

**ANÁLISE ESTATÍSTICA BASEADA NA CONSTRUÇÃO DE
ÁRVORES DE CLASSIFICAÇÃO E REGRESSÃO DE MICRORNAS
DE PACIENTES COM ADENOCARCINOMA PULMONAR**

Bethina da Rocha Camargo

Dissertação apresentada à Universidade Estadual Paulista “Júlio de Mesquita Filho” para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro – 2020

**ANÁLISE ESTATÍSTICA BASEADA NA CONSTRUÇÃO DE
ÁRVORES DE CLASSIFICAÇÃO E REGRESSÃO DE MICRORNAS
DE PACIENTES COM ADENOCARCINOMA PULMONAR**

Bethina da Rocha Camargo

Orientador: Prof. Dr. **Rogério Antonio de Oliveira**

Coorientadora: Prof.^a Dr.^a **Patrícia Pintor dos Reis**

Dissertação apresentada à Universidade Estadual Paulista “Júlio de Mesquita Filho” para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro – 2020

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Camargo, Bethina da Rocha.
Análise estatística baseada na construção de árvores de
classificação e regressão de microRNAs de pacientes com
adenocarcinoma pulmonar / Bethina da Rocha Camargo. -
Botucatu, 2020

Dissertação (mestrado) - Universidade Estadual Paulista
"Júlio de Mesquita Filho", Instituto de Biociências de
Botucatu
Orientador: Rogério Antonio de Oliveira
Coorientador: Patrícia Pintor dos Reis
Capes: 10202080

1. Pulmões - Câncer. 2. Análise de sobrevivência
(Biometria). 3. Adenocarcinoma. 4. MicroRNAs.

Palavras-chave: Análise de sobrevivência; Bioestatística;
Biomarcadores; Câncer de pulmão; Genética.

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE BETHINA DA ROCHA CAMARGO, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA, DO INSTITUTO DE BIOCIÊNCIAS - CÂMPUS DE BOTUCATU.

Aos 19 dias do mês de fevereiro do ano de 2020, às 14:00 horas, no(a) Departamento de Bioestatística, reuniu-se a Comissão Examinadora da Defesa Pública, composta pelos seguintes membros: Prof. Dr. ROGERIO ANTONIO DE OLIVEIRA - Orientador(a) do(a) Depto. de Bioestatística / IB/Botucatu - Unesp, Profa. Dra. HILDETE PRISCO PINHEIRO do(a) Departamento de Estatística / Universidade Estadual de Campinas - UNICAMP (Por meio de videoconferência), Profa. Dra. DANIELE CRISTINA CATANEO do(a) Departamento de Cirurgia e Ortopedia / Faculdade de Medicina de Botucatu - UNESP, sob a presidência do primeiro, a fim de proceder a arguição pública da DISSERTAÇÃO DE MESTRADO de BETHINA DA ROCHA CAMARGO, intitulada **ANÁLISE ESTATÍSTICA BASEADA NA CONSTRUÇÃO DE ÁRVORES DE CLASSIFICAÇÃO E REGRESSÃO DE MICRORNAS DE PACIENTES COM ADENOCARCINOMA PULMONAR**. Após a exposição, a discente foi arguida oralmente pelos membros da Comissão Examinadora, tendo recebido o conceito final: APROVADA. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelos membros da Comissão Examinadora.

Prof. Dr. ROGERIO ANTONIO DE OLIVEIRA

Profa. Dra. HILDETE PRISCO PINHEIRO

Profa. Dra. DANIELE CRISTINA CATANEO

Mas quando você orar, vá para seu quarto, feche a porta e ore a seu Pai, que está no secreto. Então seu Pai, que vê no secreto, o recompensará.

Mateus 6:6

Dedicatória

Dedico primeiramente a Deus que me guiou durante esta trajetória, a meus pais Jarbas e Marlene, meu irmão Benedito e meu namorado Felipe. Meu carinho e admiração por vocês é eterna, amo vocês.

Agradecimentos

Primeiramente agradeço a Deus, por ser essencial em minha vida, autor do meu destino, meu guia durante esta caminhada e socorro nos momentos difíceis.

Em especial ao meu pai Jarbas, minha mãe Marlene, meu irmão Benedito e meu namorado Felipe que dividem comigo todos os seus dias, são meus maiores exemplos de vida, inspiração e motivação. A vocês o meu muito obrigada, por me apoiarem e estarem ao meu lado ajudando na realização deste sonho.

Ao meu orientador Prof. Dr. Rogério Antonio de Oliveira e coorientadora Prof.^a Dr.^a Patrícia Pintor dos Reis pelo aprendizado, auxílio, motivação, dedicação e confiança.

À banca examinadora da dissertação Prof.^a Dr.^a Hildete P. Pinheiro e Prof.^a Dr.^a Daniele C. Cataneo pelas sugestões e comentários relevantes.

A todos os professores do departamento de Bioestatística UNESP/IBB. Em especial a Prof.^a Dr.^a Liciane e Prof.^a Dr.^a Miriam pelas contribuições na qualificação.

Em especial aos meus amigos Elizabete, Juliana, Gustavo, Eduardo e também, a todos os colegas de departamento pelos excelentes momentos que compartilhamos.

Aos Funcionários Arthur, Júnior e Luiz por sempre estarem dispostos a ajudar.

Ao Departamento de Bioestatística do IBB pelos ambientes adequados que possibilitaram a realização deste trabalho e ao Programa de Pós-Graduação em Biometria.

Agradeço apoio da Coordenação de Aperfeiçoamento Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Sumário

	Página
LISTA DE FIGURAS	ix
LISTA DE TABELAS	xi
RESUMO	xii
ABSTRACT	xiii
1 INTRODUÇÃO	1
1.1 Câncer de pulmão	2
1.2 MicroRNAs e o câncer	4
1.3 MicroRNAs e alvos moleculares terapêuticos	6
1.4 Árvore de classificação e regressão	7
2 OBJETIVOS	9
3 TESTES E MODELOS ESTATÍSTICOS	11
3.1 Teste de Wilcoxon-Mann-Whitney	11
3.2 Regressão logística	12
3.2.1 Teste qui-quadrado de Pearson	15
3.2.2 Estatística C	17
3.3 Análise de sobrevivência	17
3.3.1 Funções associadas ao tempo de sobrevivência	19
3.3.2 Relações entre as funções	20

	viii
3.3.3 Estimador de Kaplan Meier	21
3.3.4 Teste de log-rank	22
3.4 Construção de árvores de decisão	24
3.4.1 O método CART	26
3.4.2 Árvore de classificação	26
3.4.3 Árvore de regressão	28
3.4.4 Crescimento e poda	29
4 ANÁLISE DE DADOS E RESULTADOS	31
4.1 Análise exploratória	31
4.2 Teste de Wilcoxon-Mann-Whitney	35
4.3 Regressão logística	41
4.4 Análise do tempo de seguimento do paciente	43
4.5 Árvores de classificação e regressão	48
4.5.1 Árvore de classificação	48
4.5.2 Árvore de sobrevivência	54
4.6 Análise biológica	60
5 CONCLUSÕES	64
REFERÊNCIAS BIBLIOGRÁFICAS	66

Lista de Figuras

	Página
1 Anatomia e função dos pulmões humano.	1
2 Ilustração de uma árvore de decisão.	25
3 Algoritmo de divisão do CART.	27
4 Distribuição da idade e dos dias até o último acompanhamento.	32
5 Boxplot da expressão dos microRNAs por tecido normal e tumoral para o teste de Wilcoxon-Mann-Whitney.	37
6 Boxplot da expressão dos microRNAs por tecido normal e tumoral para o teste de Wilcoxon-Mann-Whitney.	38
7 Boxplot da expressão dos microRNAs por tecido normal e tumoral para o teste de Wilcoxon-Mann-Whitney.	39
8 Boxplot da expressão dos microRNAs por tecido normal e tumoral para o teste de Wilcoxon-Mann-Whitney.	40
9 Boxplot da expressão do microRNA por tecido normal e tumoral para o teste de Wilcoxon-Mann-Whitney.	41
10 Boxplot da expressão dos microRNAs por tecido normal e tumoral para a regressão logística.	42
11 Boxplot da expressão dos microRNAs por tecido normal e tumoral para a regressão logística.	43
12 Curvas de sobrevivência estimadas por meio de Kaplan Meier para os grupos de estágio do tumor.	44
13 Curvas de sobrevivência estimadas por meio de Kaplan Meier para as etnias.	45

14	Curvas de sobrevivência por meio do estimador de Kaplan Meier para o miR-34b-3p e miR-185-3p.	46
15	Curvas de sobrevivência por meio do estimador de Kaplan Meier para o miR-1247-3p e miR-4661-5p.	47
16	Boxplot da expressão do microRNA por tecido normal e tumoral para a análise do tempo de seguimento do paciente.. . . .	48
17	Árvore de classificação sem poda, em que, as cores azul e cinza representam tecido tumoral e normal, respectivamente.	49
18	Gráfico do custo-complexidade, erro de validação cruzada e número de divisões para árvore de classificação.	51
19	Gráfico do número de divisões pelo erro de validação cruzada para árvore de classificação.	52
20	Árvore de classificação pós poda, em que, as cores azul e cinza representam tecido tumoral e normal, respectivamente.	53
21	Boxplot da expressão dos microRNAs por tecido normal e tumoral para a árvore de classificação pós poda.	54
22	Árvore de sobrevivência sem poda.	55
23	Gráfico do custo-complexidade, erro de validação cruzada e número de divisões para árvore de sobrevivência.	57
24	Gráfico do número de divisões pelo erro de validação cruzada para árvore de sobrevivência.	57
25	Árvore de sobrevivência pós poda. RR: Risco relativo.	58
26	Boxplot da expressão dos microRNAs por tecido normal e tumoral para a árvore de sobrevivência pós poda.	59
27	Boxplot da expressão do microRNA por tecido normal e tumoral para a árvore de sobrevivência pós poda.	60
28	Genes para a via molecular metabolismo celular do nicotinato e nicotinamida.	62

