostras. Para simplificar a afirmação de separação completa, seja um vetor de n elementos tais que $\hat{y}_i = 1$ quando $y_i = 1$ e $\hat{y}_i = -1$ quando $y_i = 0$ para $i = 1, \ldots, N$ e define $\hat{X} = diag(\hat{y})X$. Então um vetor β dá separação completa entre os pontos de amostra se $\hat{X}\beta > 0$. Esse tipo de separação é ilustrada na Figura 3.8, ou seja, a presença de separabilidade linear entre grupos implica na separação completa entre os grupos de amostras.

3.5 Avaliação do Classificador

Para avaliar o desempenho de um classificador, pode-se utilizar a validação cruzada [103]. *A valid*ação cruzada é uma técnica amplamente utilizada em problemas envolvendo a modelagem e a predição de um desfecho. A validação cruzada envolve o particionamento do conjunto original de dados em subconjuntos mutualmente exclusivos, e posteriormente, agrega-se alguns destes subconjuntos para a estimação dos

Figura 3.8: Exemplo de separação completa: os pontos de amostra do grupo E1 são representados por círculos abertos e os pontos de amostra do grupo E2 por círculos preenchidos.

parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo. Um método bastante usado é o k-*fold*, onde particiona-se o conjunto de dados em k subconjuntos, agrega-se k-1 desses para formar o conjunto de treinamento e o k-ésimo conjunto é usado como conjunto teste. Repete-se o processo k vezes, sempre deixando como conjunto de teste um outro subconjunto. Dessa forma, pode-se obter as médias dos valores de acurácia, sensibilidade, especificidade e índice Kappa obtidos pelos k conjuntos de teste.. As curvas ROC ("Receiver-Operating Characteristic") são obtidas para cada um dos classificadores. Através dessas curvas podem ser calculadas as médias da área abaixo da curva (AUC - "Area Under Curve").

As principais medidas de avaliação podem ser calculadas a partir da matriz de confusão para duas classes, onde são representados quatro tipos de classificação, segundo o resultado do preditor (Tabela 3.1):

Tabela 3.1: Matriz de confusão para duas classes.

		Valor Verdadeiro	
		Positivo	*Negativo*
Valor preditivo	*Positivo*	*Verdadeiro Positivo (VP)*	*Falso positivo (FP)*
	Negativo	*Falso negativo (FN)*	*Verdadeiro Negativo (VN)*

- Classificados como positivos e pertencentes à classe positiva (verdadeiros positivos, VP);
- Classificados como negativos, mas pertencentes à classe positiva (falsos negativos, FN);
- Classificados como positivos, mas pertencentes à classe negativa (falsos positivos, FP);
- Classificados como negativos e pertencentes à classe negativa (verdadeiros negativos, VN).

A partir dessas classificações, podem-se definir três medidas de desempenho comuns: acurácia, sensibilidade e especificidade.

A acurácia (A) é definida como a proporção de acertos do modelo. Ela é dada pela fórmula:

$$A = \frac{VN + VP}{VP + FN + FP + VN} \tag{3.22}$$

Para determinar o número de acertos do modelo final é necessário estabelecer uma probabilidade, denominada ponto de corte. Probabilidades estimadas pelo modelo que sejam maiores ou iguais a esse ponto de corte recebem valor 1, enquanto que probabilidades estimadas pelo modelo que sejam menores do que esse ponto de corte recebem valor 0.

Considerando uma aplicação a uma doença, a probabilidade de um teste diagnóstico produzir um resultado positivo, dado que o indivíduo é portador da doença, é chamada sensibilidade do teste. Por outro lado, a probabilidade do teste de produzir um resultado negativo, dado que o indivíduo não é portador da doença, é chamada especificidade.

A sensibilidade (S) é dada pela proporção de verdadeiros positivos em relação ao total da classe positiva. A sensibilidade é calculada por:

$$S = \frac{VP}{VP + FN} \tag{3.23}$$

Um modelo, ao apresentar alta sensibilidade, raramente classificará como pertencente à classe negativa um valor da classe positiva. Ou seja, este fornecerá uma pequena taxa de falsos negativos.

A especificidade (E) compreende a proporção de verdadeiros negativos em relação ao total da classe negativa e é calculada através da seguinte equação:

$$E = \frac{VN}{FP + VN} \tag{3.24}$$

Um modelo, ao apresentar alta especificidade, dificilmente classificará como pertencente à classe positiva um valor da classe negativa, fornecendo, dessa forma, poucos resultados falsos positivos.

Outra medida para avaliar o desempenho de um modelo de classificação é conhecida como curva ROC (Receiver-Operating Characteristic). Essa curva é obtida traçando-se aos pares a sensibilidade e a taxa de falsos positivos (1 - E) a cada ponto. Ela mostra a severidade do balanço entre a sensibilidade e a especificidade de um teste e pode ser utilizada na decisão do melhor ponto de corte. A área sob a curva ROC é uma medida resumo usual de precisão, já que ela é estimada levando em consideração as sensibilidades e especificidades relativas a cada um dos valores estipulados [104].

Na curva ROC, o eixo vertical é a função verdadeiramente positiva (sensibilidade) e o eixo horizontal é em função da fração falsamente positiva (especificidade), para diferentes pontos de corte. Cada ponto da curva ROC representa o par sensibilidade/especificidade correspondente a um limiar particular. Quanto maior a área da curva, maior a precisão do teste [104], sendo que a área abaixo da curva ROC (Area Under Curve, AUC) é um dos índices mais utilizados para verificar a "qualidade" da curva, estando associada ao poder discriminante do teste. Quando se obtém um valor de AUC igual a 0,5 considera-se um desempenho ao acaso, enquanto um valor próximo de 1,0 indica uma predição quase perfeita. A Figura 3.9 mostra uma curva ROC hipotética para casos de AUC de valores iguais a 1, 0,85 e 0,5.

Cohen [105] define Kappa (K) como um coeficiente de concordância entre dados da classificação e verdade de campo para escalas nominais. O índice Kappa mede o grau de concordância entre as diferentes técnicas além do que seria esperado pelo acaso. Ele é calculado pela divisão da diferença entre a concordância esperada e a concordância observada e a diferença entre a concordância absoluta e a concordância esperada. Como esta última diferença representa a maior concordância possível entre a esperada e a observada, quanto maior é o índice Kappa, maior é a concordância entre as observações. O coeficiente de concordância K pode ser determinado pela equação:

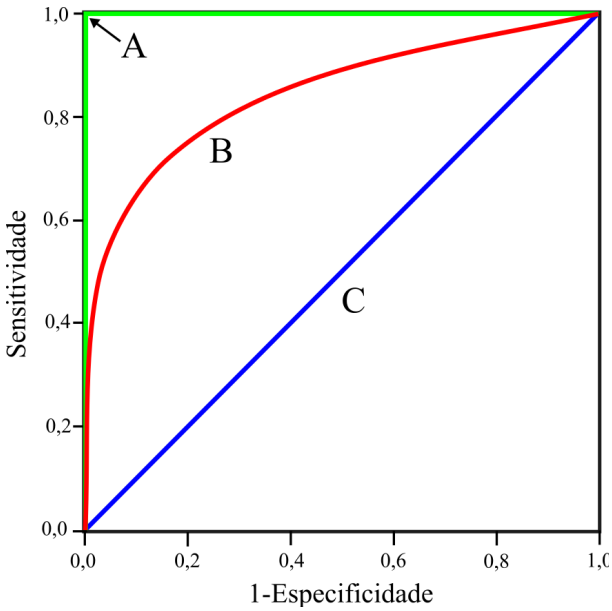


Figura 3.9: Curva ROC hipotética. A curva A representa AUC=1; a curva B representa um AUC=0,85; e a curva C representa um AUC=0,5.

$$K = \frac{P_0 - P_c}{1 - P_c} \tag{3.25}$$

onde P_0 é a exatidão global (proporção de unidades que concordam), representada pela equação:

$$P_c = \frac{\sum_{i=1}^{M} n_{i+} n_{+i}}{N^2} \tag{3.26}$$

onde M é o número de classes; n_{i+} é o total de elementos classificados para uma categoria i ; n_{+i} é o total de elementos de referência amostrados para uma categoria i; e N é número total de amostras.

Landis e Koch [106] definiram diferentes faixas para os valores de Kappa, segundo o grau de concordância que eles sugerem, formando a chamada escala de Landis. O índice Kappa pode variar de –1, indicando discordância completa, a +1, sugerindo concordância completa. Valores entre 0,80 e 1,0 indicam concordância quase perfeita; entre 0,60 e 0,79 representam concordância considerável; entre 0,40 e 0,59 indicam concordância moderável; entre 0,20 e 0,39 uma concordância razoável; entre 0 e 0,19 uma concordância baixa e valores menores que 0 não indicam concordância. As faixas de desempenho do índice Kappa estão resumidas na Tabela 3.2. O resultado da estatística Kappa normalmente é comparado aos valores contidos na Tabela 3.2, proposta por Landis e Koch [106].

Tabela 3.2: Tabela de índice *Kappa.*

Índice Kappa	Desempenho
$k \leq 0$	Péssimo
$0 < k \leq 0,2$	Ruim
$0,2 < k \leq 0,4$	Razoável
$0,4 < k \leq 0,6$	Bom
$0,6 < k \leq 0,8$	Muito Bom
$0,8 < k \leq 1$	Excelente

Fonte: Adaptada de Landis e Koch [106].

Capítulo 4

Materiais e Métodos

Nesta seção apresentam-se os dados e a sequência de métodos utilizados no presente trabalho. O diagrama de fluxo apresentado na Figura 4.1 resume os procedimentos realizados. A descrição de cada etapa é apresentada abaixo.

4.1 Conjunto de Dados

Para o presente trabalho foram utilizados três conjuntos de dados de microarranjo com amostras de indivíduos com hanseníase. Esses conjuntos de dados foram selecionados a partir do banco de dados do sítio NCBI (*Gene Expression Omnibus* - GEO [107]).

O primeiro conjunto de dados [81], é composto por 13 amostras de indivíduos portadores de hanseníase, sendo que destes 7 foram classificados com a forma lepra lepromatosa (LL) e 6 que evoluíram para a forma reacional eritema nodoso hansênico (ENH). Esse conjunto de dados foi utilizado para fazer as análises de identificação de genes diferencialmente expressos e para construir o modelo preditivo de classificação. Para construir um conjunto único para a validação do modelo, utilizamos dois conjuntos de dados. Do primeiro conjunto de dados [108], foram utilizadas as 6 amostras de ENH e do segundo [109], foram utilizadas as 7 amostras de LL. Os conjuntos contêm as expressões gênicas de 54.675 sondas ativas, as quais representam 20.721 genes únicos. Em todos os conjuntos de dados, a matriz utilizada para a hibridação de cada amostra foi *Affymetrix Human Genome U133 Plus 2.0 Array*. Os conjuntos de dados podem ser encontrados em GEO sob os números GSE17763, GSE16844 e GSE8489, e pode ser acessado pelo endereço `https://www.ncbi.nlm.nih.gov/geo`.

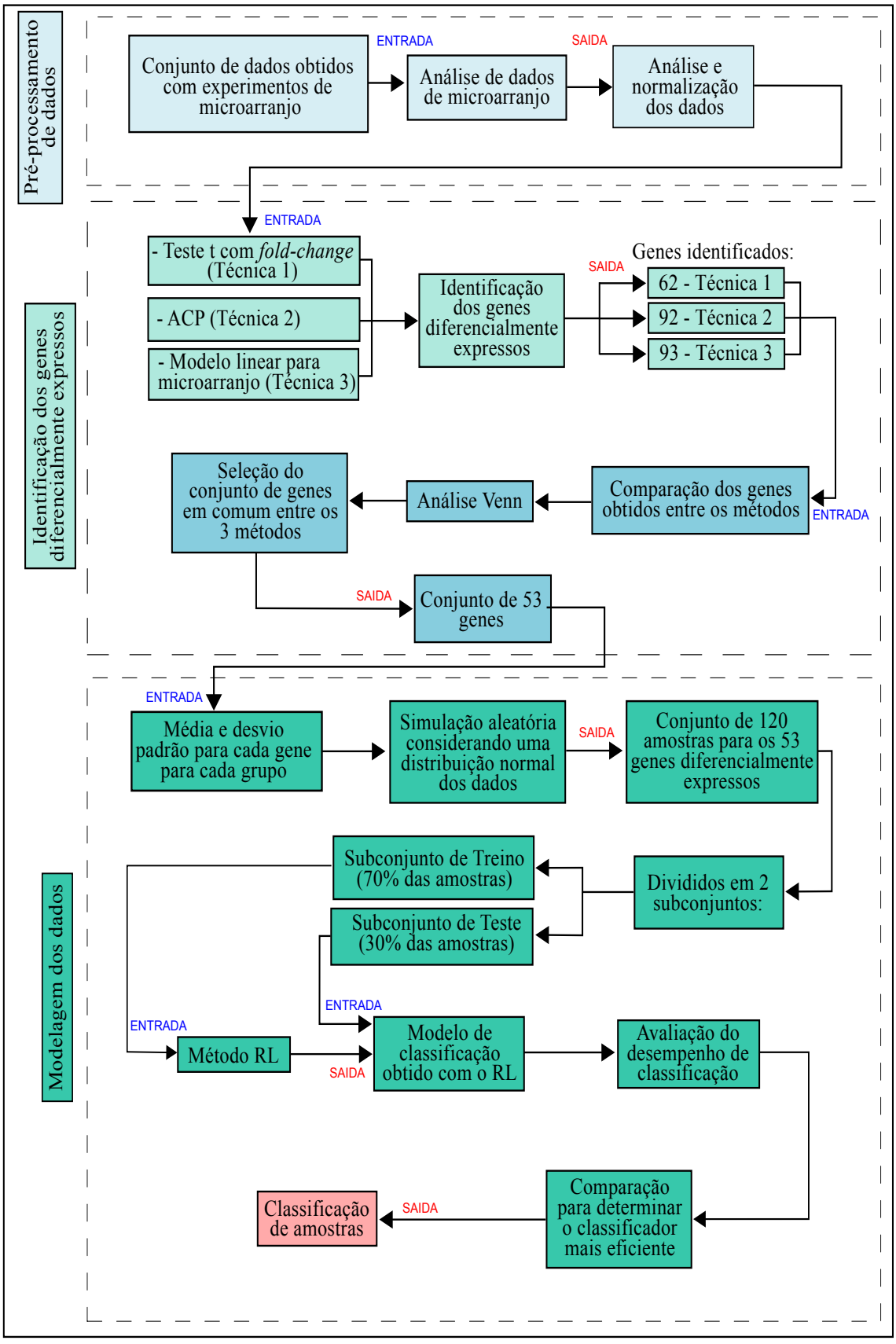


Figura 4.1: Fluxograma que apresenta a sequência das análises efetuadas.

As matrizes de dados são formadas por $m = 54.675$ genes e $n = 13$ amostras, podendo ser representada pela matriz 4.1 de expressão gênica:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \vdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \tag{4.1}$$

onde x_{mn} é a quantificação do nível de expressão do gene m na amostra n.

4.2 Metodologia

As principais etapas metodológicas apresentadas no presente trabalho foram: análise estatística descritiva (seção 4.2.1), identificação de genes diferencialmente expressos (seção 4.2.2) e classificação dos dados (seção 4.2.3). Para implementar a metodologia, foi utilizado o *software* R versão 3.3.1 `https://cran.r-project.org/`.

4.2.1 Pré-processamento de dados

A análise descritiva dos dados é a etapa inicial da presente metodologia, a qual foi utilizada para descrever e resumir os dados. Para a análise descritiva foram utilizados o histograma e o *boxplot*, como análise gráfica.

Normalização

Foi utilizado o método RMA (*Robust Multichip Average*) [110] para a correção do ruído de fundo, normalização e sumarização das sondas de microarranjo de DNA da plataforma Affymetrix. Essa escolha foi feita pelo fato do RMA ser um dos métodos mais utilizados para analisar dados dessa plataforma [111], além de apresentar alta especificidade e sensibilidade quando comparado com outros métodos [112].