Lista de Tabelas

	Página
1 Tabela de contingência para o tempo t_j	23
2 Análises descritivas das variáveis qualitativas dos dados demográficos e características clínico-patológicas.	33
3 Análises descritivas das variáveis quantitativas dos dados demográficos e características clínico-patológicas.	33
4 Análises descritivas das variáveis qualitativas de acordo com o tecido e valor p	34
5 Análises descritivas das variáveis quantitativas de acordo com o tecido e valor p	35
6 Expressão dos microRNAs significativos por tecido normal ou tumoral e valor p	36
7 Estimativas dos parâmetros do modelo de regressão logística.	42
8 Resumo do modelo para a poda da árvore de classificação.	50
9 Resumo do modelo para a poda da árvore de sobrevivência.	56
10 Vias moleculares identificadas a partir dos genes-alvo de microRNAs associados a árvore de classificação.	61
11 Vias moleculares identificadas a partir dos genes-alvo de microRNAs associados a árvore de sobrevivência.	63

ANÁLISE ESTATÍSTICA BASEADA NA CONSTRUÇÃO DE ÁRVORES DE CLASSIFICAÇÃO E REGRESSÃO DE MICRORNAS DE PACIENTES COM ADENOCARCINOMA PULMONAR

Autora: BETHINA DA ROCHA CAMARGO

Orientador: Prof. Dr. ROGÉRIO ANTONIO DE OLIVEIRA

Coorientadora: Prof.^a Dr.^a PATRÍCIA PINTOR DOS REIS

RESUMO

O adenocarcinoma pulmonar é um problema mundial na saúde pública e representa uma das maiores causas de morte por câncer no mundo. Os microRNAs são grandes reguladores e têm sido propostos como biomarcadores em diversos tipos de cânceres. O objetivo desta pesquisa é encontrar possíveis microRNAs para melhorar a classificação dos tecidos (normal ou tumoral) e a sobrevivência dos pacientes. Utilizou-se o banco de dados do projeto Atlas do Genoma do Câncer (TCGA) para pacientes com adenocarcinoma pulmonar (LUAD). A análise estatística empregada foi baseada na construção de árvores de classificação, que encontrou o miR-21-5p, miR-133a-3p, miR-1287-3p e let-7g-3p estatisticamente significativos, e árvore de regressão para a sobrevivência dos pacientes, que encontrou o miR-887-3p, miR-1271-5p, miR-128-1-5p, miR-493-3p e miR-4999-5p estatisticamente significativos.

STATISTICAL ANALYSIS BASED ON THE CONSTRUCTION OF CLASSIFICATION TREES AND REGRESSION OF MICRORNAS FROM PATIENTS WITH PULMONARY ADENOCARCINOMA

Author: BETHINA DA ROCHA CAMARGO

Adviser: Prof. Dr. ROGÉRIO ANTONIO DE OLIVEIRA

Coadviser: Prof.^a Dr.^a PATRÍCIA PINTOR DOS REIS

ABSTRACT

Pulmonary adenocarcinoma is a worldwide public health problem and represents one of the biggest causes of cancer death in the world. MicroRNAs are strong regulators and have been proposed as biomarkers in several types of cancer. The objective of this research is to find possible microRNAs to improve the classification of tissues (normal or tumor) and patient survival. We used the database of the Atlas of Cancer Genome project (TCGA) for patients with pulmonary adenocarcinoma (LUAD). A statistical analysis employed was based on the construction of classification trees, which found miR-21-5p, miR-133a-3p, miR-1287-3p and let-7g-3p, with statistical classification, and regression tree for the situation of patients, who found miR-887-3p, miR-1271-5p, miR-128-1-5p, miR-493-3p and miR-4999-5p are statistically used.

1 INTRODUÇÃO

O pulmão é o órgão do sistema respiratório responsável pelas trocas gasosas entre o sangue e o ambiente. Sua principal função é oxigenar o sangue e eliminar o dióxido de carbono (CO_2) do corpo. Os seres humanos possuem dois pulmões que estão localizados na cavidade torácica e são cobertos por uma membrana protetora chama pleura. Cada pulmão, direito ou esquerdo, é dividido em partes ou lobos, o pulmão direito é maior e possui três lobos e o esquerdo possui apenas dois.

A Figura 1 ilustra a anatomia e a função dos pulmões humano, pode-se ressaltar que o órgão tem o formato piramidal, cor rosada e consistência esponjosa.

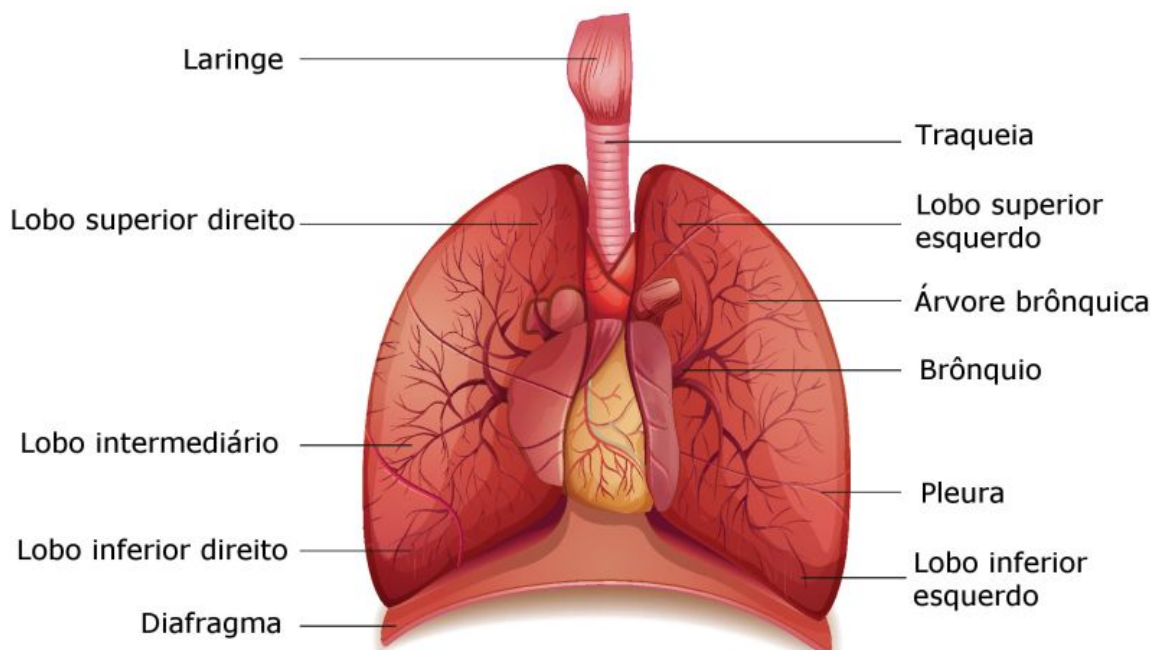


Figura 1: Anatomia e função dos pulmões humano.

Fonte: <https://www.infoescola.com/anatomia-humana/pulmoes/>

1.1 Câncer de pulmão

O câncer de pulmão é um dos cânceres mais populares e líder em mortalidade, com aproximadamente 1,8 milhões de casos novos e cerca de 1,6 milhões de óbitos por ano, mundialmente (Globocan, 2018). No Brasil, as estimativas de incidência para os anos de 2018 e 2019 apontaram o câncer de pulmão como o segundo mais frequente, com ocorrência de aproximadamente 31 mil casos novos, sendo a maioria em homens (18.740 homens e 12.530 mulheres). No sudeste do Brasil a incidência de carcinomas de pulmão em 2018 foi a terceira mais frequente com 19,22 casos a cada cem mil homens e 12,72 casos a cada cem mil mulheres (INCA, 2018).

O carcinoma pulmonar é geralmente diagnosticado em estágio avançado da doença, principalmente porque os sintomas não são percebidos nas fases iniciais de seu desenvolvimento, o diagnóstico tardio é apontado como responsável pela baixa sobrevivência dos pacientes, de 10% a 15% dos pacientes diagnosticados (INCA, 2018).

O tabagismo está relacionado como a principal causa de câncer de pulmão, na maioria da população os casos tabaco-relacionados representam cerca de 85% dos casos desse câncer (Ferlay et al., 2013). Além disso, há outros fatores de risco para este carcinoma, como a radiação e exposição à carcinógenos ocupacionais, como agentes químicos, e ambientais como, a poluição do ar, por exemplo (Schmid et al., 2010). Desta forma, esta doença pode ocorrer em indivíduos fumantes, ex-fumantes e não fumantes.

O câncer pulmonar é uma preocupante doença da saúde pública, assim deve ser priorizado a prevenção da doença e o desenvolvimento de melhores estratégias de tratamento.

Para decisões referentes a possíveis tratamentos do câncer de pulmão deve-se atentar aos subtipos histológicos¹ envolvidos neste tipo de carcinoma, que pode ser dividido em dois principais grupos: o carcinoma pulmonar de células pequenas (SCC, em inglês *small cell carcinoma*) e o carcinoma pulmonar de células não pe-

¹Estudos da estrutura microscópica, composição e função dos tecidos vivos.

quenas (NSCLC, em inglês *non-small cell lung carcinoma*). O primeiro ocorre em 15% dos pacientes, caracteriza-se por tumores mais agressivos e está associado a um prognóstico pobre com baixa sobrevivência. O segundo, é o tipo histológico mais comum, com 85% dos casos, é subdividido em vários subtipos histológicos, sendo o adenocarcinoma (AD, em inglês *adenocarcinoma*) o mais frequentemente diagnosticado (Herbst et al., 2008).

O AD é um problema mundial na saúde pública, representa um dos grandes motivos para as causas de morte por câncer no mundo, forma-se na periferia dos pulmões e é o subtipo mais encontrado de câncer de pulmão entre os não-fumantes, principalmente em mulheres (Chang et al., 2015).

Os pacientes diagnosticados com NSCLC podem fazer o tratamento da doença por meio de cirurgia, quimioterapia, radioterapia e, atualmente por meio de terapias com alvos moleculares, que nos últimos anos obtiveram vários marcadores moleculares e modelos propostos pelo mundo. Entretanto, devido à heterogeneidade observada nos pacientes com AD, o desenvolvimento de tratamentos aprimorados, como, por exemplo biomarcadores para o prognóstico da doença e para a prevenção do aparecimento da doença, ainda é um grande problema devido à complexidade biologia do câncer (Souza, 2016).

Estudos têm sido realizados com a intenção de identificar novas mutações ou alterações no carcinoma pulmonar e tais estudos têm identificado mutações em NSCLC. As alterações moleculares responsáveis pelo processo de formação do carcinoma pulmonar, especialmente o subtipo AD, ainda não são muito conhecidas. Considerando, então, que as terapias com alvos moleculares têm beneficiado poucos pacientes, ainda é necessária a identificação de novos alvos moleculares relacionados a este carcinoma (Chang et al., 2015).

O estudo sobre os microRNAs serve como base para o esclarecimento das causas de doenças e dos mecanismos que atuam sobre o organismo, também tem propiciado utilização para diagnóstico, prognóstico e terapia do câncer (Chamorro et al., 2016).

Nos últimos anos, confirmou-se que a desregulação na expressão dos microRNAs está relacionada à várias enfermidades em humanos, como problemas cardíacos e em cânceres.

A colaboração entre o Instituto Nacional do Câncer (NCI, em inglês *National Cancer Institute*) e o Instituto Nacional de Pesquisa do Genoma Humano (NHGRI, em inglês *National Human Genome Research Institute*) gerou a caracterização do perfil mutacional do pulmão, inclusive da expressão dos microRNAs. Nomeado de Atlas do Genoma do Câncer (TCGA, em inglês *The Cancer Genome Atlas*), é um banco de dados *online*, está publicamente disponível e tem sido vastamente utilizado pela comunidade de pesquisa, para realização de estudos sobre a desregularização da expressão dos microRNAs em vários tipos de cânceres.

1.2 MicroRNAs e o câncer

O primeiro microRNA encontrado em pesquisas foi por Lee et al. (1993). O microRNA lin-4, foi relacionado com o tempo de progresso do *nematóide Caenorhabditis elegans*, por meio da modulação da proteína lin-14.

Em seguida, definiu-se que sequências de pequenos RNAs, não codificadores, de 19 a 22 nucleotídeos de extensão estavam, também, expressas em outros tipos de organismos, como em *Homo sapiens*. Estas sequências foram nomeadas de microRNAs, são encontradas em diferentes espécies e cada espécie apresenta sua particularidade a respeito da expressão do microRNA (Iorio & Croce, 2009).

A nomenclatura do microRNA é composta pela espécie e o número do microRNA, nesta ordem, de acordo com a sua descoberta. Assim, o hsa-miR-21 refere-se ao microRNA 21 em *Homo sapiens*.

Encontrar a expressão dos microRNAs em cânceres humanos não criou expectativas para possíveis tratamentos até Calin et al. (2002) clonarem os miR-15a e miR-16-1, que inibiam a expressão de um gene anti apoptótico, em pacientes com leucemia linfocítica crônica. Os microRNAs agiam como supressores tumorais e estavam com a expressão diminuída em 70%.

Após este fato, outros estudos surgiram descrevendo diferentes perfis da expressão dos microRNAs em outras categorias de câncer, entre eles, o câncer de pulmão, cólon, próstata, entre outros. Além disso, também pode haver outras doenças relacionadas a diminuição ou aumento da expressão dos microRNAs. É importante o reconhecimento da função regulatória destes, para assim, obter clareza em seu papel no desenvolvimento das doenças (Liu et al., 2012).

Evidenciou-se a expressão alterada dos microRNAs ao comparar tecidos normais e tumorais e, também, em alguns cânceres específicos os microRNAs mostraram expressões diferentes nas fases da doença, ou seja, desde a formação do câncer até a metástase (Medina & Slack, 2008).

As expressões dos microRNAs podem ser alteradas em relação as variações do câncer, condições clínicas, estágio e perfil genético do paciente. Percebe-se, de um modo geral, que há uma redução significativa da expressão de microRNAs na presença do câncer, ainda assim, existem muitos microRNAs que podem estar superexpressos, ou seja, com a expressão do microRNA aumentada em relação as expressões consideradas normais. Cada tecido possui um valor referência para a expressão dos microRNA, logo a expressão do microRNA é capaz de estar aumentada ou diminuída em um tipo de câncer e o oposto em outro (Lee & Dutta, 2009).

Em alguns casos, o aparecimento e crescimento de tumores está relacionado a superexpressão dos microRNAs, tem-se como exemplos o miR-221 em glioblastomas², miR-372 e miR-373 em neoplasias do testículo e miR-221, miR-222 e miR-146 em tumores de tireoide (Lee & Dutta, 2009).

A literatura mostra que os microRNAs podem ser responsáveis pelos processos de invasão e migração de tecidos tumorais, caracterizando tumores malignos. Pesquisas encontraram superexpressão, do miR-21 em cânceres. O miR-21 é caracterizado por promover a mobilidade e invasão de tecidos tumorais, pois liga-se e bloqueia os genes supressores de tumor (Lee & Dutta, 2009).

Um estudo clínico sobre o papel do let-7 (let-7a ao let-7i) em câncer de

²Tipo mais comum e agressivo de tumor maligno cerebral que acomete os seres humanos.

pulmão, mostrou um diagnóstico ruim quando este está subexpresso, ou seja, com a expressão do microRNAs diminuída em relação as expressões consideradas normais, e a introdução do let-7 como um supressor tumoral (Takamizawa et al., 2004).

Outras pesquisas discutiram a importância da expressão dos microRNAs para o processo de formação do câncer e para o desenvolvimento do câncer de mama, já que os hormônios podem apresentar expressões diferentes nestas fases da doença (Nakano et al., 2013). Diferentes expressões de microRNAs têm sido retratadas ao comparar tecidos tumorais e normais em vários os tipos de cânceres (Yin et al., 2014).

Tsunoda et al. (2011) em sua pesquisa constatou que a inibição da morte celular programada está ligada com a superexpressão do miR-221 e miR-222 em pacientes com câncer colorretal.

Com a grande quantidade de informações a respeito dos microRNAs, surgiu a necessidade de organizar as informações em banco de dados. Hoje, existem vários bancos de dados virtuais que fornecem informações sobre os microRNAs. O miR-base, foi criado em 2002, representa o principal depósito online que contém todas as informações sobre os microRNAs publicados pelo mundo.

1.3 MicroRNAs e alvos moleculares terapêuticos

No decorrer dos anos, tem-se percebido crescimento nas pesquisas e no desenvolvimento de técnicas de biologia molecular, mostrando uma tendência em cada vez mais entender e desenvolver terapias-alvos específicas, com o objetivo de bloquear vias moleculares desreguladas nos tecidos tumorais.

Desde que os microRNAs foram descobertos, ampliou-se as possibilidades de tê-los como biomarcadores para o diagnóstico e prognóstico de cânceres, principalmente nas fases iniciais, em que não é possível detectar a doença com exames convencionais. Esses tratamentos têm como objetivo atuar seletivamente nas células tumorais que contém alterações moleculares (Cinegaglia, 2015).

A maioria dos estudos utilizam do soro para a identificação dos níveis de

expressão dos microRNAs, entretanto, outros estudos relatam que a identificação destes níveis pode ser realizada por meio da urina, saliva, suco pancreático ou fluido cístico. Encontrar os microRNAs nestes tipos de fluidos corporais são fundamentais para detecção de possíveis biomarcadores e representam uma forma não-invasiva para o diagnóstico da doença, principalmente em pacientes com a saúde mais fraca (Wang et al., 2014).

Algumas pesquisas a respeito de terapias baseadas em microRNAs possuem a intenção de restaurar ou inibir a função dos microRNAs usando microRNAs *mimics* ou *antagomirs*, que são como uma imitação dos microRNAs. Estudos *in vitro* e *in vivo* tem mostrado a eficácia do uso de microRNAs *mimics* contra o desenvolvimento tumoral (Ulivi et al., 2013). Porém pesquisas estão sendo desenvolvidas para definir a segurança na utilização dos microRNAs para este fim. Primeiramente são identificados os microRNAs relacionados ao tipo de câncer e posteriormente estudadas as necessidades biológicas para a ação.

A compreensão sobre os processos associados a formação do câncer humano, a caracterização do perfil genético, as alterações na expressão e mutações de microRNAs em adenocarcinoma, poderá contribuir para o desenvolvimento de biomarcadores com potencial aplicações clínicas. Biomarcadores moleculares serão úteis na evolução de estratégias para identificação precoce de doença, diagnóstico, prognóstico e terapêuticas mais precisas para o tratamento de pacientes.

1.4 Árvore de classificação e regressão

A aplicação de técnicas com segmentação ou agrupamentos na construção de árvores foi desenvolvida para solucionar os problemas complexos que envolvem conjunto de dados de dimensão elevada, composto por muitas variáveis qualitativas e quantitativas.

O método de árvores de classificação e regressão apareceu inicialmente no trabalho de Morgan & Sonquist (1963). Posteriormente nos trabalhos de Hunt et al. (1966), Messenger & Mandell (1972) e Morgan & Messenger (1973), que apresenta-

ram este método para problemas de classificação. Entretanto, as modificações tiveram uma grande contribuição para a popularização de árvores binárias em problemas de classificação por meio dos trabalhos realizados por Quinlan (1986) e também nos seus trabalhos dos anos 1979 e 1983, e na mesma época por Breiman et al. (1984).

Os problemas de regressão utilizaram a árvore primeiramente no trabalho de Morgan & Sonquist (1963) com o programa AID (*Automatic Interaction Detection*) e posteriormente apareceram generalizações como em Sonquist (1970), Sonquist et al. (1973) e em Van Eck (1980).

O algoritmo para árvores de classificação e regressão (CART, em inglês *Classification and Regression Trees*) foi apresentado por Breiman et al. (1984) e hoje está implementado em muitos programas estatísticos, como o *software* R. O algoritmo é uma técnica não-paramétrica que gera árvores de classificação ou de regressão, de acordo variável, qualitativa ou quantitativa, respectivamente.

O CART apresenta algumas vantagens em sua utilização, como: capacidade de encontrar relações entre os dados; os resultados são gerados no formato de árvore de decisão, implicando em facilidade, legibilidade e simplicidade com a interpretação dos resultados (Fonseca, 1994). As árvores de decisão que o algoritmo CART gera possuem respostas binárias e podem ser percorridas do nó pai até os nós finais, respondendo a perguntas com respostas de “sim” ou “não”.

Com o intuito de encontrar microRNAs como potenciais biomarcadores utiliza-se como método estatístico a construção de árvores de classificação e regressão, por ser uma técnica que gera resultados de fácil interpretação conforme descrito no Capítulo 3.