Cada microarranjo pode introduzir um viés nas intensidades obtidas pelo experimento. Para diminuir os efeitos destas variações indesejadas e tornar as intensidades observadas comparáveis entre diferentes microarranjos, são necessários métodos de normalização. Em particular, o RMA aplica a normalização quantílica. Esse método de normalização assume que as intensidades verdadeiras em cada microarranjo são oriundas da mesma distribuição. Assim, para reduzir os vieses dos experimentos, a normalização quantílica transforma os dados de forma que cada quantil seja o mesmo em todos os microarranjos [113].

4.2.2 Identificação dos genes diferencialmente expressos

Após a normalização dos dados, o próximo passo foi a identificação dos genes diferencialmente expressos. Como existem diversas técnicas para tal finalidade utilizamos três delas para obter um resultado mais robusto.

A primeira técnica corresponde ao teste t-Student, que é usado para determinar se o gene tem expressão diferencial em duas condições. O valor de p do resultado do teste indica a probabilidade de um gene ser diferencialmente expresso. Como tem-se diversos testes t, usamos as correções para taxas de falso positivo (FDR) de acordo com a metodologia proposta por Benjamini-Hocheberg [90]. Para controlar a taxa de erro global se utilizou o ajuste de FDR com o nível de significância (α) igual a 5%.

Como segunda técnica de identificação, utilizamos a modelagem linear para dados de microarranjo. Para essa técnica utilizou-se o pacote *limma* [91] do projeto Bioconductor. O *limma* é um pacote para análise de dados de expressão gênica em dados de microarranjos. Ele utiliza modelos lineares para fazer as análises, permitindo fazer comparações de muitas amostras simultaneamente e tem características que tornam a análise estável até mesmo em experimentos com pequeno número de amostras.

Com a técnica limma, foi ajustado um modelo linear fixo para cada gene, em que os valores da variável resposta y_{gij} são iguais ao $\log_2\left(\frac{ENL_{gij}}{LL_{gij}}\right)$, sendo considerado o seguinte modelo:

$$y = X\alpha + \varepsilon \tag{4.2}$$

em que y é uma matriz de valores M para o gene g, X é uma matriz de delineamento 2x1, $\alpha_g = (ENL - LL)$ é um vetor 1x1 de parâmetros desconhecidos e ε é o erro aleatório.

Como o número de amostras utilizadas é pequeno utilizou-se o método bayesiano empírico [91], pois a estimativa de variância da estatística t moderada utilizada nesse método é mais estável que a estatística t ordinária. A estatística de teste usada para avaliar se o coeficiente $\beta = 0$ foi a estatística t-moderada. Calculou-se para cada gene o valor de p ajustado para múltiplos testes, utilizando o método FDR (*False Discovery Rate*) [90]. Isto teve como objetivo aumentar o poder estatístico e simultaneamente reduzir o risco de falsos positivos [91]. Os valores selecionados serão aqueles que obtiveram um valor de p igual ou menor que 5%.

O terceiro método usado foi o MD-ACP, utilizando as duas abordagens propostas para obtenção da ACP: eigengene (EG) e eigenassay (EA) [96]. Foram determinadas as componentes principais (EG e EA) com base na matriz de correlação. Para obter as matrizes de escores, o procedimento é padronizar X para obter os escores de EG,

mas não para os escores de EA, uma vez que não padronizamos X para determinar EA.

Após a obtenção da CPs para as duas propostas (EA e EG), selecionamos a CP que visualmente apresente uma maior separação entre os grupos. A análise visual é feita utilizando os gráficos com os valores das componentes para os EGs e gráficos com os valores da matriz de escores para os EAs, de acordo com as equações 3.6 e 3.8 descritas na fundamentação teórica. A CP que apresenta um perfil com maior separação entre as classes é selecionada para determinação dos genes diferencialmente expressos. A última etapa é obtenção da diferença de contribuições gênicas de acordo com a equação 3.11, se CP escolhida for a obtida pela análise EG ou de acordo com a equação 3.14, se CP escolhida foi a da abordagem EA.

Obtidos as diferenças $\bar{d}_g$ para cada um dos genes, essas são ordenadas e apresentados em gráfico que possibilitará a identificação do ponto em que as diferenciação não sejam diferentes de zero. Nesse trabalho optamos por um método não estatístico para determinar o ponto de corte separando genes diferencialmente expresso dos não diferencialmente expressos. A observação do joelho da curva como ponto de corte para seleção dos genes de interesse foi utilizada. Implementou-se um algoritmo no R para determinar a segunda derivada numérica e, a partir dessa, o ponto de corte ou joelho da curva.

4.2.3 Modelagem

Foi utilizada uma técnica de modelagem estatística para a classificação da hanseníase: Regressão Logística (RL). Para o desenvolvimento dos modelos, foi considerado como variável resposta a presença ou não presença da forma reacional eritema nodoso hansênico (ENH). Os indivíduos que apresentaram ENH foram codificados com valor 1, enquanto aqueles que apresentavam a forma LL foram codificados com valor 0. Devido a que o número de amostras contidas no conjunto de dados original tem um tamanho reduzido, que não permitiria a estimativa do modelo, visto termos $n \ll p$, sendo n o número de amostras e p número de variáveis/genes, foram realizadas simulações numéricas com base no conjunto de dados original. Os modelos foram aplicados a um conjunto de dados simulados e no conjunto de dados originais, onde a capacidade de distinção das classificações fora observada e medidas. O conjunto simulado contém n=120 observações para cada um dos genes selecionados. As réplicas para cada gene i foram produzidas como observações independentes de uma distribuição normal $N\left(\mu_i, \sigma_i^2\right), \;\; i = 1, \ldots, N$, como mostrado na equação 4.3:

$$M_{ij} \,|\, \mu_i, \sigma_i \;\sim\; N\left(\mu_i, \sigma_i^2\right) \tag{4.3}$$

sendo considerados M_{ij}como variáveis aleatórias de uma distribuição normal com média μ_i e variância σ_i^2. Estes valores são encontrados empiricamente como sendo aproximadamente o dado real, mas supondo que i e j são independentes. Essa simulação foi utilizada para aumentar o número de amostras e ser possível fazer o modelo preditivo.

Regressão Logística

Para a modelagem estatística foi utilizada a regressão logística (RL). Para implementar o modelo estatístico RL usamos um conjunto de treino, o qual foi considerado como 70% das amostras do conjunto total dos dados simulados. Para avaliar o modelo estatístico RL usamos o mesmo em um conjunto de teste, o qual foi considerado como 30% das amostras do conjunto total dos dados simulados. Além deste, utilizamos um novo conjunto de dados reais para validar o modelo preditivo.

O poder preditivo dos modelos foi avaliado segundo os parâmetros de acurácia, sensibilidade, especificidade, área sob a curva ROC e índice *Kappa*. Para ajudar a decidir se o resultado de um teste é positivo ou negativo, um ponto de corte deve ser escolhido. Em geral, o ponto de corte utilizado nos estudos é o de 0,5, em que valores acima são considerados positivos e valores abaixo são classificados como negativos. Neste presente trabalho, além da utilização do ponto de corte de 0,5, para cada grupo de modelos foi escolhido um melhor ponto de corte. Este foi definido como a melhor combinação entre os valores de sensibilidade e especificidade, correspondendo à maior soma entre os pares.

4.2.4 Anotações funcionais: Ontologia dos genes (GO) e análise de enriquecimento

Todos os genes selecionados pelo modelo foram submetidos à analise funcional, usando o banco de dados DAVID v 6.8 (*Database for Annotation, Visualization and Integrated Discovery* - `http://david.abcc.ncifcrf.gov`) [114]. Esse método foi utilizado para identificar processos biológicos e para um melhor entendimento das relações biológicas representadas por genes diferencialmente expressos, integrando dados de diversos bancos para a realização de um enriquecimento funcional [115]. Além disso, foram selecionadas somente categorias dos bancos internos de *Gene Ontology* (GO). Uma categoria foi considerada significativa caso tivesse no mínimo três genes e um valor $\leq 0,05$ com correção de FDR.

Capítulo 5

Resultados

5.1 Análise dos dados de microarranjo

Os dados de microarranjo obtidos foram analisados considerando as etapas tal como descrito na Seção 4.2.1. Na continuação são apresentados os resultados da normalização.

A normalização permite realizara correção dos erros gerados a partir dos seguintes fatores: processo de fabricação da matriz de microarranjo, eficiência da incorporação do material genético marcado, erros sistemáticos no processo de hibridação / lavagem do material, e os defeitos na matriz através de diversos tipos de contaminação. Portanto, todos os dados de microarranjo foram normalizados para ajustar para estes efeitos e permitir uma comparação dos níveis de expressão obtidos a partir de diferentes matrizes. Foi utilizada a abordagem normalização quantil, que é um método simples e rápido. Ao aplicar esta normalização, todas as lâminas obtiveram uma distribuição de intensidades comum.

A Figura 5.1 apresenta os gráficos *boxplot* dos arranjos pré-normalização para os dados do conjunto completo de microarranjo. As amostras de ENH e LL com maior mediana foram as de ENH3, ENH4, ENH5, ENH6, LL5, LL6 e LL7. Os menores valores foram para as amostras de ENH1, ENH2, LL1, LL2, LL3 e LL4.

A normalização foi feita entre todos as amostras, pelo método quantile para o ajuste das distribuições de probabilidade. A Figura 5.2 apresenta os *boxplot* dos dados normalizados e mostra que os microarranjo de todas as amostras de todos os indivíduos , possuem posições e comprimentos similares. Observando a variabilidade das amostras, pode se notar que as de ENH tendem a um valor um pouco maior do que as de LL.

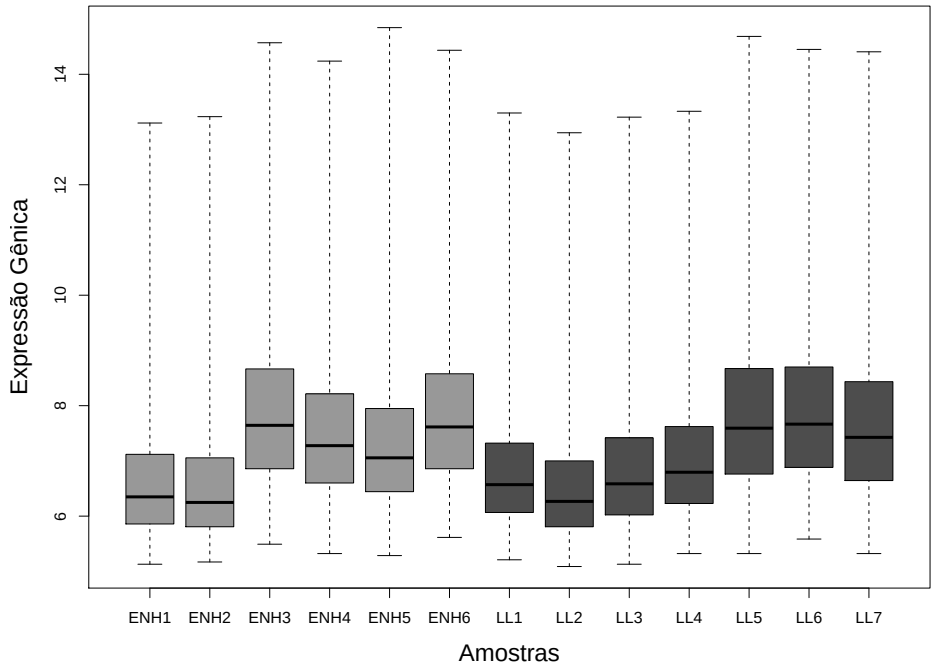


Figura 5.1: *Boxplot* da pré-normalização dos arranjos das amostras.

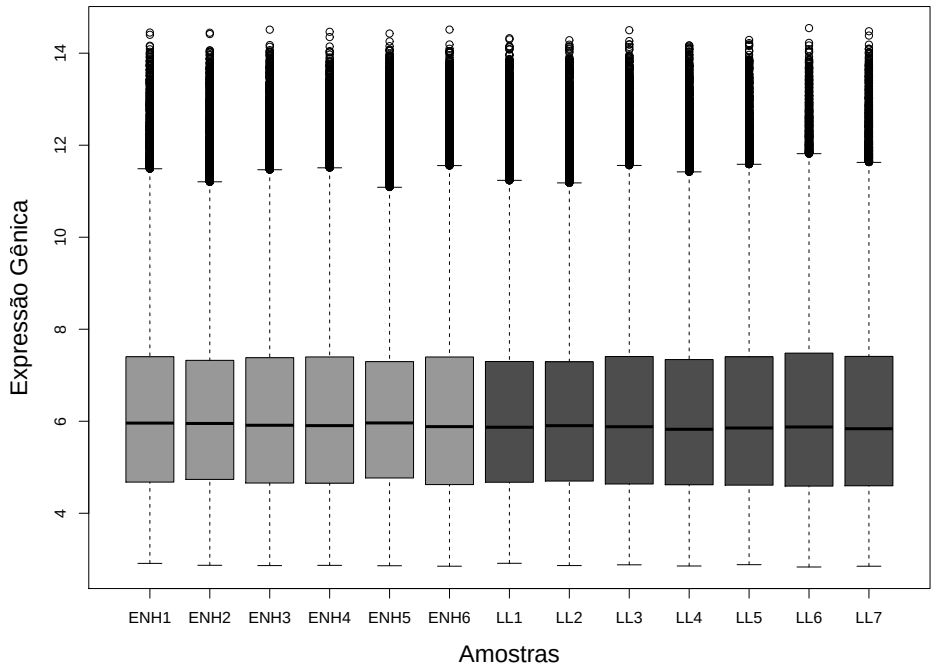


Figura 5.2: Gráfico *boxplot* de todos as amostras de microarranjo utilizados neste trabalho, após a normalização pela metodologia quantile

Após a normalização, foi obtido o histograma para mostrar o padrão dos sinais de intensidade da expressão gênica e identificar a existência de possíveis valores extremos. O histograma é mostrado na Figura 5.3.

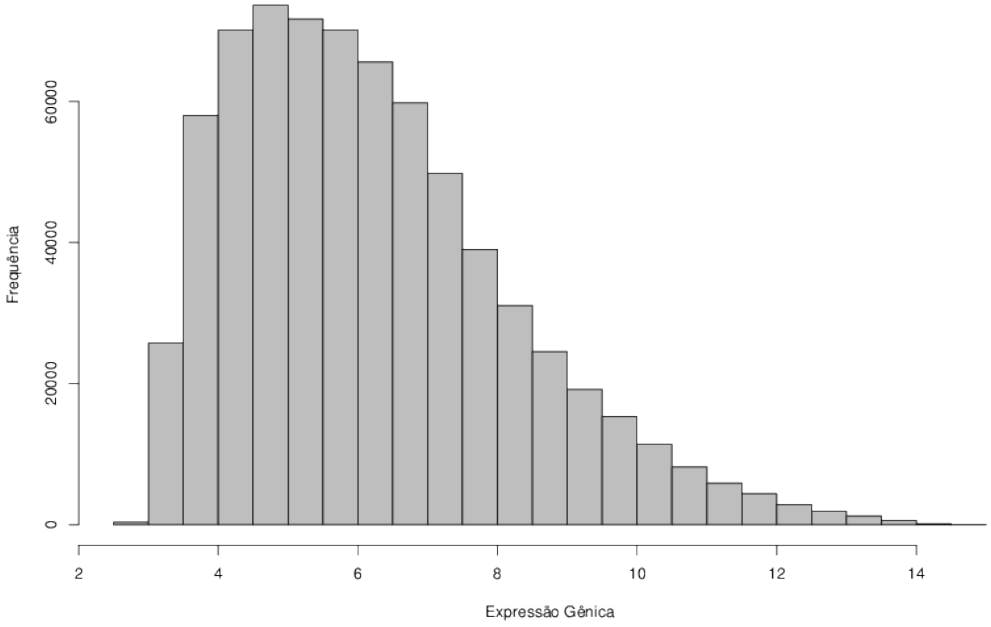


Figura 5.3: Histograma de intensidade do sinal de expressão após a normalização

5.2 Identificação de genes diferencialmente expressos

A primeira técnica utilizada para realizar a identificação dos genes diferencialmente expressos foi o teste t com *fold-change*. Na tabela 5.1 é apresentada uma comparação do número de genes rejeitados para as hipóteses de nulidade testadas. Para isto, foi considerado o nível de significância =5%, com e sem aplicação de FDR. Sem o ajuste FDR, 12.926 hipóteses nulas foram rejeitadas e com o ajuste FDR esse número foi reduzido a 1.501 genes. Através dessa comparação, foram selecionados como diferencialmente expressos 2,74% dos genes (1.501), os quais foram testados considerando o ajuste FDR e a significância =5%.

Tabela 5.1: Número genes rejeitados para as hipóteses de nulidade testadas, considerando 5% de significância, com e sem aplicação de *false discovery rate*.

	5%
Sem FDR	12.926
Com FDR	1.501

Para um melhor refinamento desses genes, o método *fold-change* foi aplicado. Na Figura 5.4 vemos o diagrama de dispersão do *fold-change* versus o logaritmo do valor de p. O diagrama corresponde ao tipo de gráfico volcano de análise dos dados. O eixo horizontal mostra os valores do *fold-change* e o eixo vertical representa a média logarítmica do valor de p. No gráfico é apresentada a dispersão dos genes analisados. Os pontos destacados acima e sobre a linha horizontal representam os genes com valores-p menores ou iguais ao ponto de corte calculado pelo método FDR. Isto é, genes com expressão diferencial a um nível de significância =5% com FDR. As duas linhas verticais delimitam os valores de *fold-change* utilizados nesta análise, que foi atribuído o valor de $\leqq -2$ ou $\geqq 2$. Os pontos à esquerda da primeira linha vertical representam genes que estão se expressando duas vezes mais nas amostras ENH, em relação as amostras LL; e os pontos à direita da segunda linha vertical representam genes que estão se expressando duas vezes mais nas amostras LL, em relação à amostra ENH.

Sendo assim, através do método do teste t e *fold-change* foram selecionados 62 genes, os quais correspondem aos pontos destacados mostrados na Figura 5.4. Dos genes encontrados, a Tabela 5.2 apresenta uma lista dos 10 primeiros genes com menores valores de p (mais diferencialmente expressos) e dos 10 primeiros genes com maiores valores de *fold-change*. A tabela com a listagem completa de genes pode ser encontrada no Apêndice A.

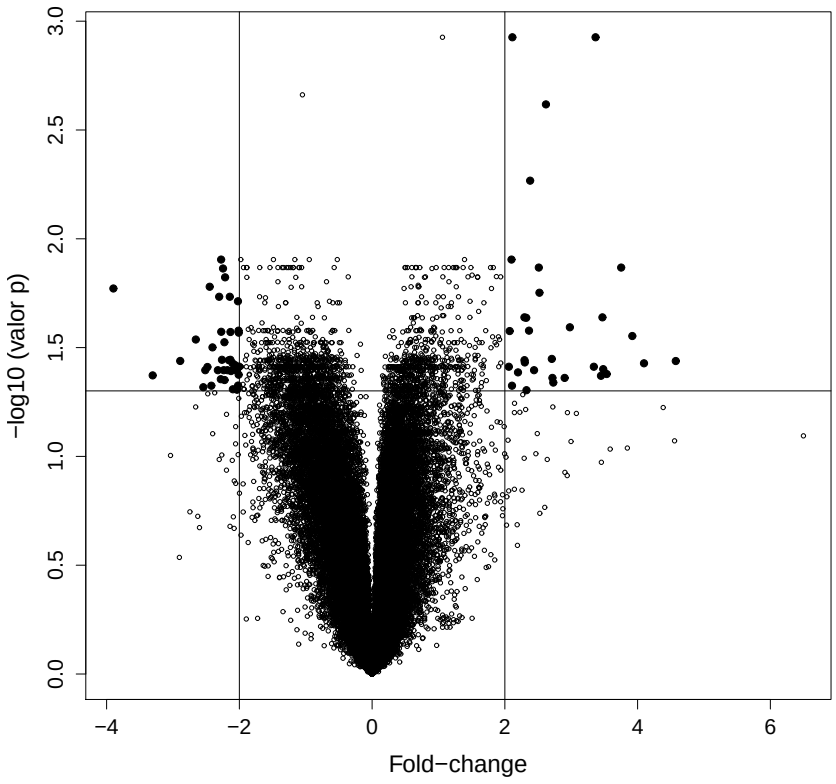


Figura 5.4: Diagrama de dispersão do logaritmo do valor de p versus o valor do *fold-change*, com os valores obtidos pela primeira técnica. Os pontos em evidência representam os genes que expressaram um valor de *fold-change* $\leqq -2$ ou $\geqq 2$ e com valor de p $\leq 0,05$.

Tabela 5.2: Lista dos 10 primeiros genes com menores valores de p e dos 10 primeiros genes com maiores *fold-change*.

Ordenado pelo valor de p			Ordenado pelo *fold-change*		
Gene	*Fold-change*	Valor dep	Gene	*Fold-change*	Valor dep
TNFAIP6	3,37	0,0012	S100A7	4,58	0,0365
ANGPTL4	2,11	0,0012	KRT6A	4,10	0,0373
PXDN	2,61	0,0024	S100A12	3,92	0,0280
NNMT	2,38	0,0054	TMEM255A	-3,90	0,0169
ALDH5A1	2,10	0,0125	AKR1B10	3,75	0,0136
SLC39A14	-2,27	0,0125	PTX3	3,53	0,0418
AKR1B10	3,75	0,0136	S100A8	3,48	0,0397
THBS1	2,51	0,0136	LTF	3,45	0,0426
PLCL1	-2,24	0,0137	TNFAIP6	3,37	0,0012
SNCA	-2,21	0,0150	PI15	3,34	0,0387

A segunda técnica utilizada foi a modelagem linear para dados de microarranjo (limma). Da mesma forma que foi utilizado na técnica anterior, foram utilizados o valor de p e os valores de *fold-change* para selecionar os genes diferencialmente expressos. Os resultados obtidos com essa técnica são apresentados em um gráfico do tipo volcano na Figura 5.5. No gráfico é apresentada a dispersão dos genes analisados. Os pontos em destaque estão enfatizando os genes que expressaram um valor de *fold-change* $\leqq -2$ ou $\geqq 2$ e com valor de $p \leq 0,05$.

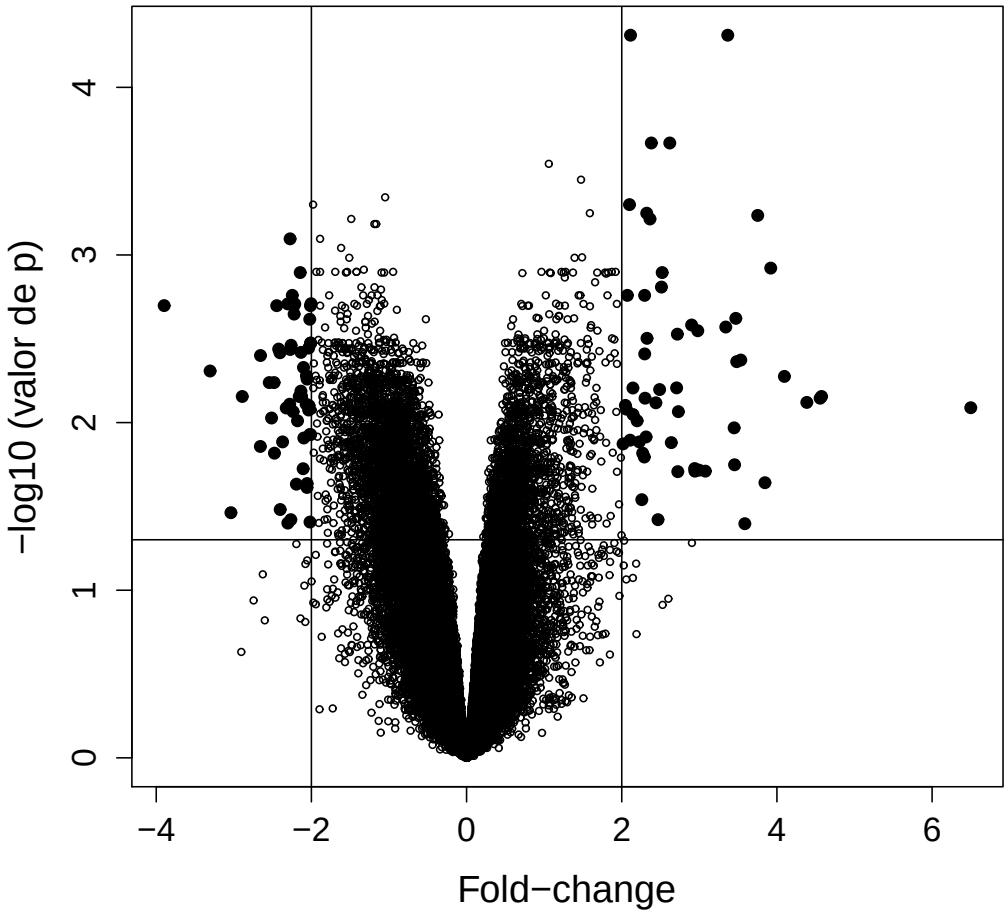


Figura 5.5: Diagrama de dispersão do logaritmo do valor de p versus o valor do *fold-change*, com os valores obtidos pela técnica limma. Os pontos em evidência representam os genes que expressaram um valor de *fold-change* $\leqq -2$ ou $\geqq 2$ e com valor de $p \leq 0,05$.

Com a técnica limma, utilizando como ponto de corte o valor de $p \leq 0,05$, e o *fold-change,* com pontos de corte os valor $\leqq -2$ ou $\geqq 2$, foram selecionados 92 genes diferencialmente expressos. A lista dos 10 primeiros genes com menores valores de p (mais diferencialmente expressos) são apresentados na Tabela 5.3. Podem ser

observados os genes com seus respectivos valores de p e $|logFC| > 2$. A tabela com a listagem completa de genes pode ser encontrada no Apêndice B.

Tabela 5.3: Lista dos 10 primeiros genes com menores valores de p e $|logFC| > 2$.

Ordenado por valor de p com o modelo linear		
Gene	Valor-p	LogFC
ANGPTL4	0,00005	2,11
TNFAIP6	0,00005	3,37
NNMT	0,00021	2,38
PXDN	0,00021	2,62
SLC39A14	0,00050	2,10
KCNJ15	0,00056	2,32
AKR1B10	0,00058	3,75
PDPN	0,00061	2,36
ALDH5A1	0,00080	-2,27
S100A12	0,00120	3,92

A terceira técnica utilizada para a identificação dos genes diferencialmente expressos foi a MD-ACP.

A Figura 5.6 apresenta as cinco primeiras CPs, obtidas através da abordagem EG. Cada componente está representada em um gráfico, no qual o eixo x representa o número da amostra e o eixo y o valor da carga. Subsequentemente, na Figura 5.7 são apresentadas as cinco primeiras CPs, obtidas através da abordagem EA.

A componente principal com a maior distância dos valores dos Grupos A e B, no caso da abordagem EG, estavam na CP2, ou seja, foi a única que apresentou a separação entre os grupos ENH e LL, tal como mostrada na Figura 5.8. Essa CP foi selecionada para se obter a lista ordenada de genes que contribuíram para a diferenciação dos grupos. Da mesma forma, para a abordagem EA, o gráfico da CP que mostra a separação entre os grupos ENH e LL corresponde a CP1, tal como se mostra na Figura 5.9.

Dentre as componentes selecionadas, pelos dois métodos, optamos para seguir as análises com a CP1, da abordagem EA. Essa escolha foi feita através do valor da proporção da variância que foi de 31% contra 1,3% da CP2 na abordagem EG. Para identificar os genes que diferenciam as amostras ENL e LL, com o método MD-ACP usamos as contribuições dos genes individualmente de ambos os grupos de amostras. A contribuição diferencial é o resultado da subtração simples da média dos valores de contribuição para cada gene, em cada grupo de amostras (ENL - LL).

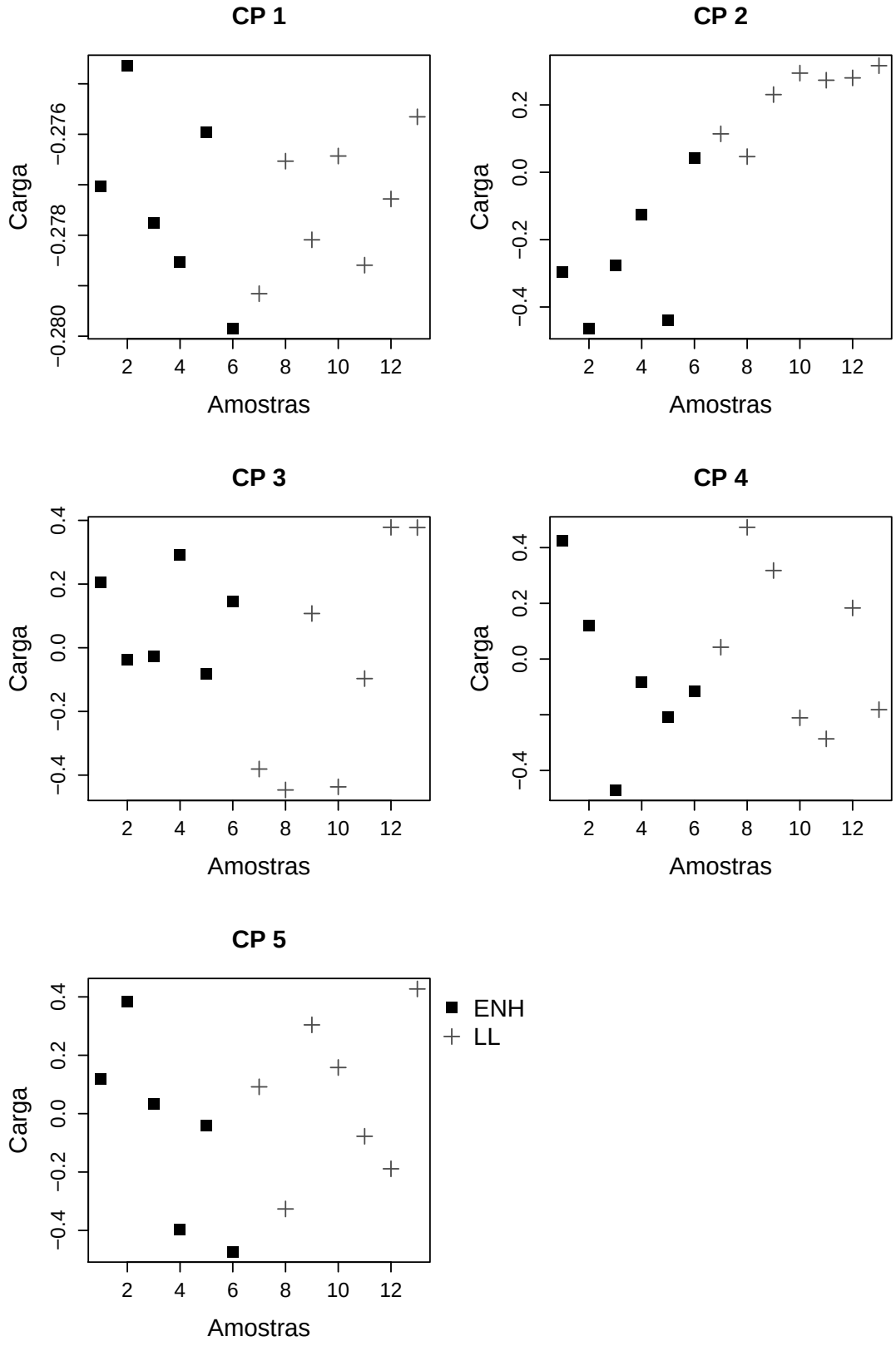


Figura 5.6: Gráfico das cinco primeiras componentes principais (CPs) obtidas com a abordagem *Eigengene*. Os quadrados representam as amostras ENH e as cruzes representam as amostras LL.