2 OBJETIVOS

Os objetivos do trabalho foram divididos em dois tópicos: objetivo principal e objetivos específicos.

- **Objetivo principal**

O trabalho tem como objetivo principal encontrar microRNAs com potencial para possíveis biomarcadores e, assim, melhorar as estratégias de tratamento dos pacientes com adenocarcinoma pulmonar.

- **Objetivos específicos**

O trabalho tem como objetivos específicos encontrar os microRNAs estatisticamente significativos por meio de quatro métodos estatísticos detalhados abaixo:

- O teste de Wilcoxon-Mann-Whitney é utilizado para encontrar os microRNAs estatisticamente significativos nas amostras em relação a classificação do tecido.
- A regressão logística é utilizada com o objetivo de encontrar microRNAs estatisticamente significativos e avaliá-los por meio das estimativas do método.
- A análise de sobrevivência dos grupos de microRNAs é utilizada para verificar se os microRNAs superexpressos ou subexpressos apresentam diferença estatisticamente significativa, utilizando o teste de log-rank.
- A árvore de classificação é construída com a intenção de encontrar microRNAs que relacionem a classificação do tecido histologicamente em normal ou tumoral e a árvore de sobrevivência é construída com o objetivo de encontrar os microRNAs estatisticamente significativos relacionados a sobrevivência dos pacientes com câncer.

Após encontrar os microRNAs estatisticamente significativos, procura-se entender o que estes regulam por meio de vias moleculares utilizando ferramentas de bioinformática. No Capítulo 3 abordam-se materiais e métodos estatísticos utilizados na análise de dados. No Capítulo 4 apresentam-se os resultados encontrados. As conclusões e considerações finais são abordadas no Capítulo 5.

3 TESTES E MODELOS ESTATÍSTICOS

Os tópicos deste capítulo apresentam os testes e os métodos estatísticos que foram utilizadas nesse trabalho para analisar e interpretar os microRNAs do banco de dados.

3.1 Teste de Wilcoxon-Mann-Whitney

O teste de Wilcoxon-Mann-Whitney ou teste Mann-Whitney foi inicialmente apresentado por Wilcoxon (1945), para duas amostras independentes de tamanhos iguais. Posteriormente, Mann & Whitney (1947) generalizaram a técnica para amostras de tamanhos diferentes.

O teste muitas vezes é considerado uma versão não paramétrica do teste t de Student, é utilizado quando se tem um par de amostras independentes com o intuito de testar se as populações destas duas amostras podem ser consideradas semelhantes ou não, por meio da análise da igualdade das medianas dos grupos.

O teste de Mann-Whitney é fundamentado nos valores obtidos dos postos por meio da combinação das duas amostras, que é obtido pela ordenação dos valores no formato crescente. Os valores que são calculados pelo teste possuem a intenção de avaliar o quão entrelaçado o par de amostras está pós ordenação.

Segundo Siegel & Castellan Jr (1975) o método para realização do cálculo para o teste é:

- Determina-se os escores dos postos do grupo 1 (n_1) e no grupo 2 (n_2);
- Coloca-se em conjunto os escores dos grupos, em que o escore que for menor é atribuído o posto 1. Os valores dos postos podem variar de 1 a N em que $N =$

$n_1 + n_2$. Em caso de empate é atribuído a média dos postos correspondentes;

- Determina-se, $U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$, R_1 é a soma dos postos do grupo 1 (n_1).

As hipóteses estatísticas a serem testadas são:

- H_0 : não existe diferença nas medianas dos grupos (tecidos histologicamente normal e tumoral);
- H_1 : existe diferença estatisticamente significativa nas medianas dos grupos, considerando o nível de significância do teste $\alpha = 5\%$.

3.2 Regressão logística

Dias Filho et al. (2007) relatam em seu trabalho que o método da regressão logística apareceu inicialmente em 1960, consolidou-se para a previsão de dados e classificação, quando a variável reposta dispõe de duas categorias, conhecidas como binárias ou dicotômicas.

A técnica de regressão logística é um caso especial de modelos lineares generalizados (MLG). Ressalta-se que os MLG são compostos por três componentes: uma componente aleatória, uma componente sistemática e uma função de ligação. As funções de ligação mais conhecidas são: *logit*, o complemento *log-log* e a *probit*, segundo Figueira (2006).

Hosmer Jr et al. (2013) relatam que a regressão logística é flexível, fácil de interpretar e de aplicar, tendo assim, bons motivos para a sua utilização, tais como: facilidade de lidar com variáveis independentes categorizadas ou contínuas, os resultados podem ser expressos como probabilidades, facilidade para classificar os indivíduos em categorias, demanda um número pequeno de suposições e possui o alto grau de confiabilidade.

Assim, a regressão logística estima a probabilidade relacionada ao acontecimento de um determinado evento de interesse, dado um conjunto de variáveis exploratórias. O modelo de regressão logístico é dado desta forma:

$$\ln \left[\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})} \right] = g(\mathbf{X}),$$

em que

$$g(\mathbf{X}) = \beta' \mathbf{X},$$

ou equivalentemente,

$$\pi(\mathbf{X}) = \frac{e^{\beta' \mathbf{X}}}{1 + e^{\beta' \mathbf{X}}}.$$

É fundamental realizar a transformação *logit* em $\pi(x)$ para o estudo da regressão logística. Define-se:

$$g(\mathbf{X}) = \ln \left[\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})} \right] = \beta' \mathbf{X}. \quad (1)$$

Caso, $\beta < 0$ quando x for grande a $P(Y = 1)$, probabilidade do tecido ser tumoral, para o contexto do trabalho, é pequena. Porém, se $\beta > 0$ quando x é grande, a $P(Y = 1)$ é grande. O *logit*, $g(\mathbf{X})$, tem propriedades interessantes já que possui seus parâmetros lineares, pode ser contínuo e dependendo do valor de x , varia de $-\infty$ a $+\infty$.

A estimação dos parâmetros do modelo é realizada pelo método de máxima verossimilhança, que se destaca por obter valores para os parâmetros que não são conhecidos. No trabalho de Hosmer Jr et al. (2013) é mostrado o desenvolvimento para encontrar os parâmetros de um modelo de regressão logística.

A Expressão 1 mostra o modelo para a modelagem univariada, porem uma boa modelagem é aquela que consegue modelar várias variáveis. Para a regressão logística múltipla com p variáveis o modelo é:

$$g(\mathbf{X}) = \ln \left[\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})} \right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p,$$

e equivalentemente,

$$\text{logit } \pi(\mathbf{X}) = \ln \left[\frac{P(Y = 1)}{1 - P(Y = 1)} \right] = \ln \left[\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})} \right] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = \beta' \mathbf{X}_p$$

desta forma o modelo de regressão logística é:

$$\pi(\mathbf{X}) = \frac{e^{g(\mathbf{X})}}{1 + e^{g(\mathbf{X})}}.$$

Assim, quando

- $\beta' \mathbf{X}_p \rightarrow +\infty$, então $P(Y = 1) \rightarrow 1$;
- $\beta' \mathbf{X}_p \rightarrow -\infty$, então $P(Y = 1) \rightarrow 0$.

No contexto do trabalho, $\pi(\mathbf{X})$ é a probabilidade do tecido ser tumoral ($Y = 1$), em que, $(\mathbf{Y})$ é a variável dependente, os vetores são os coeficientes da regressão logística, e $\mathbf{X}_p$ são as variáveis independentes associadas a p-ésima célula.

A primeira suposição para usar o modelo de regressão logística é a linearidade entre a razão de probabilidade de ocorrência e não ocorrência do evento de interesse, dada pela equação:

$$\frac{P(Y = 1)}{P(Y = 0)} = e^{\beta_0 + \beta_1 \mathbf{X}_1 + \dots + \beta_p \mathbf{X}_p},$$

e, por consequência,

$$\ln \left[\frac{P(Y = 1)}{P(Y = 0)} \right] = \beta_0 + \beta_1 \mathbf{X}_1 + \dots + \beta_p \mathbf{X}_p.$$

Por isso, para interpretar os coeficientes da regressão logística, usa-se e^β e não diretamente β . Entretanto, ao utilizar o modelo logístico para a distinção entre grupos, não se possui muito interesse na interpretação dos coeficientes (Garson, 2004).

A regra para a utilização do modelo de regressão logístico na distinção de dois grupos, que é o caso deste trabalho, é:

- se $P(Y = 1) > 0,5$, classifica-se $Y = 1$ (tecido histologicamente tumoral);
- caso contrário, classifica-se $Y = 0$ (tecido histologicamente normal).

Hosmer Jr et al. (2013) realizaram um trabalho a respeito das principais vantagens, desvantagens e características de vários testes utilizados para testar a bondade e a homogeneidade do ajuste do modelo de regressão logística. Neste trabalho será utilizada para avaliar o ajuste do modelo o teste qui-quadrado de Pearson.

3.2.1 Teste qui-quadrado de Pearson

O teste qui-quadrado de Pearson (χ^2) foi explorado inicialmente por Pearson (1900), é um teste de hipótese utilizado para dados categóricos com o intuito de avaliar a associação que há entre variáveis qualitativas.

O teste de χ^2 é usado para averiguar se a frequência de um determinado evento de interesse na amostra desviam-se com significância ou não da frequência esperada e compara a distribuição de vários eventos em distintas amostras, para julgar se as proporções observadas apresentam ou não diferenças significativas, ou então, se as amostras mostram diferenças significativas em relação às proporções deste acontecimento (Correa et al., 2006).

Os testes qui-quadrado podem ser utilizados para avaliar a associação entre duas variáveis categóricas (teste de independência), para verificar se as proporções dos atributos observados em dois grupos ou mais não apresentam diferenças estatisticamente significativas (teste de homogeneidade) e também podem ser aplicados para avaliar a adequação do ajuste de um determinado modelo estatístico a um conjunto de dados (teste de bondade de ajuste).

As condições necessárias para aplicar o teste segundo Correa et al. (2006) são que: as amostras precisam ser independentes, as observações dessas amostras são selecionadas aleatoriamente, as observações precisam ser frequências ou contagens, cada observação deve pertencer a uma categoria e a amostra deve ser relativamente grande (no mínimo 10 observações se tiverem apenas duas categorias).