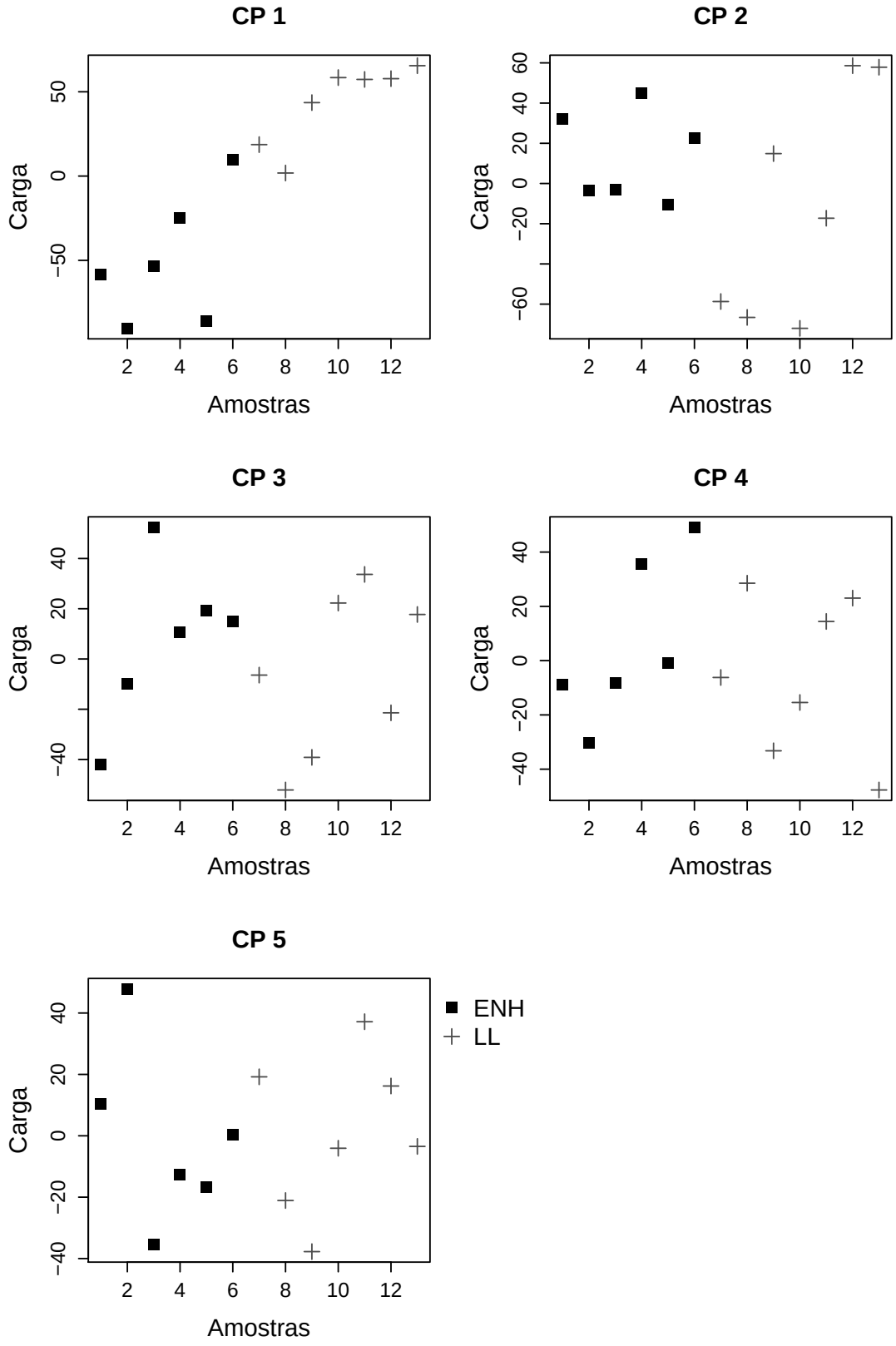


Figura 5.7: Gráfico das cinco primeiras componentes principais obtidas com a abordagem *Eigenassay*. Os quadrados representam as amostras ENH e as cruzes representam as amostras LL.

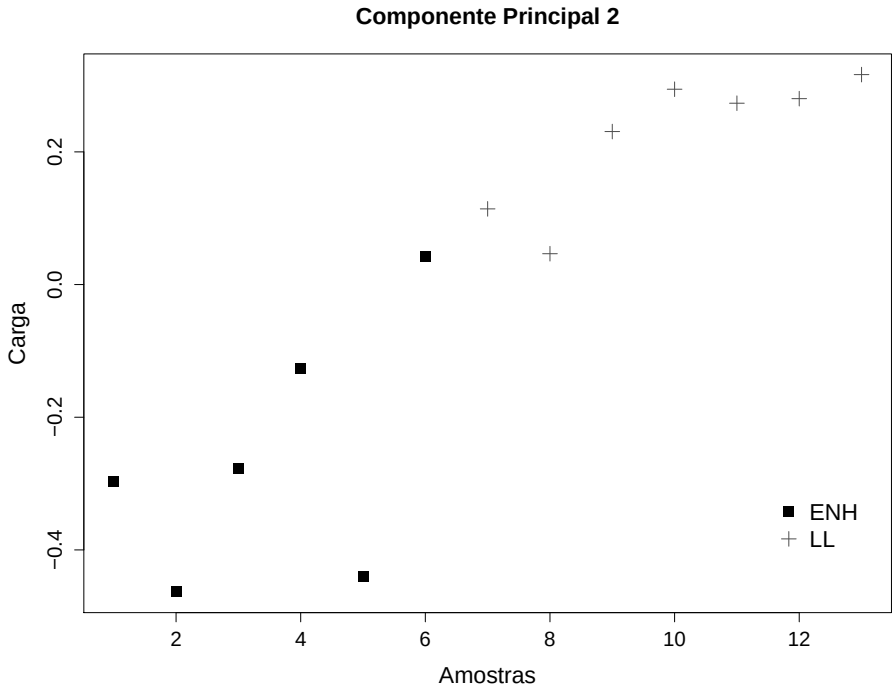


Figura 5.8: Gráfico da componente principal 2 selecionada na abordagem *Eigengene*. O grupo de amostras LL se distingue do grupo de amostras ENH.

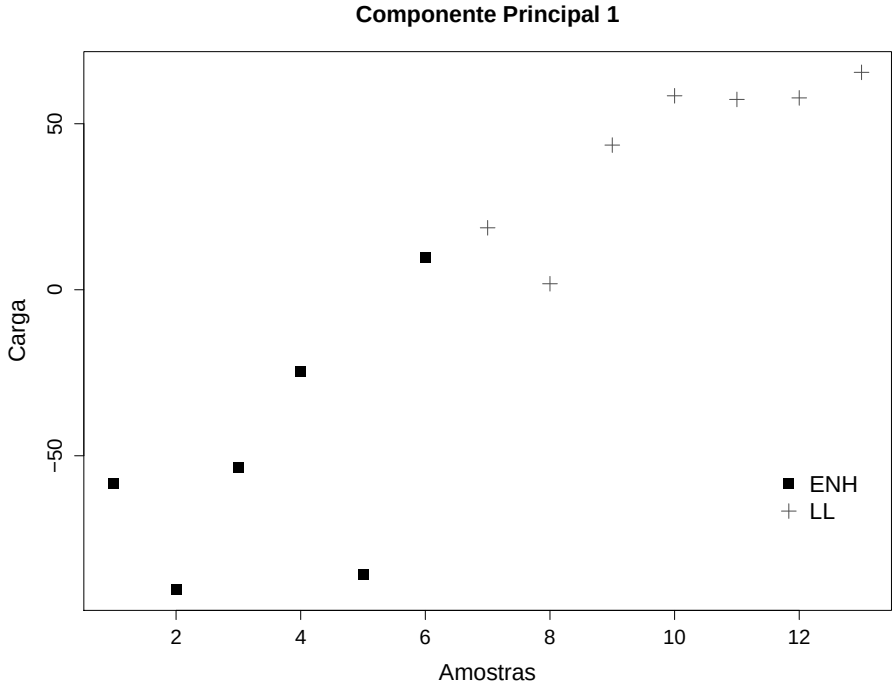


Figura 5.9: Gráfico da componente principal 1 selecionada na abordagem *Eigenassay*. O grupo de amostras LL se distingue do grupo de amostras ENH.

A Figura 5.10 mostra o gráfico de diferenças dos genes pelas abordagens EA. Nessa figura, o eixo x representa o número de genes e o eixo vertical o valor da diferença.

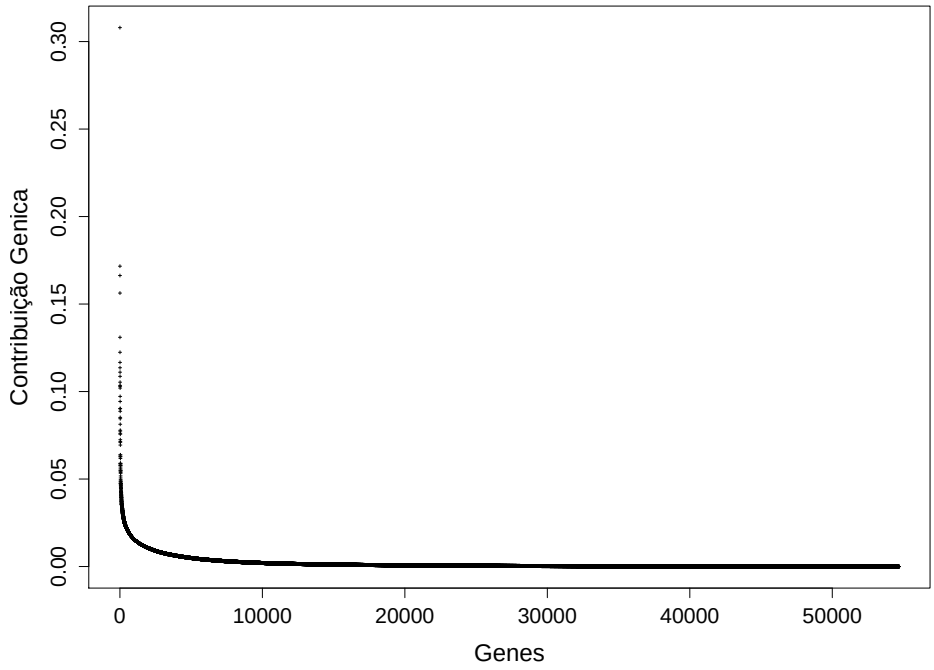


Figura 5.10: Gráficos de diferença mostrando o ranking de contribuição de genes do método *Eigenassay*.

Os genes de interesse são aqueles com os valores de diferença mais altos já que estes correspondem a um alto grau de diferença de carga de expressão entre os dois grupos de amostras. Para selecionar os genes utilizamos o método de ponto do joelho (*knee point*). Os valores da contribuição foram suavizados, por terem uma variabilidade alta, através da função não paramétrica robusta *loess*. A função *loess* permite ainda a interferência e ratificação do parâmetro de suavização (f) para a obtenção de uma suavização ideal para cada coordenada. Foram selecionados os genes que foram indicados pelo ponto do joelho por meio da segunda derivada. Na Figura 5.11 é apresentado o gráfico com os pontos do valor da diferença de contribuição para cada gene com a curva suavizada através da função *loess* e o ponto de corte indicado pelo ponto de joelho.

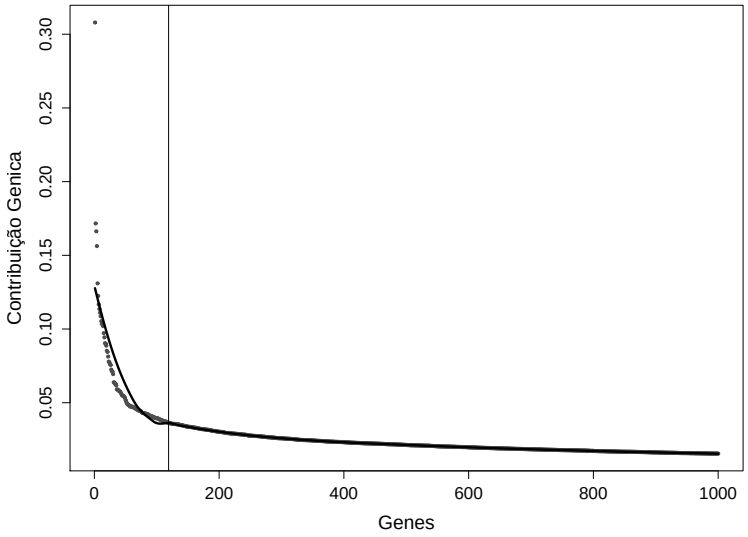


Figura 5.11: Gráfico demonstrando a curva se suavização dos valores da diferença de contribuição da Componente Principal 1 do joelho através da função "loess" com f=0,22 e a linha vertical demonstra o ponto de corte dos genes selecionados.

Através da análise com o método MD-ACP foram possíveis selecionar 93 genes únicos. As listas dos 10 primeiros genes que obtiveram maior diferença de contribuição, são mostradas na Tabela 5.4. Pode-se observar o gene identificado assim como o valor da diferença respectivo para as abordagens por EA e EG. A tabela com a lista completo de genes pode ser encontrada no Apêndice C.

Tabela 5.4: Lista dos 10 primeiros genes que obtiveram maior diferença de contribuição entre ENH e LL, na abordagen *Eigenassay*.

Lista de EA	
Gene	Valor diferença
MMP1	0,308
SERPINB4	0,172
S100A7	0,166
MMP3	0,156
S100A12	0,131
TMEM255A	0,122
S100A7A	0,117
KRT6A	0,114
AKR1B10	0,111
TDO2	0,109

5.3 Comparação dos resultados

Os conjuntos de genes diferencialmente expressos selecionados com cada um dos três métodos apresentados na seção 5.2 foram comparados. Para isto utilizou-se o diagrama de venn, com o objetivo de facilitar a observação das relações de união e intersecção entre conjuntos de genes resultantes. O diagrama é apresentado na Figura 5.12. Pode-se verificar que 53 genes foram encontrados pelas três técnicas.

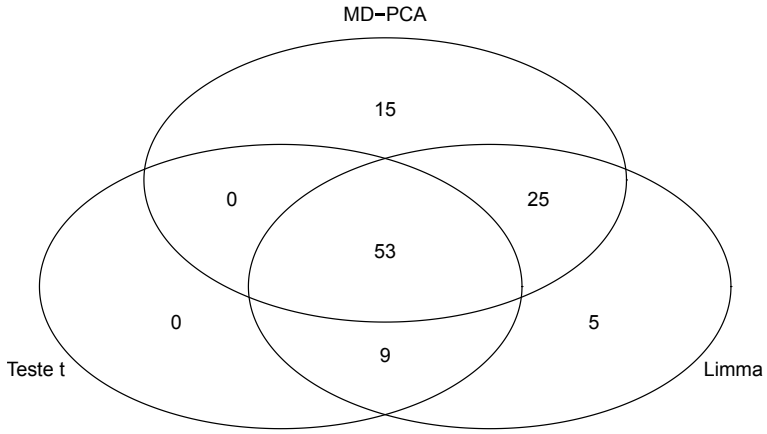


Figura 5.12: Diagrama de Venn contendo os números de genes diferencialmente expressos e compartilhados entre as técnicas de análises utilizadas. Nas intersecções são ressaltados os números de genes específicos em cada tipo de técnica.

Os genes em comum encontrados na Figura 5.12 estão descritos na Tabela 5.5.

A representação dos 53 genes diferencialmente expressos selecionados está apresentada na Figura 5.13 por meio de um dendograma. Este gráfico mostra o agrupamento hierárquico não supervisionado das amostras de indivíduos ENH e LL, segundo o perfil de expressão gênica dos 56 genes analisados. As linhas do dendograma mostram cada um dos genes encontrados. As colunas representam as amostras de ENH e LL. As intensidades de cada célula em cada amostra foram normalizadas pelo *z-score*. O dendograma de barras verticais mostra o agrupamento inter-amostral e o dendograma de barras horizontais mostra o agrupamento intra-amostral dos genes.

A Figura 5.13 mostra o heatmap típico resultante da análise de partição, que proporciona uma visão global dos dados de todos os indivíduos dos dois tipos da doença.

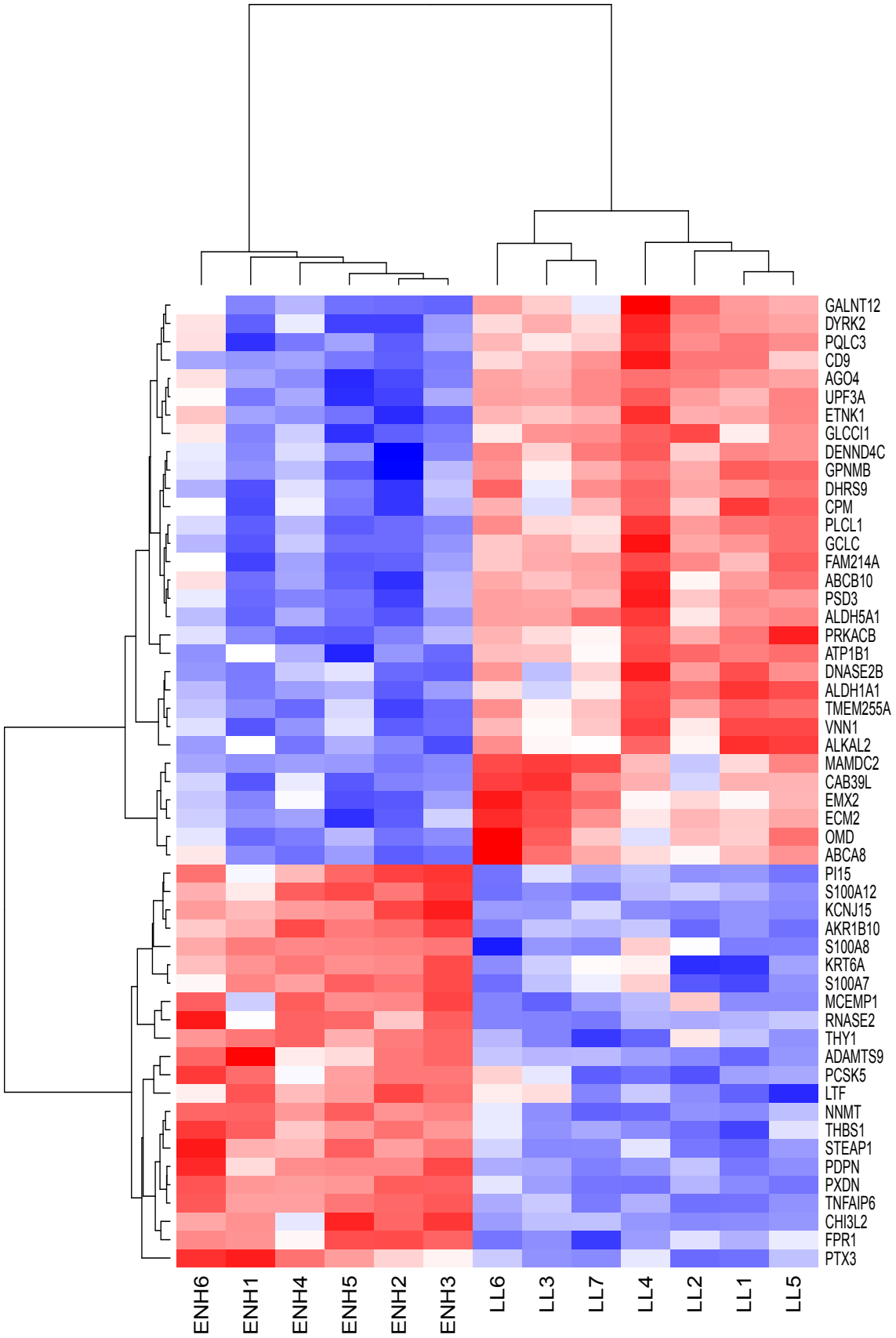


Figura 5.13: Heatmap representando o agrupamento das amostras de indivíduos hansênicos, segundo o perfil de expressão gênica dos 53 genes diferencialmente expressos selecionados.

Tabela 5.5: Lista dos 53 genes diferencialmente expressos encontrados em comum entre os três métodos.

Símbolo	Nome
ABCA8	ATP binding cassette subfamily A member 8
ABCB10	ATP binding cassette subfamily B member 10
ADAMTS9	ADAM metallopeptidase with thrombospondin type 1 motif 9
AGO4	argonaute 4, RISC catalytic component
AKR1B10	aldo-keto reductase family 1 member B10
ALDH1A1	aldehyde dehydrogenase 1 family member A1
ALDH5A1	aldehyde dehydrogenase 5 family member A1
ATP1B1	ATPase Na+/K+ transporting subunit beta 1
CAB39L	calcium binding protein 39 like
CD9	CD9 molecule
CHI3L2	chitinase 3 like 2
CPM	carboxypeptidase M
DENND4C	DENN domain containing 4C
DHRS9	dehydrogenase/reductase 9
DNASE2B	deoxyribonuclease 2 beta
DYRK2	dual specificity tyrosine phosphorylation regulated kinase 2
ECM2	extracellular matrix protein 2
EMX2	empty spiracles homeobox 2
ETNK1	ethanolamine kinase 1
FAM150B	family with sequence similarity 150 member B
FAM214A	family with sequence similarity 214 member A
FPR1	formyl peptide receptor
GALNT12	polypeptide N-acetylgalactosaminyltransferase 12
GCLC	glutamate-cysteine ligase catalytic subunit
GLCCI1	glucocorticoid induced 1
GPNMB	glycoprotein nmb
KCNJ15	potassium voltage-gated channel subfamily J member 15
KRT6A	keratin 6A
LTF	lactotransferrin
MAMDC2	MAM domain containing 2
MCEMP1	mast cell expressed membrane protein 1
NNMT	nicotinamide N-methyltransferase
OMD	osteomodulin
PCSK5	proprotein convertase subtilisin/kexin type 5

Tabela 5.5: Lista dos 53 genes diferencialmente expressos encontrados em comum entre os três métodos. (continuação)

Símbolo	Nome
PDPN	podoplanin
PI15	peptidase inhibitor 15
PLCL1	phospholipase C like
PQLC3	PQ loop repeat containing 3
PRKACB	protein kinase cAMP-activated catalytic subunit beta
PSD3	pleckstrin and Sec7 domain containing 3
PTX3	pentraxin 3
PXDN	peroxidasin
RNASE2	ribonuclease A family member 2
S100A12	S100 calcium binding protein A12
S100A7	S100 calcium binding protein A7
S100A8	S100 calcium binding protein A8
STEAP1	STEAP family member 1
THBS1	thrombospondin 1
THY1	Thy-1 cell surface antigen
TMEM255A	transmembrane protein 255A
TNFAIP6	TNF alpha induced protein 6
UPF3A	UPF3 regulator of nonsense transcripts homolog A (yeast)
VNN1	vanin 1

Finalmente, o conjunto de genes foi submetido à análise de enriquecimento funcional. Na Tabela 5.6, pode-se observar dados relevantes de cada gene como as funções gênicas e características da sua ontologia (GO). A anotação funcional da lista de genes selecionada foi realizada através da plataforma DAVID. Esta foi utilizado para identificar categorias de ontologia gênica e vias significativas com um valor de $p \leq 0,05$.

Tabela 5.6: Tabela ilustrando as categorias significativas por agrupamentos gênicos identificados. Foram dispostas categorias significativas do *Gene Ontology*, identificados pela análise de enriquecimento funcional DAVID.

Ontologia Gênica	Genes
GO:0005178~Ligação integrina	CD9, THY1, ECM2, GPNMB, THBS1
GO:0007155~ Adesão celular	ATP1B1, CD9, TNFAIP6, THY1, GPNMB, OMD, THBS1
GO:0009897~Lado externo da membrana plasmática	CD9, THY1, PDPN, THBS1
GO:0050786~Ligação do receptor RAGE	S100A12, S100A7, S100A8, FPR1
GO:0006954~Resposta inflamatória	S100A12, S100A8, TNFAIP6, FPR1, PTX3, THBS1, VNN1
GO:0045087~Resposta imune inata	S100A12, S100A7, S100A8, PTX3, VNN1

5.4 Modelagem

Para fazer a classificação da reação hansênica estudada neste trabalho (ENH e LL) é necessário um conjunto grande de amostras. Porém, como os experimentos de microarranjo são caros, uma alternativa consiste em realizar uma simulação estatística a fim de aumentar o número de amostras.

No presente trabalho foi realizado um estudo de simulação com N=120 réplicas de amostras. Com esta simulação foi possível dividir os dados em dois conjuntos: um para treino (70%) e outro para teste (30%), tal como descrito na Seção 4.2.3. Através da simulação será possível melhor descrever e avaliar os modelos aqui propostos. Com os dados simulados foi implementado o modelo de regressão logística (RL) e com o conjunto de amostras real foi feita uma validação do modelo final. Os resultados obtidos com esses modelos são apresentados a seguir.

5.4.1 Regressão logística

A partir do conjunto de treino, foi possível gerar um modelo linear generalizado (neste caso um modelo de regressão logística). O modelo completo gerado, ou seja, aquele com todos os 53 genes (variáveis) selecionados, na seção 5.3, apresentou um AIC = 114. As amostras apresentaram um problema de separação completa, ou seja, resultando em que as probabilidades foram ajustadas em 0 ou 1. Esse tipo de situação induz que para os dados utilizados no treinamento gera uma acurácia de 100%.

Esse problema ocorreu, provavelmente, por causa do tipo de dados. Isso quer dizer que um ajuste perfeito é possível dentro da parametrização do modelo: uma probabilidade $P(Y = 1) = 0$ é ajustada para casos em que o Y observado = 0; e uma probabilidade $P(Y = 1) = 1$ é ajustada para casos em que o Y observado = 1.

A Figura 5.14 mostra a análise de resíduos do modelo completo para as amostras. Os eixos x e y representam as amostras e o valor dos resíduos, respectivamente. Pode-se observar que os resíduos apresentam uma aleatoriedade, com um padrão muito próximo do valor zero, além de não ultrapassarem um limite de 2 desvios padrões $(\pm 0, 2)$.

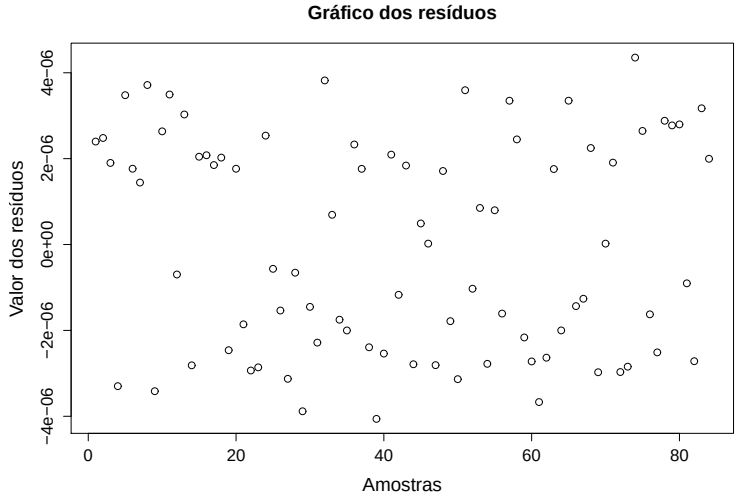


Figura 5.14: Gráfico dos resíduos do modelo para as amostras.

Para avaliar o modelo final, este foi aplicado ao conjunto de teste. Os resultados obtidos estão apresentados em uma matriz de confusão, tal como mostrado na Tabela 5.7. Nesta tabela pode ser observado o número de amostras ENH e LL preditas pelo modelo e comparadas com as amostras observadas. O ponto de corte do modelo corresponde a 0,5.

Tabela 5.7: Matriz de confusão das classificações do modelo final.

	Observado		
Preditos	ENH	LL	Total
ENH	5	0	5
LL	1	7	8
Total	6	7	13

Na Tabela 5.8, além do ponto de corte, apresentam-se os resultados de sensibilidade, especificidade, acurácia, área sob a curva ROC e índice *Kappa*. O modelo apresentou uma área sobre a curva ROC de 0,92, uma sensibilidade de 100%, uma especificidade de 83,3% e o índice *Kappa* apresentou valor igual a 0,84, indicando

concordância excelente. A curva ROC correspondente ao modelo final pode ser observada na Figura 5.15. Os eixos x e y da figura representam a taxa de falso positivo e a taxa de verdadeiro positivo, respectivamente. Essas taxas variam de 0 a 1.

Tabela 5.8: Desempenho do modelo preditivo final.

Ponto de corte	0,50
Acurácia	92%
Área sob a curva ROC	0,92
Sensibilidade	100%
Especificidade	83,3%
Índice *Kappa*	0,84

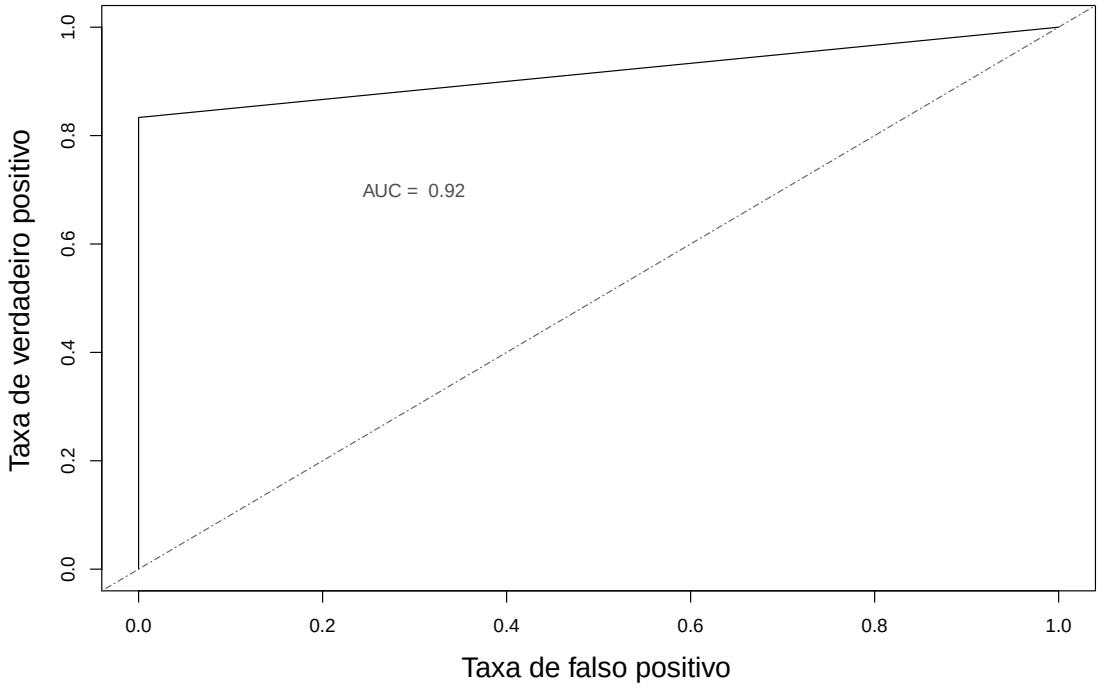


Figura 5.15: Curva ROC. A linha tracejada indica 50% da área total do gráfico.