Para o cálculo do teste, tem-se os seguintes passos:

1) Pearson (1900) apresentou a formula para calcular a estatística do teste:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i},$$

em que, χ^2 é a estatística de teste cumulativa de Pearson que assintoticamente se aproxima da distribuição qui-quadrado, O_i é o número de observações da categoria i e E_i é a frequência esperada da categoria i , que é calculada como o número total

da amostra multiplicado pela proporção observada na categoria da variável.

2) Após o calcular a estatística do teste, determina-se os graus de liberdade (gl) dessa estatística, da seguinte maneira, para:

- **Teste de qualidade de ajuste (bondade de ajuste):** os gl são o número de categorias menos o número de parâmetros da distribuição ajustada.
- **Teste de homogeneidade:** os gl são dados da seguinte forma $gl = (\text{número de categorias} - 1) \times (\text{número de grupos independentes} - 1)$.

3) Determina-se o valor do nível de significância (α) desejado.

4) É realizada a comparação entre a estatística calculada χ^2 em relação ao valor crítico da distribuição qui-quadrado com gl e o nível de significância, em que, normalmente há uma boa aproximação da distribuição χ^2 .

5) Verifica-se a hipótese nula, de que as frequências observadas são as mesmas da teórica, dependendo do teste utilizado.

As hipóteses a serem testadas para o teste de qualidade de ajuste:

- H_0 : O modelo estatístico está bem ajustado ao conjunto de dados.
- H_1 : O modelo estatístico não está bem ajustado ao conjunto de dados.

As hipóteses a serem testadas para o teste de homogeneidade:

- H_0 : Não existe diferença nas proporções das categorias observadas nos grupos avaliados.
- H_1 : Existe diferença nas proporções das categorias observadas nos grupos avaliados.

Para avaliar a qualidade na discriminação dos tecidos histologicamente normais e tumorais na regressão logística será utilizado a estatística C, conforme apresentado a seguir.

3.2.2 Estatística C

A estatística C é também conhecida como estatística de concordância ou índice C, sendo utilizada no contexto da regressão logística com resultados dicotômicos como uma medida de discriminação (Young, 1941).

O princípio desta estatística é que um sujeito que experimenta um determinado resultado tem uma probabilidade prevista mais alta do que um sujeito que não o experimenta. Assim, o índice C pode ser calculado como a proporção de pares de indivíduos cujos resultados observados e previstos concordam entre todos os pares possíveis em que um indivíduo possui o resultado de interesse e o outro indivíduo, não (Guo et al., 2018).

Quanto maior o valor da estatística C, melhor é modelo para discriminação, sendo maior o poder de discriminação do modelo estatístico utilizado no conjunto de dados. No contexto da análise de sobrevivência, várias estatísticas C foram formuladas para dados censurados à direita (Harrell Jr, 1986).

A estatística de concordância segundo Young (1941), é dada pela fórmula:

$$C = 1 - \frac{\sum_1^{n-1} (X_i - X_{i+1})^2}{2 \sum_1^n (X_i - \bar{X})^2},$$

em que, X_i são elementos em estudo.

3.3 Análise de sobrevivência

A análise de sobrevivência foi uma área que recebeu um grande destaque no fim do século passado devido ao avanço de técnicas de estatística e a evolução das tecnologias dos computadores, principalmente na área da saúde.

Na análise de sobrevivência a variável resposta em estudo, normalmente, é representada pelo tempo até a ocorrência de determinado evento de interesse, denominado tempo de falha. Esta variável não é negativa e a resposta é por natureza uma medida do tempo de seguimento. Há muitos exemplos no qual a análise de sobrevivência pode ser aplicada, como na área da saúde, a morte de um paciente, a

cura ou o reaparecimento de uma doença e até mesmo a falha de um determinado equipamento.

A principal característica dos dados em análise de sobrevivência é a presença de censura no conjunto de dados analisado. A censura é definida como a observação parcial da resposta. Colosimo & Giolo (2006) apresentaram em seu livro a importância de analisar dados censurados pois, mesmo que as informações dos dados censurados sejam incompletas, ainda sim, fornecem informações relevantes sobre o tempo de vida do paciente e a omissão dos dados censurados podem acarretar em conclusões viciadas no cálculo das estatísticas.

Seja t_i , para $i = 1, \dots, n$ o tempo de falha do i -ésimo indivíduo que admite o menor valor entre t_i^* , tempo de sobrevivência até a falha e u_i , tempo de falha, isto é,

$$t_i = \min(t_i^*, u_i), \text{ para } i = 1, \dots, n.$$

A variável que indica a censura, falha, é determinada por,

$$\delta_i = \begin{cases} 0, & \text{se } t_i \text{ é um tempo de censura;} \\ 1, & \text{se } t_i \text{ é tempo de falha.} \end{cases}$$

As principais classificações para os tipos de censuras, segundo Colosimo & Giolo (2006), são: a censura à direita, a censura à esquerda e a censura intervalar. A censura à direita caracteriza-se pelo tempo do acontecimento do evento estar à direita do tempo registrado. Ela pode ser classificada em:

- Tipo I: ao final do estudo alguns indivíduos ainda não falharam;
- Tipo II: Após um número pré estabelecido de falhas ocorrer o estudo é finalizado;
- Aleatória: Por motivos externos, alguns indivíduos foram retirados do estudo e ainda não tinha ocorrido a falha.

A censura à esquerda acontece se o tempo registrado for maior que o tempo da falha, ou seja, o evento de interesse já tinha acontecido quando o indivíduo foi

observado. E a censura intervalar acontece quando os indivíduos são acompanhados com consultas frequentes e sabe-se que o evento de interesse aconteceu em determinado intervalo de tempo.

3.3.1 Funções associadas ao tempo de sobrevivência

Considere T como o tempo até a ocorrência de um determinado evento, assim, T é uma variável aleatória que não é negativa e contínua, dado que a distribuição pode ser definida pelas funções de densidade de probabilidade, sobrevivência ou a função risco. Definidas a seguir.

- Função Densidade de Probabilidade (f.d.p.)

O limite da probabilidade de um indivíduo falhar em um determinado intervalo de tempo $(t, t + \Delta t)$ é dado pela f.d.p. abaixo:

$$\begin{aligned} f(t) &= \lim_{\Delta t \rightarrow 0} \frac{P(\text{um indivíduo falhar no intervalo } (t, t + \Delta t))}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{P(T \in (t, t + \Delta t))}{\Delta t} \end{aligned}$$

tal que $f(t) \geq 0$ para todo $t \geq 0$.

Adicionalmente, a função de distribuição acumulada:

$$F(t) = P(T < t) = \int_0^t f(u) du, \text{ para } t > 0.$$

- Função de Sobrevivência

A probabilidade de um evento de interesse acontecer além do tempo t é dado pela função de sobrevivência a seguir:

$$S(t) = P(T \geq t) = 1 - P(T \leq t) = 1 - F(t). \quad (2)$$

Neste caso, $F(t)$ representa a probabilidade de morte até o tempo t . Satisfazendo às seguintes propriedades:

- $S(0) = 1$;
- $\lim_{t \rightarrow \infty} S(t) = 0$;
- $S(t)$ é decrescente.

- Função de Risco

Para indicar a probabilidade de um indivíduo falhar em um curto intervalo de tempo, tem-se a função de risco, dada por:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (3)$$

A função de risco acumulada, que pode ser usada para estimações não paramétricas, fornece a taxa de falha acumulada do indivíduo até certo momento, como abaixo:

$$\Lambda(t) = \int_0^t h(u) du.$$

3.3.2 Relações entre as funções

A partir das definições anteriores, é possível obter-se relações que poderão tornar-se úteis, tais como:

Das funções (2) e (3), tem-se que:

$$\lambda(t) = \frac{f(t)}{S(t)}, \quad (4)$$

utilizando a densidade e a distribuição, tem-se:

$$f(t) = -S'(t). \quad (5)$$

Manipulando (5) em (4), observa-se:

$$\lambda(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log(S(t)),$$

solucionando a equação diferencial tem-se,

$$S(t) = \exp \left(- \int_0^t \lambda(u) du \right). \quad (6)$$

De (6) e (4),

$$f(t) = \lambda(t) \exp \left(- \int_0^t \lambda(x) dx \right).$$

3.3.3 Estimador de Kaplan Meier

O estimador de Kaplan e Meier (EKM) é uma das técnicas mais conhecidas e utilizadas para estimação da curva de sobrevivência na presença de censura, foi proposto por Kaplan & Meier (1958). Há, também, outros estimadores conhecidos como o estimador de Nelson-Aalen e a tabela de vida ou aturial.

O EKM é estabelecido como:

$$\hat{S}(t) = \prod_{j:t_j < t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j < t} \left(1 - \frac{d_j}{n_j} \right)$$

em que o $\hat{S}(t)$ simboliza uma função discreta com saltos, $t_1 < t_2 \dots < t_k$, são os k tempos distintos e ordenados de falhas, o d_j é número de falhas em t_j , $j = 1, \dots, k$ e n_j o número de indivíduos sob risco em t_j .

Para Colosimo & Giolo (2006) há vantagens na utilização do EKM, como: é um estimador de máxima verossimilhança de $S(t)$, não é viciado para amostras consideradas grandes, converge assintoticamente para um processo gaussiano e é fracamente consistente.

É preciso avaliar a exatidão do EKM para realizar a construção de intervalos de confiança e testar as hipóteses. A variância assintótica do estimador, obtida por meio das propriedades de máxima verossimilhança, é estimada pela fórmula de Greenwood, expressa por:

$$\widehat{Var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:t_j < t} \left(\frac{d_j}{n_j (n_j - d_j)} \right).$$

Dado $\hat{S}(t)$, com t fixo, tem a distribuição assintoticamente Normal, o intervalo aproximado de $100(1 - \alpha)\%$ de confiança para $S(t)$ é apresentado por:

$$\hat{S}(t) \pm z_{\alpha/2} \sqrt{\widehat{Var}(\hat{S}(t))}, \quad (7)$$

no qual, $z_{\alpha/2}$ é o percentil $(1 - \alpha/2)100\%$ da distribuição normal padrão.