Capítulo 6

Discussão

Nesse trabalho foi realizada a identificação dos genes diferencialmente expressos fazendo uso de três técnicas diferentes. Como esperado, cada uma dessas identificou um número de genes diferente. Usando a intersecção entre os distintos conjuntos selecionamos um número menor de genes. Observamos que desses 53 genes selecionados, 85,5% corresponde aos genes selecionados pelo teste t, e que para os outros dois métodos esses genes representaram apenas cerca de 57% dos genes selecionados por cada um dos métodos. Esse resultado sugere que o teste t foi capaz de ser mais restritivo na seleção dos genes, resultando em um número total menor que os outros dois métodos.

A técnica MD-ACP que foi mais recentemente sugerida na literatura mostrou-se adequada visto a concordância entre os métodos. É uma técnica que pode ser útil em casos em que as distribuições não sejam simétricas e que dificultem o uso das técnicas envolvendo modelos paramétricos e lineares. No problema específico desse trabalho, não houve vantagens evidentes no uso dessa técnica.

A partir dos genes selecionados foi possível analisar as funções biológicas que possam auxiliar na diferenciação entre ENH e LL. Utilizando a plataforma DAVID., obtivemos as ontologias gênicas que são utilizadas para categorizar genes e processos gênicos por meio de suas funções moleculares, envolvimento em processos biológicos e localização celular.

Alguns dos 53 genes são encontrados em 6 Ontologias Gênica que são significantes, são eles: GO:0005178 (Ligação integrina) que engloba os genes: CD9, THY1, ECM2, GPNMB, THBS1; GO:0007155 (Adesão celular) que engloba os genes: ATP1B1, CD9, TNFAIP6, THY1, GPNMB, OMD, THBS1; GO:0009897 (Lado externo da membrana plasmática) que engloba os genes: CD9, THY1, PDPN, THBS1; GO:0050
786 (Ligação do receptor RAGE) que engloba os genes: S100A12, S100A7, S100A8, FPR1; GO:0006954 (Resposta inflamatória) que engloba os genes: S100A12,

S100A8, TNFAIP6, FPR1, PTX3, THBS1, VNN1; e GO:0045087 (Resposta imune inata) que engloba os genes: S100A12, S100A7, S100A8, PTX3, VNN1. Alguns desses genes foram descritos na literatura como genes associados a hanseníase e / ou a reação ENH.

A proteína codificada pelo gene S100A12 é um membro da família S100 de proteínas. As proteínas S100 são localizadas no citoplasma e / ou núcleo de uma vasta gama de células e envolvidas na regulação de vários processos celulares tais como a progressão e diferenciação do ciclo celular. Sugere-se que esta proteína esteja envolvida em vias de transdução de sinal dependentes de cálcio específicas e o seu efeito regulador em componentes citoesqueléticos pode modular várias atividades de neutrófilos. A proteína inclui um peptídeo antimicrobiano que tem atividade antibacteriana. O gene S100A12 foi descrito como suficiente para matar diretamente o M. leprae. Há informações limitadas sobre o papel do S100A12 em doenças infecciosas, levando-nos a testar a hipótese de que S100A12 contribui para a defesa do hospedeiro contra a infecção por micobactérias em humanos [88, 116].

O gene LTF é um membro da família de genes transferrina e seu produto proteico é encontrado nos grânulos secundários de neutrófilos. Essa proteína é uma importante proteína de ligação ao ferro no leite e secreções corporais com uma atividade antimicrobiana, tornando-o um componente importante do sistema imunológico não específico. A proteína demonstra um amplo espectro de propriedades, incluindo a regulação da homeostase do ferro, a defesa do hospedeiro contra uma ampla gama de infecções microbianas, a atividade anti-inflamatória, a regulação do crescimento e diferenciação celular e a proteção contra o desenvolvimento do câncer e metástases. Os níveis elevados do gene LTF tem uma associação significativa com as reações ENH na hanseníase. Os níveis livres de lactoferrina em lágrimas são significativamente maiores em pacientes com lepra em comparação com controles normais [117, 118].

O gene THBS1 codifica uma proteína que é uma subunidade de uma proteína homotrimérica ligada a dissulfureto. Esta proteína é uma glicoproteína adesiva que intervém às interações célula a célula e célula-matriz, além de poder se ligar ao fibrinogênio, fibronectina, laminina, colágeno tipo V e integrinas alfa-V / beta-1. Esta proteína demonstrou desempenhar papéis na agregação plaquetária, angiogênese e tumorigênese. O THBS1 já foi encontrado em outro trabalho que comparava os genes diferencialmente expressos entre ENH e outra reação hansênica (RR) [118].

Alguns outros genes também já foram descritos na literatura com alguma associação à hanseníase, como o TNFAIP6 [118, 119]; RNASE2 e FPR1 [118].

Nossos resultados obtiveram um número significativo de genes confirmados pela literatura, o que permite inferir que o modelo preditivo pudesse predizer com boa acurácia as amostras LL e ENH. Essa modelagem e predição poderá a auxiliar na

possibilidade de que a reação ENH seja diagnosticada antes mesmo dos sinais físicos aparecerem.

A capacidade dos genes de prever o fenótipo clínico das amostras foi avaliada utilizando modelos de regressão logística. Este tipo de abordagem visa encontrar modelos que melhor descrevem a relação entre variáveis dependentes (ou de resposta) binárias e um conjunto de variáveis independentes. Como foi identificado que os genes selecionados eram linearmente separáveis, isto é, um conjunto desses sempre resultaria em acerto de 100% no conjunto teste simulado, não foi usado método para redução de variáveis dado que sempre teríamos apenas um gene selecionado. O modelo nesse caso, quanto usado para novos conjuntos de dados que não tivessem essa característica de serem perfeitamente linearmente separáveis, poderia resultar em alto índice de predições errôneas.

O modelo com os 53 genes foi capaz de classificar com acurácia de 92% amostras de dados diferente ao utilizado para as análises iniciais. Esse conjunto de avaliação, tem 13 amostras e o modelo errou a classificação de apenas uma amostra.

O presente estudo foi capaz de trazer novos conhecimentos sobre os genes presentes na evolução da lepra lepromatosa para a reação eritema nodoso hansênico. Foram identificados padrões de expressão que podem auxiliar na predição da reação ENH, possivelmente evitando situações com sequelas irreversíveis. Além disso, o estudo confirma alguns genes que foram recentemente demonstrados na literatura por outros estudos estarem associados a hanseníase e suas diferentes formas, e selecionou genes-candidatos, cuja expressão pode estar relacionada ao prognóstico das reações ENH. Esses genes ainda necessitam de avaliação através de técnicas da biologia molecular, como por exemplo, o uso de PCR em tempo real.

Trabalhos futuros

Baseado nos resultados obtidos neste trabalho, podem-se propor os seguintes estudos para aprofundar o conhecimento e estender a aplicação da metodologia proposta.

- Obter maiores informações dos genes obtidos, analisando suas funções e vias metabólicas em indivíduos com Eritema Nodoso Hansênico;
- Validar os modelos de classificação com outros conjuntos de dados;
- Aplicar a metodologia considerando diferentes grupos de classificação e reações hansênicas;

- Estender a aplicação da metodologia para analisar conjuntos de dados de outras doenças.

Capítulo 7

Conclusão

Neste trabalho foi proposta uma metodologia para a análise e classificação gênica através de dados de microarranjo, tendo como aplicação a identificação do episódio reacional Eritema Nodoso Hansênico.

Foram identificados diversos genes diferencialmente expressos entre a classificação LL e a reação ENH da hanseníase. Outros genes identificados abrem a possibilidade de identificar novos marcadores envolvidos no processo de evolução para o ENH que poderão ser úteis como marcadores prognósticos e alvos terapêuticos.

Foram identificados 53 genes diferencialmente expressos entre LL e ENH. Destes, 31 genes estavam sobre expressos em LL e 22 genes estavam sobre expressos em ENH.

Com os genes obtidos, foi desenvolvido um modelo estatístico preditivo para classificação de amostras LL e ENH, utilizando modelagem de regressão logística (RL). O modelo final adotado, incluía todas as 53 variáveis selecionadas visto as características dos dados originais, que apresentaram uma separação linear que impossibilitou o uso de técnicas de seleção de variáveis convencionais quando do uso da RL.

O uso de técnicas de reconhecimento de genes diferencialmente expressos numa lista de 50000 genes obtidos a partir de um experimento de microarranjo para avaliar pacientes com LL e ENH, permitiu o ranqueamento de 53 genes que foram analisados quanto as funções biológicas buscando relações com a hanseníase e reações como o ENH. O modelo de regressão logística proposto usando esses genes mostrou-se eficiente quando utilizado em um novo conjunto de dados. Embora essa amostra para avaliação tenha somente 13 casos, o resultado foi bastante significativo e encorajador para novos avanços na direção de buscar modelos eficientes para apoiar o diagnóstico clínico. Os achados relativos à lista de genes diferencialmente expressos, e resultado positivo do modelo de RL para predição, associados com pesquisas adicionais a serem desenvolvidas poderão gerar novas ferramentas para diagnóstico e tratamento de indivíduos com hanseníase.

Referências Bibliográficas

[1] WHO, W. H. O. "Global Leprosy Situation; Global leprosy update, 2014", *Weekly Epidemiological Record*, v. 90, n. 36, pp. 461–473, September 2015.

[2] HAN, X. Y., AUNG, F. M., CHOON, S. E., et al. "Analysis of the Leprosy Agents Mycobacterium leprae and Mycobacterium lepromatosis in Four Countries", *American Journal of Clinical Pathology*, v. 142, n. 4, pp. 524–532, 2014. ISSN: 0002-9173.

[3] RIDLEY, D., JOPLING, W. "Classification of leprosy according to immunity. A five-group system." *International Journal of Leprosy*, v. 34, n. 3, pp. 255–273, 1966.

[4] LEPROSY, V. I. C. "Classification. Technical Resolutions", *International Journal of Leprosy*, v. 21, pp. 504–16, 1953. Madrid Leprosy Classification.

[5] SCOLLARD, D. M., MARTELLI, C. M., STEFANI, M. M., et al. "Risk factors for leprosy reactions in three endemic countries", *The American Journal of Tropical Medicine and Hygiene*, v. 92, n. 1, pp. 108–114, 2015.

[6] RIDLEY, D. "Reactions in leprosy." *Leprosy Review*, v. 40, n. 2, pp. 77–81, 1969.

[7] KAMATH, S., VACCARO, S. A., REA, T. H., et al. "Recognizing and managing the immunologic reactions in leprosy", *Journal of the American Academy of Dermatology*, v. 71, n. 4, pp. 795–803, 2014.

[8] MONOT, M., HONORÉ, N., GARNIER, T., et al. "Comparative genomic and phylogeographic analysis of Mycobacterium leprae", *Nature Genetics*, v. 41, n. 12, pp. 1282–1289, 2009.

[9] SOUSA, A. L. M., FAVA, V. M., SAMPAIO, L. H., et al. "Genetic and immunological evidence implicates interleukin 6 as a susceptibility gene for leprosy type 2 reaction", *Journal of Infectious Diseases*, v. 205, n. 9, pp. 1417–1424, 2012.

[10] DHARMADI, Y., GONZALEZ, R. "DNA microarrays: experimental issues, data analysis, and application to bacterial systems", *Biotechnology Progress*, v. 20, n. 5, pp. 1309–1324, 2004.

[11] SCHENA, M., SHALON, D., DAVIS, R. W., et al. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", *Science*, v. 270, n. 5235, pp. 467–470, 1995.

[12] GOLUB, T. R., SLONIM, D. K., TAMAYO, P., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, v. 286, n. 5439, pp. 531–537, 1999.

[13] MITRA, R. D., SHENDURE, J., OLEJNIK, J., et al. "Fluorescent in situ sequencing on polymerase colonies", *Analytical Biochemistry*, v. 320, n. 1, pp. 55–65, 2003.

[14] WORTMAN, J. R., HAAS, B. J., HANNICK, L. I., et al. "Annotation of the Arabidopsis genome", *Plant Physiology*, v. 132, n. 2, pp. 461–468, 2003.

[15] CLOSE, T. J., WANAMAKER, S. I., CALDO, R. A., et al. "A new resource for cereal genomics: 22K barley GeneChip comes of age", *Plant Physiology*, v. 134, n. 3, pp. 960–968, 2004.

[16] PODDER, S., GHOSH, T. C. "Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human", *Molecular Biology and Evolution*, v. 27, n. 4, pp. 934–941, 2010.

[17] VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B., et al. "Serial analysis of gene expression", *Science*, v. 270, n. 5235, pp. 484, 1995.

[18] HAYWARD-LESTER, A., OEFNER, P. J., SABATINI, S., et al. "Accurate and absolute quantitative measurement of gene expression by single-tube RT-PCR and HPLC." *Genome Research*, v. 5, n. 5, pp. 494–499, 1995.

[19] MCCARTHY, D. J., CHEN, Y., SMYTH, G. K. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation", *Nucleic Acids Research*, p. gks042, 2012.

[20] FAGERBERG, L., HALLSTROM, B. M., OKSVOLD, P., et al. "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics", *Molecular & Cellular Proteomics*, v. 13, n. 2, pp. 397–406, 2014.

[21] TARCA, A. L., ROMERO, R., DRAGHICI, S. "Analysis of microarray experiments of gene expression profiling", *American Journal of Obstetrics and Gynecology*, v. 195, n. 2, pp. 373–388, 2006.

[22] NADON, R., SHOEMAKER, J. "Statistical issues with microarrays: processing and analysis", *Trends in Genetics*, v. 18, n. 5, pp. 265–271, 2002.

[23] SMYTH, G. K., SPEED, T. "Normalization of cDNA microarray data", *Methods*, v. 31, n. 4, pp. 265–273, 2003.

[24] ROLLINS, D. K., TEH, A. "An extended data mining method for identifying differentially expressed assay-specific signatures in functional genomic studies", *BioData Mining*, v. 3, n. 1, pp. 1–18, 2010. ISSN: 1756-0381.

[25] DIBOUN, I., WERNISCH, L., ORENGO, C. A., et al. "Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma", *BMC Genomics*, v. 7, n. 1, pp. 252, 2006.

[26] YU, Y., LUO, Y., ZHENG, Y., et al. "Exploring the mechanism of non-small-cell lung cancer cell lines resistant to epidermal growth factor receptor tyrosine kinase inhibitor", *Journal of Cancer Research and Therapeutics*, v. 12, n. 1, pp. 121, 2016.

[27] EFRON, B., TIBSHIRANI, R., STOREY, J. D., et al. "Empirical Bayes analysis of a microarray experiment", *Journal of the American Statistical Association*, v. 96, n. 456, pp. 1151–1160, 2001.

[28] KENDZIORSKI, C., NEWTON, M., LAN, H., et al. "On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles", *Statistics in Medicine*, v. 22, n. 24, pp. 3899–3914, 2003.

[29] HYDUKE, D. R., JARBOE, L. R., TRAN, L. M., et al. "Integrated network analysis identifies nitric oxide response networks and dihydroxyacid dehydratase as a crucial target in Escherichia coli", *Proceedings of the National Academy of Sciences*, v. 104, n. 20, pp. 8484–8489, 2007.

[30] MEI, R., DI, X., RYDER, T., et al. "Analysis of high density expression microarrays with signed-rank call algorithms", *Bioinformatics*, v. 18, n. 12, pp. 1593–1599, 2002.

[31] LIAO, J., CHIN, K.-V. "Logistic regression for disease classification using microarray data: model selection in a large p and small n case", *Bioinformatics*, v. 23, n. 15, pp. 1945–1951, 2007.

[32] LU, Y., HAN, J. "Cancer classification using gene expression data", *Information Systems*, v. 28, n. 4, pp. 243–268, 2003.

[33] LI, T., ZHANG, C., OGIHARA, M. "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression", *Bioinformatics*, v. 20, n. 15, pp. 2429–2437, 2004.

[34] SCHOLKOPFT, B., MULLERT, K.-R. "Fisher discriminant analysis with kernels", *Neural Networks for Signal Processing IX*, v. 1, n. 1, pp. 1, 1999.

[35] ROSENFELD, N., AHARONOV, R., MEIRI, E., et al. "MicroRNAs accurately identify cancer tissue origin", *Nature Biotechnology*, v. 26, n. 4, pp. 462–469, 2008.

[36] RAMASWAMY, S., TAMAYO, P., RIFKIN, R., et al. "Multiclass cancer diagnosis using tumor gene expression signatures", *Proceedings of the National Academy of Sciences*, v. 98, n. 26, pp. 15149–15154, 2001.

[37] GUYON, I., WESTON, J., BARNHILL, S., et al. "Gene selection for cancer classification using support vector machines", *Machine Learning*, v. 46, n. 1-3, pp. 389–422, 2002.

[38] ZHANG, H., BERG, A. C., MAIRE, M., et al. "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, v. 2, pp. 2126–2136. IEEE, 2006.

[39] KHAN, J., WEI, J. S., RINGNER, M., et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", *Nature Medicine*, v. 7, n. 6, pp. 673–679, 2001.

[40] MADADI, M., ZHANG, S., YEARY, K. H. K., et al. "Analyzing factors associated with women's attitudes and behaviors toward screening mammography using design-based logistic regression", *Breast Cancer Research and Treatment*, v. 144, n. 1, pp. 193–204, 2014.

[41] LEE, S. H., YU, D., BACHMAN, A. H., et al. "Application of fused lasso logistic regression to the study of corpus callosum thickness in early Alzheimer's disease", *Journal of Neuroscience Methods*, v. 221, pp. 78–84, 2014.

[42] HOSMER, D. W., HOSMER, T., LE CESSIE, S., et al. "A comparison of goodness-of-fit tests for the logistic regression model", *Statistics in medicine*, v. 16, n. 9, pp. 965–980, 1997.

[43] EL SANHARAWI, M., NAUDET, F. "Comprendre la regression logistique", *Journal francais d'ophtalmologie*, v. 36, n. 8, pp. 710–715, 2013.

[44] MONTGOMERY, D. C., RUNGER, G. C. *Applied statistics and probability for engineers*. Fifth ed. New Jersey, John Wiley & Sons, 2010.

[45] PARDOE, I. *Applied regression modeling: a business approach*. Second ed. New Jersey, John Wiley & Sons, 2012.

[46] HOSMER JR, D. W., LEMESHOW, S., STURDIVANT, R. X. *Applied logistic regression*, v. 398. Third ed. New Jersey, John Wiley & Sons, 2013.

[47] FIGUEIRA, C. V. *Modelos de regressão logística*. Tese de Mestrado, Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, 2006.

[48] WHO, W. H. O. "Global Leprosy Situation; Global leprosy update, 2015", *Weekly Epidemiological Record*, v. 91, n. 02, pp. 405–420, September 2016.

[49] WHO, W. H. O. "Progress towards the elimination of leprosy as a public health problem", *Weekly Epidemiological Record*, v. 72, n. 26, pp. 181–186, 1995.

[50] NOORDEEN, S. "Vaccination against leprosy; recent advances and practical implications." *Leprosy Review*, v. 56, n. 1, pp. 1–3, 1985.

[51] BRITTON, W. J., LOCKWOOD, D. N. "Leprosy", *The Lancet*, v. 363, n. 9416, pp. 1209–1219, 2004. ISSN: 0140-6736.

[52] REES, R., MCDOUGALL, A. "Airborne infection with Mycobacterium leprae in mice", *Journal of Medical Microbiology*, v. 10, n. 1, pp. 63–68, 1977.

[53] MATOS, H. J. D., DUPPRE, N., ALVIM, M. F. S., et al. "Epidemiologia da hanseníase em coorte de contatos intradomiciliares no Rio de Janeiro (1987-1991)", *Cadernos de Saúde Pública*, v. 15, n. 3, pp. 533–542, 1999.

[54] SALES, A. M., DE LEON, A. P., DUPPRE, N. C., et al. "Leprosy among patient contacts: a multilevel study of risk factors", *PLoS Neglected Tropical Diseases*, v. 5, n. 3, pp. e1013, 2011.

[55] SCHAIBLE, U. E., HAGENS, K., FISCHER, K., et al. "Intersection of group I CD1 molecules and mycobacteria in different intracellular compartments of dendritic cells", *The Journal of Immunology*, v. 164, n. 9, pp. 4843–4852, 2000.

[56] ALMEIDA, E. C. D., MARTINEZ, A. N., MANIERO, V. C., et al. "Detection of Mycobacterium leprae DNA by polymerase chain reaction in the

blood and nasal secretion of Brazilian household contacts", *Memórias do Instituto Oswaldo Cruz*, v. 99, n. 5, pp. 509–511, 2004.

[57] MOET, F. J., PAHAN, D., SCHURING, R. P., et al. "Physical distance, genetic relationship, age, and leprosy classification are independent risk factors for leprosy in contacts of patients with leprosy", *Journal of Infectious Diseases*, v. 193, n. 3, pp. 346–353, 2006.

[58] LAZARO, F. P., WERNECK, R. I., MACKERT, C. C., et al. "A major gene controls leprosy susceptibility in a hyperendemic isolated population from north of Brazil", *Journal of Infectious Diseases*, v. 201, n. 10, pp. 1598–1605, 2010.

[59] MORAES, M. O., CARDOSO, C. C., VANDERBORGHT, P. R., et al. "Genetics of host response in leprosy", *Leprosy Review*, v. 77, n. 3, pp. 189, 2006.

[60] NOORDEEN, S., BRAVO, L. L., SUNDARESAN, T. "Estimated number of leprosy cases in the world." *Bulletin of the World Health Organization*, v. 70, n. 1, pp. 7, 1992.

[61] RIDLEY, D. "The bacteriological interpretation of skin smears and biopsies in leprosy", *Transactions of the Royal Society of Tropical Medicine and Hygiene*, v. 49, n. 5, pp. 449–452, 1955.

[62] ASSOCIATION, I. L. *Report of the International Leprosy Association Technical.* Paris, France, LEPRA, 25-28 February 2002.

[63] BRASIL. "Guia para o controle da hanseníase", 2002.

[64] LOCKWOOD, D., REID, A. "The diagnosis of leprosy is delayed in the United Kingdom." *QJM: Monthly Journal of the Association of Physicians*, v. 94, n. 4, pp. 207–212, april 2001.

[65] MOURA, R. S. D., CALADO, K. L., OLIVEIRA, M. L. W., et al. "Sorologia da hanseníase utilizando PGL-I: revisão sistemática", *Revista da Sociedade Brasileira de Medicina Tropical*, v. 41, n. supl. 2, pp. 11–18, 2008.

[66] MARTINEZ, A. N., BRITTO, C. F., NERY, J. A., et al. "Evaluation of real-time and conventional PCR targeting complex 85 genes for detection of Mycobacterium leprae DNA in skin biopsy samples from patients diagnosed with leprosy", *Journal of Clinical Microbiology*, v. 44, n. 9, pp. 3154–3159, 2006.

[67] MARTINEZ, A. N., LAHIRI, R., PITTMAN, T. L., et al. "Molecular determination of Mycobacterium leprae viability by use of real-time PCR", *Journal of Clinical Microbiology*, v. 47, n. 7, pp. 2124–2130, 2009.

[68] MARTINEZ, A. N., RIBEIRO-ALVES, M., SARNO, E. N., et al. "Evaluation of qPCR-based assays for leprosy diagnosis directly in clinical specimens", *PLoS Neglected Tropical Diseases*, v. 5, n. 10, pp. e1354, 2011.

[69] JOPLING, W. H. "Indeterminate leprosy." *Proceedings of the Royal Society of Medicine*, v. 52, n. 5, pp. 370–371, maio 1959.

[70] JARDIM, M. R., ANTUNES, S. L., SIMONS, B., et al. "Role of PGL-I antibody detection in the diagnosis of pure neural leprosy", *Leprosy Review*, v. 76, n. 3, pp. 232–240, 2005.

[71] SCOLLARD, D., ADAMS, L., GILLIS, T., et al. "The continuing challenges of leprosy", *Clinical Microbiology Reviews*, v. 19, n. 2, pp. 338–381, 2006.

[72] NAAFS, B. "Current views on reactions in leprosy." *Indian Journal of Leprosy*, v. 72, n. 1, pp. 97, 2000.

[73] REGO, V., MACHADO, P. R., MARTINS, I., et al. "Type 1 reaction in leprosy: characteristics and association with hepatitis B and C viruses", *Revista da Sociedade Brasileira de Medicina Tropical*, v. 40, n. 5, pp. 546–549, 2007.

[74] LEGENDRE, D. P., MUZNY, C. A., SWIATLO, E. "Hansen's Disease (Leprosy): Current and Future Pharmacotherapy and Treatment of Disease-Related Immunologic Reactions", *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, v. 32, n. 1, pp. 27–37, 2012.

[75] KAHAWITA, I., LOCKWOOD, D. "Towards understanding the pathology of erythema nodosum leprosum", *Transactions of the Royal Society of Tropical Medicine and Hygiene*, v. 102, n. 4, pp. 329–337, 2008.

[76] POCATERRA, L., JAIN, S., REDDY, R., et al. "Clinical course of erythema nodosum leprosum: an 11-year cohort study in Hyderabad, India", *The American Journal of Tropical Medicine and Hygiene*, v. 74, n. 5, pp. 868–879, 2006.

[77] KUMAR, B., DOGRA, S., KAUR, I. "Epidemiological Characteristics of Leprosy Reactions: 15 Years Experience from North India1", *International Journal of Leprosy and Other Mycobacterial Diseases*, v. 72, n. 2, pp. 125, 2004.

[78] MANANDHAR, R., LEMASTER, J. W., ROCHE, P. W. "Risk factors for erythema nodosum leprosum", *International Journal of Leprosy and Other Mycobacterial Diseases*, v. 67, n. 3, pp. 270–278, 1999.

[79] SAUNDERSON, P., GEBRE, S., BYASS, P. "ENL reactions in the multibacillary cases of the AMFES cohort in central Ethiopia: incidence and risk factors." *Leprosy Review*, v. 71, n. 3, pp. 318–324, 2000.

[80] MAURUS, J. N. "Hansen's disease in pregnancy." *Obstetrics & Gynecology*, v. 52, n. 1, pp. 22–25, 1978.

[81] LEE, D. J., LI, H., OCHOA, M. T., et al. "Integrated pathways for neutrophil recruitment and inflammation in leprosy", *Journal of Infectious Diseases*, v. 201, n. 4, pp. 558–569, 2010.

[82] WHO. "Estratégia global aprimorada para reduçã adicional da carga da hanseníase (2011-2015)", *Brasil: Ministério da Saúde*, 2010.

[83] WANG, Y., TETKO, I. V., HALL, M. A., et al. "Gene selection from microarray data for cancer classification - a machine learning approach", *Computational Biology and Chemistry*, v. 29, n. 1, pp. 37–46, 2005.

[84] DUGGAN, D. J., BITTNER, M., CHEN, Y., et al. "Expression profiling using cDNA microarrays", *Nature Genetics*, v. 21, pp. 10–14, 1999.

[85] QUACKENBUSH, J. "Computational analysis of microarray data", *Nature Reviews Genetics*, v. 2, n. 6, pp. 418–427, 2001.

[86] SCHENA, M., HELLER, R. A., THERIAULT, T. P., et al. "Microarrays: biotechnology's discovery platform for functional genomics", *Trends in Biotechnology*, v. 16, n. 7, pp. 301–306, 1998.

[87] LAUSTED, C., DAHL, T., WARREN, C., et al. "POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer", *Genome Biology*, v. 5, n. 8, pp. 1, 2004.

[88] KIM, K.-J., CHO, S.-B. "Ensemble classifiers based on correlation analysis for DNA microarray classification", *Neurocomputing*, v. 70, n. 1, pp. 187–199, 2006.

[89] BALDI, P., LONG, A. D. "A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes", *Bioinformatics*, v. 17, n. 6, pp. 509–519, 2001.

[90] BENJAMINI, Y., HOCHBERG, Y. "Controlling the false discovery rate: a practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[91] SMYTH, G. K. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments", *Statistical Applications in Genetics and Molecular Biology*, v. 3, June 2004.

[92] PEARSON, K. "Principal components analysis", *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, v. 6, n. 2, pp. 559, 1901.

[93] HOTELLING, H. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology*, v. 24, n. 6, pp. 417, 1933.

[94] JOLLIFFE, I. T. *Principal component analysis*. Second ed. New York, Springer-Verlag, 2002.

[95] LANDGREBE, J., WURST, W., WELZL, G. "Permutation-validated principal components analysis of microarray data", *Genome Biology*, v. 3, n. 4, pp. 1, 2002.

[96] ROLLINS, D. K., ZHAI, D., JOE, A. L., et al. "A novel data mining method to identify assay-specific signatures in functional genomic studies", *BMC Bioinformatics*, v. 7, n. 1, pp. 1, 2006.

[97] KRZANOWSKI, W. J. *An Introdution to Statistical Modelling*. First ed. London, United Kingdom, John Wiley and Sons Ltd, 1998.

[98] ZHU, J., HASTIE, T. "Classification of gene microarrays by penalized logistic regression", *Biostatistics*, v. 5, n. 3, pp. 427–443, 2004.

[99] AKAIKE, H. "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, v. 19, n. 6, pp. 716–723, 1974.

[100] SANTOS, A. M., DE SEIXAS, J. M., DE BRAGANÇA PEREIRA, B., et al. "Usando redes neurais artificiais e regressão logística na predição da hepatite A", *Revista Brasileira de Epidemiologia*, v. 8, n. 2, pp. 117–26, 2005.

[101] KONIS, K. *Linear programming algorithms for detecting separated data in binary logistic regression models*. Tese de Doutorado, University of Oxford, Oxford, 2007.

[102] RIPLEY, B. D. *Pattern recognition and neural networks*. New York, Cambridge University Press, 2007.

[103] KOHAVI, R. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Appears in the International Joint Conference on Artificial Intelligence*, v. 14, pp. 1137–1145. Stanford, CA, 1995.

[104] ZWEIG, M. H., CAMPBELL, G. "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." *Clinical Chemistry*, v. 39, n. 4, pp. 561–577, 1993.

[105] COHEN, J. "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." *Psychological Bulletin*, v. 70, n. 4, pp. 213, 1968.

[106] LANDIS, J. R., KOCH, G. G. "The measurement of observer agreement for categorical data", *Biometrics*, pp. 159–174, 1977.

[107] EDGAR, R., DOMRACHEV, M., LASH, A. E. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", *Nucleic Acids Research*, v. 30, n. 1, pp. 207–210, 2002.

[108] INKELES, M. S., TELES, R. M., POULDAR, D., et al. "Cell-type deconvolution with immune pathways identifies gene networks of host defense and immunopathology in leprosy", *JCI Insight*, v. 1, n. 15, 2016.

[109] MONTOYA, D., CRUZ, D., TELES, R. M., et al. "Divergence of macrophage phagocytic and antimicrobial programs in leprosy", *Cell Host & Microbe*, v. 6, n. 4, pp. 343–353, 2009.

[110] IRIZARRY, R. A., HOBBS, B., COLLIN, F., et al. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data", *Biostatistics*, v. 4, n. 2, pp. 249–264, 2003.

[111] MCCALL, M. N., BOLSTAD, B. M., IRIZARRY, R. A. "Frozen robust multiarray analysis (fRMA)", *Biostatistics*, v. 11, n. 2, pp. 242–253, 2010.