Porém, no caso de valores extremos para t o intervalo de confiança dado pela Expressão 7, pode apresentar o limite superior maior que 1 ou o inferior negativo. Para resolver este problema tem-se a opção de realizar a transformação logística, sugerida por Kalbfleisch & Prentice (2011), desta forma a variância assintótica é representada por:

$$\widehat{\text{Var}}(\hat{U}(t)) = \frac{\sum_{j:t_j < t} \frac{d_j}{n_j(n_j - d_j)}}{[\log \hat{S}(t)]^2}.$$

O intervalo aproximado de confiança de $100(1 - \alpha)\%$ para $S(t)$, é dada por:

$$[\hat{S}(t)]^{\exp\{\pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{U}(t))}\}}. \quad (8)$$

Para o EKM, os gráficos das curvas de sobrevivência podem ser feitos para verificar visualmente se existe desigualdade entre as categorias. Porém, deve-se utilizar testes de hipóteses para verificar se as diferenças encontradas são realmente estatisticamente significativas ou não.

Entre todos os testes não paramétricos, um dos mais utilizados é o teste de log-rank (Mantel, 1966). O teste é adequado quando a razão das funções de taxa de falha dos grupos a serem confrontados é constante (Colosimo & Giolo, 2006). O teste de log-rank coloca o mesmo peso ao número de indivíduos para qualquer que seja o tempo de seguimento.

3.3.4 Teste de log-rank

O teste de log-rank (Mantel, 1966) é mais utilizado na comparação de sobrevivência para dados censurados. É mais apropriado quando os grupos comparados possuem propriedades de taxas de falhas proporcionais, caracteriza-se pela diferença entre o número observado de falhas em cada grupo e o número esperado de falhas sob a hipótese nula de igualdade das curvas de sobrevivência (Argenton et al., 2013).

Primeiramente, no livro de Colosimo & Giolo (2006), tem-se o teste para a

igualdade de duas funções de sobrevivência dado por $S_1(t)$ e $S_2(t)$. Os tempos de falhas $t_1 < t_2 < \dots < t_k$ são diferentes da amostra formada por meio da combinação de outras duas amostras diferentes. Supondo que no tempo t_j ocorra d_j falhas, que indivíduos estejam sob risco n_j em um tempo menor a t_j na amostra combinada. Nessa forma, o tempo de falha t_j poderá ser representada por meio de uma tabela de contingência 2x2 com d_{ij} falhas e $n_{ij} - d_{ij}$ sobreviventes na coluna i como mostrado na Tabela 1.

Tabela 1: Tabela de contingência para o tempo t_j .

	Grupo 1	Grupo 2	Totais
Falha	d_{1j}	d_{2j}	d_j
Não Falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Totais	n_{1j}	n_{2j}	n_j

A distribuição D_{2j} é uma hipergeométrica, relativo à falha e censura até o tempo t_j fixando-se as marginais da coluna e ao número de falhas no tempo t_j fixando-se as marginais de linha, tem-se:

$$P(D_{2j} = d_{2j}) = \frac{\binom{n_{1j}}{d_{1j}} \binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}}.$$

A média de D_{2j} é $w_{2j} = \frac{n_{2j}d_j}{n_j}$ e a variância de D_{2j} pela distribuição da hipergeométrica é:

$$(V_j)_2 = \frac{n_{2j}(n_j - n_{2j})d_j(n_j - d_j)}{n_j^2(n_j - 1)}.$$

A estatística $D_{2j} - w_{2j}$ obtém a média com valor zero e variância $(V_j)_2$. Caso as tabelas de contingência k , forem independentes um teste adequado para equivalência das duas funções de sobrevivência pode ter como base:

$$T_L = \frac{\left[\sum_{j=1}^k (d_{2j} - w_{2j}) \right]^2}{\sum_{j=1}^k (V_j)_2}.$$

As hipóteses a serem testadas são:

- H_0 : As curvas de sobrevivência dos grupos não apresentam diferença nos tempos.
- H_1 : As curvas de sobrevivência dos grupos apresentam diferença estatisticamente significativa, ao nível de significância de $(\alpha = 5\%)$.

Considerando apenas dois grupos para qualquer t no espaço de acompanhamento, obtém uma distribuição aproximada qui-quadrado com um grau de liberdade. Rejeita-se H_0 , se valor $p \leq \alpha$.

3.4 Construção de árvores de decisão

A árvore de decisão é um método de aprendizado de máquina para a construção de modelos de previsão de dados, caracteriza-se por ser muito eficiente, não paramétrico e amplamente usada em problemas de regressão quando a variável dependente é quantitativa ou para classificação quando é qualitativa.

Para construir uma árvore de decisão, os principais componentes são: definição das possíveis divisões do espaço preditor para cada variável em relação a um nó, a seleção da melhor divisão de todas, decisão de quando considerar o nó como final ou não e para cada nó final atribuir um valor resposta.

Entretanto, a principal desvantagem no uso das árvores de decisão é a instabilidade, pois poucas modificações na amostra de aprendizagem podem levar a grandes mudanças na árvore, como por exemplo, o aumento da complexidade ou a mudança nos valores escolhidos para as divisões (Argenton et al., 2013).

O método para a construção de uma árvore de decisão binária, em geral, é da seguinte forma: apresenta-se o conjunto de dados ao nó pai (ou inicial), é realizado

um teste lógico para o nó ramificar-se em dois nós filhos que serão chamados posteriormente de intermediários (nó interno ou nó decisório), o nó interno irá dividir-se em dois nós descendentes, um esquerdo e um direito, no caso em que os dados satisfazem o teste lógico do nó interno seguem para o nó da esquerda e os dados que não satisfazem seguem para o nó da direita e o procedimento será repetido até o nó final (ou folha) ser alcançado. A recursividade de uma árvore é definida pela repetição desse processo, posteriormente, se necessário pode-se utilizar técnicas de poda (Ciampi et al., 2000). A Figura 2 ilustra uma árvore de decisão.

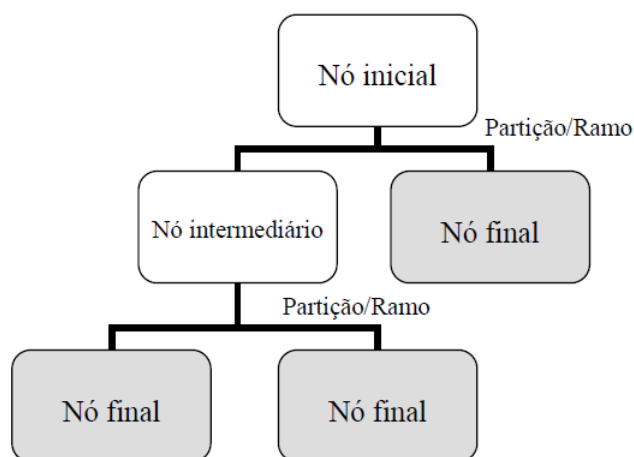


Figura 2: Ilustração de uma árvore de decisão.

A árvore de decisão pode ser de classificação ou de regressão e estas possuem semelhanças, como: ambas dividem o espaço preditor (variáveis independentes) em diferentes regiões sobrepostas e seguem uma abordagem “*top-down*” e “gananciosa” conhecida como divisão binária recursiva (do inglês, *recursive binary splitting*). *Top* porque inicia no topo da árvore quando todas as observações estão disponíveis em uma única região e *down* porque divide o espaço preditor em dois novos ramos no sentido inferior da árvore. Gananciosa devido ao algoritmo se preocupar apenas com a divisão atual (procura a melhor variável disponível) e não divisões futuras, que poderiam levar a uma árvore melhor.

3.4.1 O método CART

Existem diferentes algoritmos para a construção de árvores de decisão, todavia será usado neste trabalho o algoritmo CART (classification and regression trees) que é o mais popular e eficaz neste âmbito, foi proposto em Breiman et al. (1984).

Uma vantagem importante do CART é sua robustez a observações divergentes (*outliers*). O algoritmo de divisão isola observações divergentes em um nó individual, que provavelmente será podado posteriormente. E se caracteriza por definir todas as regras para divisão dos nós da árvore, decidir quando a árvore está completa e relacionar cada a folha a uma classe ou a um valor preditivo.

Uma outra propriedade interessante de metodologia CART é que a estrutura de suas árvores de classificação e regressão pode ser substituída pelo seu logaritmo ou raiz quadrada que a estrutura da árvore não será modificada, isso para qualquer variável (Argenton et al., 2013).

O algoritmo CART permite que a árvore cresça até o tamanho máximo, ou seja, a árvore máxima. Posteriormente, começa-se o processo de poda de alguns ramos com o intuito de aumentar o poder da generalização, a poda da árvore é realizada por meio da medida de custo-complexidade.

A medida de custo-complexidade controla o tamanho ou complexidade da árvore, e como a árvore se ajusta aos dados. Se o parâmetro de complexidade é grande, a árvore que minimiza o custo-complexidade é pequena, e conforme decresce, minimiza o custo complexidade aumenta (Sobral, 2003).

As árvores podem ser divididas em dois tipos: árvore de classificação para variáveis qualitativas e árvore de regressão para variáveis quantitativas, conforme descritas nos próximos tópicos.

3.4.2 Árvore de classificação

Inicialmente, a árvore de classificação apareceu como uma técnica alternativa para a regressão logística múltipla. Na árvore cada folha tem rotulo que mostrará para o conjunto de dados a classe predita. O algoritmo CART possui em cada interação o

particionamento passo a passo, assim utiliza somente de uma variável, este método pode ser visualizado na Figura 3.

Na Figura 3, definiu-se que a regra de divisão para os nós da árvore seria uma desigualdade, o que normalmente acontece em todos os casos, o nó esquerdo é composto pelas observações que satisfazem essa igualdade e o restante das observações seguem para nó direito. Considera-se t_p um nó pai e t_d e t_e , respectivamente, nós filhos direito e esquerdo do t_p .

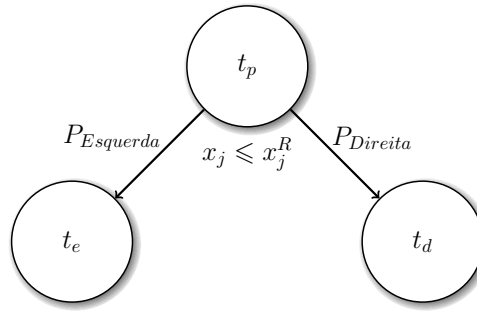


Figura 3: Algoritmo de divisão do CART.

Para a Figura 3 a divisão foi feita por meio da variável j no valor x_j^R , que se refere ao valor de referência para esta divisão, ou seja, o melhor valor para esta variável (Timofeev, 2004).