[112] IRIZARRY, R. A., BOLSTAD, B. M., COLLIN, F., et al. "Summaries of Affymetrix GeneChip probe level data", *Nucleic Acids Research*, v. 31, n. 4, pp. e15–e15, 2003.

[113] BOLSTAD, B. M., IRIZARRY, R. A., ÅSTRAND, M., et al. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias", *Bioinformatics*, v. 19, n. 2, pp. 185–193, 2003.

[114] DENNIS, G., SHERMAN, B. T., HOSACK, D. A., et al. "DAVID: database for annotation, visualization, and integrated discovery", *Genome Biology*, v. 4, n. 9, pp. 1, 2003.

[115] HUANG, D. W., SHERMAN, B. T., LEMPICKI, R. A. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources", *Nature Protocols*, v. 4, n. 1, pp. 44–57, 2009.

[116] REALEGENO, S., KELLY-SCUMPIA, K. M., DANG, A. T., et al. "S100A12 is part of the Antimicrobial network against Mycobacterium leprae in Human Macrophages", *PLoS Pathogens*, v. 12, n. 6, pp. e1005705, 2016.

[117] DANIEL, E., DURIASAMY, M., EBENEZER, G. J., et al. "Elevated free tear lactoferrin levels in leprosy are associated with type 2 reactions", *Indian Journal of Ophthalmology*, v. 52, n. 1, pp. 51, 2004.

[118] DUPNIK, K. M., BAIR, T. B., MAIA, A. O., et al. "Transcriptional changes that characterize the immune reactions of leprosy", *The Journal of Infectious Diseases*, v. 211, n. 10, pp. 1658–1676, 2014.

[119] ORLOVA, M., COBAT, A., HUONG, N. T., et al. "Gene set signature of reversal reaction type I in leprosy patients", *PLoS Genetics*, v. 9, n. 7, pp. e1003624, 2013.

Apêndice A

Resultado do método de seleção utilizando teste t

Tabela A.1: Tabela dos genes obtidos através da técnica teste t e *fold-change*.

Rank	ID	Símbolo	*Fold-change*	Valor p ajustado
1	221009_s_at	ANGPTL4	2,11	0,0012
2	206025_s_at	TNFAIP6	3,37	0,0012
3	212013_at	PXDN	2,62	0,0024
4	202237_at	NNMT	2,38	0,0054
5	203608_at	ALDH5A1	-2,27	0,0125
6	212110_at	SLC39A14	2,10	0,0125
7	206561_s_at	AKR1B10	3,75	0,0136
8	201110_s_at	THBS1	2,51	0,0136
9	205934_at	PLCL1	-2,25	0,0137
10	236081_at	SNCA	-2,21	0,0150
11	202922_at	GCLC	-2,45	0,0166
12	219895_at	TMEM255A	-3,90	0,0169
13	205542_at	STEAP1	2,52	0,0177
14	201005_at	CD9	-2,14	0,0185
15	218613_at	PSD3	-2,30	0,0185
16	224009_x_at	DHRS9	-2,02	0,0194
17	208850_s_at	THY1	2,29	0,0230
18	210119_at	KCNJ15	2,32	0,0231
19	205119_s_at	FPR1	2,98	0,0255
20	225327_at	FAM214A	-2,01	0,0264
21	221898_at	PDPN	2,36	0,0264
22	205656_at	PCDH17	2,07	0,0265
23	206101_at	ECM2	-2,27	0,0267
24	235780_at	PRKACB	-2,13	0,0268
25	205863_at	S100A12	3,92	0,0280
26	201243_s_at	ATP1B1	-2,40	0,0316
27	214370_at	S100A8	2,71	0,0356
28	1554018_at	GPNMB	-2,26	0,0360
29	205844_at	VNN1	-2,14	0,0360
30	205798_at	IL7R	2,30	0,0361
31	225915_at	CAB39L	-2,16	0,0362
32	218885_s_at	GALNT12	-2,89	0,0364
33	205916_at	S100A7	4,58	0,0365
34	205559_s_at	PCSK5	2,30	0,0370
35	209125_at	KRT6A	4,10	0,0373
36	205907_s_at	OMD	-2,10	0,0373
37	204719_at	ABCA8	-2,07	0,0380
38	206958_s_at	UPF3A	-2,00	0,0386
39	229947_at	PI15	3,34	0,0387
40	225579_at	PQLC3	-2,06	0,0387
41	219869_s_at	SLC39A8	2,06	0,0387
42	227525_at	GLCCI1	-2,48	0,0388

Tabela A.1: Tabela dos genes obtidos através da técnica teste t e *fold-change*. (continuação)

Rank	ID	Símbolo	*Fold-change*	Valor p ajustado
43	226867_at	DENND4C	-2,04	0,0400
44	238018_at	ALKAL2	-2,51	0,0402
45	220380_at	DNASE2B	-2,24	0,0402
46	221950_at	EMX2	-2,18	0,0402
47	223320_s_at	ABCB10	-2,13	0,0405
48	235368_at	ADAMTS5	2,20	0,0411
49	206157_at	PTX3	3,53	0,0418
50	228885_at	MAMDC2	-3,31	0,0424
51	202018_s_at	LTF	3,45	0,0426
52	226814_at	ADAMTS9	2,72	0,0436
53	206111_at	RNASE2	2,90	0,0436
54	212224_at	ALDH1A1	-2,28	0,0442
55	227930_at	AGO4	-2,22	0,0446
56	235568_at	MCEMP1	2,73	0,0458
57	241068_at	IGSF6	-2,02	0,0473
58	202436_s_at	CYP1B1	2,11	0,0473
59	225290_at	ETNK1	-2,54	0,0481
60	206100_at	CPM	-2,10	0,0491
61	202969_at	DYRK2	-2,04	0,0494
62	213060_s_at	CHI3L2	2,33	0,0496

Apêndice B

Resultado do método de seleção utililizando modelo linear para análise de microarranjo *(limma)*

Tabela B.1: Tabela dos genes obtidos através da técnica limma.

Rank	ID	Símbolo	Valor p ajustado	*Fold-change*
1	221009_s_at	ANGPTL4	0,00005	2,11
2	206025_s_at	TNFAIP6	0,00005	3,37
3	202237_at	NNMT	0,00021	2,38
4	212013_at	PXDN	0,00021	2,62
5	212110_at	SLC39A14	0,00050	2,10
6	210119_at	KCNJ15	0,00056	2,32
7	206561_s_at	AKR1B10	0,00058	3,75
8	221898_at	PDPN	0,00061	2,36
9	203608_at	ALDH5A1	0,00080	-2,27
10	205863_at	S100A12	0,00120	3,92
11	201005_at	CD9	0,0013	-2,14
12	205542_at	STEAP1	0,0013	2,52
13	201110_s_at	THBS1	0,0016	2,51
14	205656_at	PCDH17	0,0017	2,07
15	205934_at	PLCL1	0,0017	-2,25
16	208850_s_at	THY1	0,0017	2,29
17	206958_s_at	UPF3A	0,0020	-2,00
18	218613_at	PSD3	0,0020	-2,30
19	236081_at	SNCA	0,0020	-2,21
20	219895_at	TMEM255A	0,0020	-3,90
21	225327_at	FAM214A	0,0020	-2,01
22	202922_at	GCLC	0,0020	-2,45
23	227930_at	AGO4	0,0023	-2,22
24	224009_x_at	DHRS9	0,0024	-2,02
25	206111_at	RNASE2	0,0026	2,90
26	229947_at	PI15	0,0027	3,34
27	205119_s_at	FPR1	0,0028	2,98
28	226814_at	ADAMTS9	0,0030	2,72
29	213060_s_at	CHI3L2	0,0031	2,33
30	1554018_at	GPNMB	0,0035	-2,26
31	226867_at	DENND4C	0,0036	-2,04
32	206101_at	ECM2	0,0037	-2,27
33	235780_at	PRKACB	0,0038	-2,13
34	201243_s_at	ATP1B1	0,0038	-2,40
35	205798_at	IL7R	0,0039	2,30
36	206157_at	PTX3	0,0042	3,53
37	205907_s_at	OMD	0,0047	-2,10
38	228885_at	MAMDC2	0,0049	-3,31
39	209125_at	KRT6A	0,0053	4,10
40	225579_at	PQLC3	0,0055	-2,06
41	225290_at	ETNK1	0,0058	-2,54
42	227525_at	GLCCI1	0,0058	-2,48

Tabela B.1: Tabela dos genes obtidos através da técnica limma. (continuação)

Rank	ID	Símbolo	Valor p ajustado	*Fold-change*
43	214370_at	S100A8	0,0062	2,71
44	209800_at	KRT16	0,0062	2,14
45	214461_at	LBP	0,0064	2,49
46	205844_at	VNN1	0,0065	-2,14
47	223320_s_at	ABCB10	0,0070	-2,13
48	225915_at	CAB39L	0,0070	-2,16
49	218885_s_at	GALNT12	0,0070	-2,89
50	205916_at	S100A7	0,0070	4,58
51	205828_at	MMP3	0,0072	4,56
52	205559_s_at	PCSK5	0,0072	2,30
53	211906_s_at	SERPINB4	0,0076	4,39
54	204719_at	ABCA8	0,0078	-2,07
55	212224_at	ALDH1A1	0,0078	-2,28
56	204105_s_at	NRCAM	0,0079	2,05
57	204475_at	MMP1	0,0082	6,50
58	219869_s_at	SLC39A8	0,0083	2,06
59	202969_at	DYRK2	0,0084	-2,04
60	220380_at	DNASE2B	0,0086	-2,24
61	235568_at	MCEMP1	0,0086	2,73
62	223767_at	GPR84	0,0090	2,15
63	238018_at	ALKAL2	0,0094	-2,51
64	235368_at	ADAMTS5	0,0098	2,20
65	221950_at	EMX2	0,0098	-2,18
66	202018_s_at	LTF	0,0107	3,45
67	241068_at	IGSF6	0,0118	-2,02
68	212820_at	DMXL2	0,0118	-2,01
69	227140_at	INHBA	0,0122	2,31
70	206100_at	CPM	0,0124	-2,10
71	202436_s_at	CYP1B1	0,0127	2,11
72	231993_at	ITGBL1	0,0130	-2,37
73	219014_at	PLAC8	0,0130	2,23
74	202376_at	SERPINA3	0,0132	2,64
75	203649_s_at	PLA2G2A	0,0134	2,02
76	219908_at	DKK2	0,0139	-2,66
77	203824_at	TSPAN8	0,0152	-2,48
78	232170_at	S100A7A	0,0179	3,45
79	203691_at	PI3	0,0188	2,94
80	203789_s_at	SEMA3C	0,0188	-2,11
81	227736_at	C10orf99	0,0195	2,94
82	203535_at	S100A9	0,0195	3,08
83	203434_s_at	MME	0,0196	2,72
84	205943_at	TDO2	0,0228	3,85

Tabela B.1: Tabela dos genes obtidos através da técnica limma. (continuação)

Rank	ID	Símbolo	Valor p ajustado	*Fold-change*
85	236341_at	CTLA4	0,0231	-2,06
86	239349_at	C1QTNF7	0,0244	-2,06
87	206509_at	PIP	0,0330	-2,40
88	209613_s_at	ADH1B	0,0345	-3,04
89	202952_s_at	ADAM12	0,0379	2,47
90	208089_s_at	TDRD3	0,0392	-2,02
91	220356_at	CORIN	0,0398	-2,30
92	209720_s_at	SERPINB3	0,0401	3,59

Apêndice C

Resultado do método de seleção utililizando MD-ACP

Tabela C.1: Tabela dos genes obtidos através da técnica MD-ACP.

Rank	ID	Símbolo	Diferença de Contribuição
1	204475_at	MMP1	0,308
2	211906_s_at	SERPINB4	0,172
3	205916_at	S100A7	0,166
4	205828_at	MMP3	0,156
5	205863_at	S100A12	0,131
6	219895_at	TMEM255A	0,122
7	232170_at	S100A7A	0,117
8	209125_at	KRT6A	0,114
9	206561_s_at	AKR1B10	0,111
10	205943_at	TDO2	0,109
11	209720_s_at	SERPINB3	0,105
12	202018_s_at	LTF	0,104
13	228885_at	MAMDC2	0,102
14	203535_at	S100A9	0,097
15	236203_at	HLA-DQA1	0,094
16	203691_at	PI3	0,090
17	229947_at	PI15	0,090
18	209613_s_at	ADH1B	0,085
19	206025_s_at	TNFAIP6	0,084
20	205119_s_at	FPR1	0,081
21	206157_at	PTX3	0,076
22	227736_at	C10orf99	0,076
23	218885_s_at	GALNT12	0,072
24	202376_at	SERPINA3	0,071
25	206378_at	SCGB2A2	0,069
26	219908_at	DKK2	0,064
27	225290_at	ETNK1	0,064
28	216248_s_at	NR4A2	0,063
29	235568_at	MCEMP1	0,059
30	214370_at	S100A8	0,059
31	203434_s_at	MME	0,059
32	227525_at	GLCCI1	0,058
33	214461_at	LBP	0,058
34	206111_at	RNASE2	0,057
35	206509_at	PIP	0,056
36	209480_at	HLA-DQB1	0,055
37	203824_at	TSPAN8	0,055
38	220356_at	CORIN	0,055
39	226814_at	ADAMTS9	0,054
40	212013_at	PXDN	0,053
41	202922_at	GCLC	0,052
42	203789_s_at	SEMA3C	0,050

Tabela C.1: Tabela dos genes obtidos através da técnica MD-ACP. (continuação)

Rank	ID	Símbolo	Diferença de Contribuição
43	238018_at	ALKAL2	0,049
44	218613_at	PSD3	0,048
45	202952_s_at	ADAM12	0,048
46	1554018_at	GPNMB	0,047
47	206101_at	ECM2	0,047
48	213060_s_at	CHI3L2	0,047
49	203608_at	ALDH5A1	0,047
50	227140_at	INHBA	0,047
51	221950_at	EMX2	0,047
52	205908_s_at	OMD	0,047
53	223320_s_at	ABCB10	0,046
54	227930_at	AGO4	0,046
55	225915_at	CAB39L	0,046
56	208850_s_at	THY1	0,045
57	223767_at	GPR84	0,045
58	201243_s_at	ATP1B1	0,045
59	231993_at	ITGBL1	0,045
60	213929_at	EXPH5	0,044
61	205934_at	PLCL1	0,044
62	202237_at	NNMT	0,043
63	218432_at	FBXO3	0,043
64	1569410_at	FLG2	0,043
65	206799_at	SCGB1D2	0,043
66	210119_at	KCNJ15	0,043
67	201110_s_at	THBS1	0,043
68	205542_at	STEAP1	0,043
69	209800_at	KRT16	0,043
70	206100_at	CPM	0,042
71	206569_at	IL24	0,042
72	219890_at	CLEC5A	0,041
73	204719_at	ABCA8	0,041
74	221898_at	PDPN	0,040
75	226867_at	DENND4C	0,040
76	202969_at	DYRK2	0,040
77	212224_at	ALDH1A1	0,040
78	226529_at	TMEM106B	0,039
79	205844_at	VNN1	0,039
80	235780_at	PRKACB	0,039
81	206958_s_at	UPF3A	0,039
82	220380_at	DNASE2B	0,038
83	239349_at	C1QTNF7	0,038
84	205559_s_at	PCSK5	0,038

Tabela C.1: Tabela dos genes obtidos através da técnica MD-ACP. (continuação)

Rank	ID	Símbolo	Diferença de Contribuição
85	205064_at	SPRR1B	0,038
86	206633_at	CHRNA1	0,038
87	202704_at	TOB1	0,037
88	225579_at	PQLC3	0,037
89	212820_at	DMXL2	0,037
90	225327_at	FAM214A	0,037
91	201005_at	CD9	0,037
92	204146_at	RAD51AP1	0,037
93	223952_x_at	DHRS9	0,037