O nó t_p possui como objetivo principal gerar o t_d e t_e com máxima homogeneidade, ou seja, o intuito é ter dois filhos com homogeneidade e, desta forma, maximizar a diferença de impureza. Tem-se uma função para mediar a homogeneidade dos filhos, que é denominado de função de impureza $i(t)$. Como a impureza do t_p é constante para qualquer uma das divisões possíveis $x_j \leq x_j^R$, a homogeneidade máxima dos t_d e t_e será equivalente a maximização da função de mudança de impureza $\Delta i(t)$, segundo Timofeev (2004):

$$\Delta i(t) = i(t_p) - E[i(t_c)],$$

em que t_c são os t_d e t_e juntos. Suponha-se que P_e e P_d são as probabilidades dos nós esquerdos e direitos, respectivamente, pode-se obter essa fórmula usando-se as

probabilidades dos nós filhos. Assim,

$$\Delta i(t) = i(t_p) - P_e i(t_e) - P_d i(t_d).$$

Logo,

$$\arg \max_{x_j \leq x_j^R, j=1, \dots, M} [i(t_p) - P_e i(t_e) - P_d i(t_d)]. \quad (9)$$

Por meio da Expressão (9) percebe-se que o algoritmo CART buscará todos os valores possíveis das variáveis para encontrar a melhor “pergunta” para a divisão $x_j < x_j^r$ que maximizará $\Delta i(t)$. Assim, é necessário definir $i(t)$.

A função de impureza mais comum, que também é usada pelo algoritmo do Rpart no *software* R, é chamada regra de índice de Gini, definida pela função (Breiman et al., 1984):

$$i(t) = \sum_{k \neq l} p(k|t)p(l|t).$$

Se há K classes distintas e k e l indicam os índices das classes, $p(k|t)$ indica a probabilidade condicional da classe k desde que esteja no nó t . Substituindo esta fórmula na Expressão (9), obtém-se como resultado:

$$\arg \max_{x_j \leq x_j^R, j=1, \dots, M} \left[- \sum_{k=1}^K p^2(k|t_p) + P_e \sum_{k=1}^K p^2(k|t_e) + P_d \sum_{k=1}^K p^2(k|t_d) \right].$$

O algoritmo CART para calcular a função de impureza Gini procurará na amostra a maior classe e a isolará do restante dos dados.

3.4.3 Árvore de regressão

As árvores de regressão não possuem classes pré-atribuídas, ao contrário da árvore de classificação, assim existem vetores de respostas a Y que representa, os valores para cada observação da variável X . Então, como não existem classes o índice Gini não pode ser utilizado.

A regra de divisão para a árvore de regressão é encontrada por meio do algoritmo que minimiza os resíduos quadrados, acarretando que as variações esperadas da soma para os dois nós filhos devem ser minimizadas (Timofeev, 2004).

$$\arg \max_{x_j \leq x_j^R, j=1, \dots, p} [P_e \text{Var}(Y_e) + P_d \text{Var}(Y_d)] \quad (10)$$

o qual, $\text{Var}(Y_e)$ e $\text{Var}(Y_d)$ são os vetores de resposta para o nó filho esquerdo e direito respectivamente, $x_j \leq x_j^R$ é a referência em relação a divisão ideal que satisfaça a condição da Equação (10). Ressalta-se que na árvore de regressão os nós terminais são caracterizados por um valor médio e seu erro quadrado médio.

Existe uma relação entre o algoritmo de minimização dos resíduos quadrados e a regra de divisão de Gini, segundo Timofeev (2004) é a função de impureza do índice Gini por meio da notação de variâncias, caso atribua o valor um aos objetos da classe atual e zero ao resto dos objetos, então Y pode ser interpretado como uma variável aleatória que segue a lei de Bernoulli. Portanto, a variância da amostra desses valores seria igual para $P(k|t)[1 - P(k|t)]$. Resumindo pelo número de classes K , obtém-se a seguinte medida de impureza $i(t)$:

$$i(t) = 1 - \sum_{k=1}^K P^2(k|t).$$

Usando esse critério de divisão, pode-se aumentar a árvore de regressão, entretanto, para resolver esse problema utiliza-se regras de poda de árvores na implementação numérica. No próximo tópico, apresentam-se os métodos para poda de árvores grandes, tanto para regressão como para classificação.

3.4.4 Crescimento e poda

Para o algoritmo CART há duas observações que são importantes na partição da árvore: para cada nó intermediário como selecionar uma divisão ideal e como determinar se um nó é terminal. Entretanto, um obstáculo no algoritmo para selecionar as divisões é a confiabilidade nas estimativas dos erros usadas, desta forma, a precisão das estimativas dos erros está fortemente relacionada a qualidade da amostra analisada (Sobral, 2003).

Dado que o algoritmo divide os dados de treinamento original recursivamente, conforme acontecem as divisões, é avaliado amostras cada vez menores, isto

é, na medida que a árvore cresce tem-se menos confiabilidade nas estimativas de erro. Com o objetivo de minimizar esse problema e evitar um superajustamento dos dados com árvores extremamente complexas tem-se a técnica conhecida como poda.

O CART utiliza a pós-poda, que é o processo que a árvore cresce até o tamanho máximo, e depois métodos analisam o crescimento da árvore e a confiabilidade, estes são utilizados para identificar qual árvore podada que possui o tamanho ideal (Sobral, 2003).

Breiman et al. (1984) relatou alternativas para a seleção da árvore final, com base nas estimativas dos erros encontrados: ou opta pela árvore com menor erro estimado ou escolhe a menor árvore, em que, o erro estimado está dentro do intervalo: menor erro estimado $\pm$ erro padrão desta estimativa.

4 ANÁLISE DE DADOS E RESULTADOS

4.1 Análise exploratória

A seguir apresentam-se os resultados das análises estatísticas dos microRNAs. O banco de dados do carcinoma pulmonar de células não pequenas (LUAD) foi coletado no Atlas do Genoma do Câncer (TCGA, site: <https://portal.gdc.cancer.gov/>), está publicamente disponível e os dados foram baixados no dia 12/12/2018.

Os dados do estudo que se referem ao sequenciamento de microRNAs de alto desempenho (miRNA-Seq), é composto por 508 amostras histopatologicamente tumoral e 45 amostras histopatologicamente normal, sendo 9 variáveis clínicas e 709 microRNAs expressos nestas amostras.

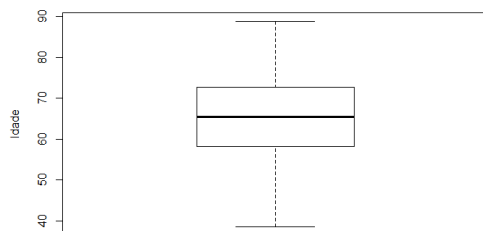
As variáveis clínicas histopatológicas e as medidas resumo para as variáveis qualitativas e quantitativas estão apresentadas na Tabela 2 e 3, respectivamente. Baseando-se nos resultados destas Tabelas, pode-se traçar o perfil dos indivíduos pertencentes a este conjunto de dados

Na Tabela 2, o tecido tumoral corresponde a 91,86% dos tecidos presentes no banco de dados, a variável sexo está dividida de forma semelhante, sendo 53,16% mulheres e 46,84% homens. Observou-se na amostra que 35,62% dos pacientes foram a óbito até o final do estudo e a grande maioria dos pacientes é de etnia branca (76,31%). Para a histologia todas as classes da Tabela 2 são tipos do adenocarcinoma pulmonar, o mais frequente foi a classificação somente como adenocarcinoma, sem especificações mais detalhadas, com 355 casos (64,20%). Em relação a localização do carcinoma, ou órgão de origem, o lobo superior apresenta 58,05% dos casos seguindo

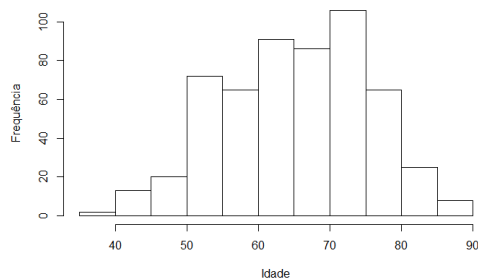
pelo lobo inferior com 33,82% dos casos. Para recordar a localização e anatomia do pulmão, tem-se a Figura 1 apresentada no Capítulo 1.

O Grupo A, composto por I, Ia/Ib e IIa/IIb descrito na Tabela 2, refere-se aos estágios iniciais do tumor, com representatividade de mais de 78%. O número elevado de amostras neste grupo ocorre porque são amostras coletadas a partir de tumores retirados cirurgicamente, apenas são submetidos a cirurgia pacientes que não estão em estágios avançados da doença. Para o Grupo B, composto pelo estagiamento IIIa/IIIb e IV descrito na Tabela 2, corresponde ao grupo que a doença se encontra em estágios mais elevados, representando 20,25%.

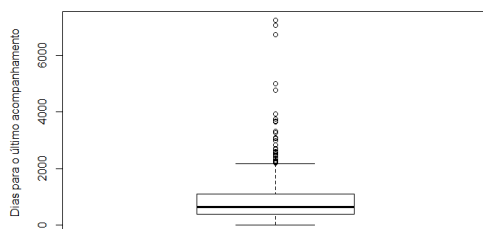
Na Tabela 3, a idade dos pacientes na amostra varia de 38,5 a 88,9 anos, sendo que metade tem mais que 65,5 anos e 25% mais do que 72,8 anos. No histograma da Figura 4(b), observa-se que a distribuição da idade é assimétrica e que a grande maioria dos pacientes estão concentrados na faixa etária de 50 a 80 anos.



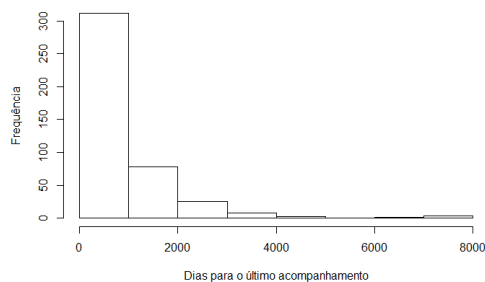
(a) Boxplot da idade.



(b) Distribuição de frequência da idade.



(c) Boxplot dos dias.



(d) Distribuição de frequência dos dias.

Figura 4: Distribuição da idade e dos dias até o último acompanhamento.

Tabela 2: Análises descritivas das variáveis qualitativas dos dados demográficos e características clínico-patológicas.

Variáveis	Descrições	Medidas resumo	
		n = 553	%
Tecido	Tumoral	508	91,86%
	Normal	45	8,10%
Sexo	Feminino	294	53,16%
	Masculino	259	46,84%
Estado Vital	Vivo=0	356	64,38%
	Morto =1 (censura)	197	35,62%
Etnia	Branca	422	76,31%
	Negro ou afro-americano	58	10,49%
	Asiático	8	1,45%
	Faltante	65	11,75%
Histologia	Adenocarcinoma	355	64,20%
	Adenocarcinoma com subtipos mistos	108	19,53%
	Adenocarcinoma papilar	23	4,16%
	Carcinoma de células acinares	22	3,98%
	Carcinoma bronquíolo-alveolar, não mucinoso	20	3,62%
	Adenocarcinoma mucinoso	14	2,53%
	Carcinoma broncoalveolar mucinoso	6	1,08%
	Adenocarcinoma bronquíolo-alveolar	3	0,54%
	Carcinoma Micropapilar	2	0,36%
Tecido ou órgão de origem	Lobo superior	321	58,05%
	Lobo inferior	187	33,82%
	Brônquio principal	24	4,34%
	Pulmão	15	2,71%
	Lesão de sobreposição de pulmão	4	0,72%
	Lóbulo médio	2	0,36%
Estágio do tumor	I, Ia/Ib, IIa/IIb (Grupo A)	434	78,48%
	IIIa/IIIb, IV (Grupo B)	112	20,25%
	Faltante	7	1,27%

Tabela 3: Análises descritivas das variáveis quantitativas dos dados demográficos e características clínico-patológicas.

Variáveis	Média	Desvio Padrão	Mínimo	Q ₁ *	Mediana	Q ₃ *	Máximo
Idade (anos)	65,32	9,99	38,50	58,20	65,50	72,80	88,90
Dias para último acompanhamento	884,42	968,11	1	377	630	1091	7248

*1º Quartil, ** 3º Quartil

Os dias até o último acompanhamento do paciente se referem ao número de dias em que o paciente foi observado até a sua última consulta ou acompanhamento. Pode-se observar na Figura 4(c) e 4(d) que há observações discrepantes e que a maioria dos pacientes tem tempo de acompanhamento inferior a 2000 dias (5,5 anos).

Nas Tabelas 4 e 5, apresentaram-se análise descritiva das variáveis qualitativas e quantitativas em relação ao tecido normal, tumoral e o valor p , foi realizado o teste qui-quadrado para as variáveis qualitativas e teste não-paramétrico de Wilcoxon para as variáveis quantitativas.

Tabela 4: Análises descritivas das variáveis qualitativas de acordo com o tecido e valor p .

		Tecido				Valor <i>p</i>
Variáveis	Descrição	Normal (n=45)		Tumoral (n=508)		
		n	%	n	%	
Sexo	Feminino	25	55,60	269	53,00	0,7373
	Masculino	20	44,40	239	47,00	
Estado Vital	Vivo	32	71,10	324	63,80	0,3249
	Óbito	13	28,90	184	36,20	
Etnia	Branco	39	86,70	383	75,40	0,0567
	Negro	6	13,30	52	10,20	
	Asiático	0	0,00	8	1,60	
	Faltante	0	0,00	65	12,80	
Histologia	Adenocarcinoma (AD)	37	82,22	308	61,11	0,2272
	AD com subtipos mistos	3	6,67	105	20,83	
	Adenocarcinoma Papilar	1	2,22	22	4,37	
	Carcinoma de células acinares	0	0,00	22	4,37	
	Carcinoma brônquio-alveolar, mucinoso	1	2,22	5	0,99	
	Adenocarcinoma mucinoso	1	2,22	13	2,58	
	Carcinoma sólido	0	0,00	6	1,19	
	Carcinoma bronquíolo-alveolar, não mucinoso	1	2,22	19	3,57	
	Adenocarcinoma bronquíolo-alveolar	0	0,00	3	0,60	
	Adenocarcinoma de células claras	1	2,22	2	0,40	
	Carcinoma micropapilar	0	0,00	2	0,39	
	Carcinoma de células de anel de sinete	0	0,00	1	0,20	
Tecido ou órgão de origem	Lobo superior	27	60,00	294	57,90	0,9324
	Lobo inferior	14	31,10	173	34,10	
	Lóbulo Médio	3	6,70	21	4,10	
	Pulmão	1	2,20	14	2,80	
	Lesão de sobreposição do pulmão	0	0,00	4	0,80	
	Brônquio principal	0	0,00	2	0,04	
Estágio do Tumor	I, Ia/Ib, II, IIa/IIb = Grupo A	39	86,67	394	77,56	0,3219
	IIIa/IIIb, IV = Grupo B	6	13,30	107	21,06	
	Faltante	0	0,00	7	1,38	

*Teste qui-quadrado.

Tabela 5: Análises descritivas das variáveis quantitativas de acordo com o tecido e valor p .

Variáveis	Descrição	Tecido		valor p^{**}
		Normal	Tumoral	
		Média (DP*)	Média (DP)	
Idade	Anos	66,13 (67)	65,25 (65,5)	0,6124
Dias do último acompanhamento	Dias	938,80 (712)	663,73 (624)	0,1559

*Desvio Padrão; **Teste não-paramétrico de Wilcoxon.

Na Tabela 4 a variável sexo apresentou maior número de amostras do sexo feminino tanto no tecido normal (55,6%) quanto no tumoral (53,0%), para a variável estado vital mais pessoas vieram a óbito no tecido tumoral (36,2%). A etnia apresentou valor $p = 0,0567$, sendo não estatisticamente significativa e seu maior grupo de amostras possuíram cor branca em ambos os tecidos, considerando nível de significância $\alpha = 0,05$. A origem do carcinoma, tanto para amostras do tecido normal, quanto para o tumoral, continuou sendo o lobo superior seguido do inferior. O estágio do tumor continuou concentrado no Grupo A, ou seja, nos estágios iniciais da doença independentemente do tipo de tecido.

Na Tabela 5, para as variáveis quantitativas a idade de concentração do tecido normal e tumoral assemelha-se (66,13 anos para tecido normal e 65,25 anos para tumoral), mas o dia do último acompanhamento para o tecido tumoral foi mais baixo (média = 663,73 dias) em relação ao tecido normal (média = 938,8 dias), entretanto não é significativo ao nível de 5% com valor $p = 0,1559$.

4.2 Teste de Wilcoxon-Mann-Whitney

Para os 709 microRNAs do banco de dados, realizou-se uma análise preliminar utilizando-se o teste de Wilcoxon-Mann-Whitney. O teste foi aplicado para verificar quais microRNAs que apresentavam diferença estatisticamente significativa

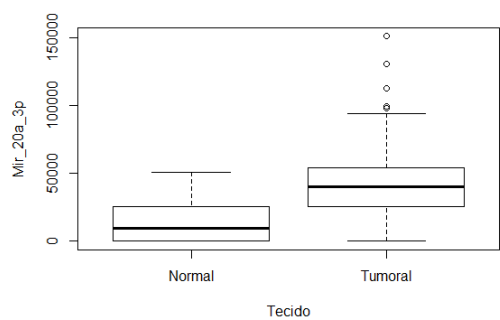
na mediana das amostras de tecidos normais e tumorais em relação a mediana do tecido normal. Os microRNAs estatisticamente significativos encontrados no teste estão apresentados na Tabela 6.

A Tabela 6 é composta pela expressão de 25 microRNAs e com o respectivo valor p . As setas $\uparrow$ ou $\downarrow$ encontradas junto de cada microRNA denotam-se a expressão do microRNA está aumentada ou diminuída, respectivamente, em relação aos valores observados para a mediana do tecido normal. As Figuras 5, 6 e 7 apresentam as distribuições de 18 microRNAs superexpressos, e as Figuras 8 e 9 as distribuições dos 7 microRNAs subexpressos que apresentaram diferenças estatisticamente significativas para os tecidos normais e tumorais.

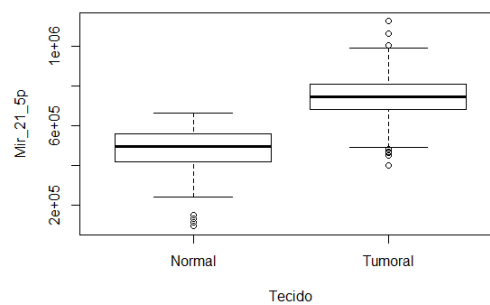
Tabela 6: Expressão dos microRNAs significativos por tecido normal ou tumoral e valor p .

MicroRNAs	Tecido Normal (n=45)					Tecido Tumoral (n=508)					valor p
	Mínimo	Q1	Mediana	Q3	Máximo	Mínimo	Q1	Mediana	Q3	Máximo	
miR-20a-3p $\uparrow^1$	0	2	9968	25624	50670	0	25626,8	40497,5	54335,25	151138	0,0001
miR-21-5p $\uparrow$	92566	417928	497239	561692	664743	398236	680936	746033,5	810729,5	1128308	0,0001
miR-155-3p $\uparrow$	0	0	0	0	9966	0	1	2	19598	54064	0,0001
miR-151b $\uparrow$	1	4780	9967	22575	30658	0	15659	25045	35011	59677	0,0001
miR-155-5p $\uparrow$	45564	112946	130779	164336	242194	61783	138481	182986	233942,25	390309	0,0001
miR-186-5p $\uparrow$	43749	118742	157599	170152	200102	44062	142579	175489	215734,75	398517	0,0001
miR-19b-1-5p $\uparrow$	0	4782	19598	29021	56281	2	17694	28166	40737,25	114979	0,0002
miR-219a-1-3p $\uparrow$	0	9967	19599	29277	54610	0	15659	26392	38234,5	75884	0,0014
miR-28-5p $\uparrow$	22191	67361	85087	101840	138715	30818	93462,3	116794	142592,5	242272	0,0001
miR-320d $\uparrow$	0	1	2	15659	49164	0	2	9968	29650,5	96984	0,002
miR-33b-5p $\uparrow$	0	0	1	3	61771	0	19933	38233	59742,5	123322	0,0001
miR-409-5p $\uparrow$	0	1	1	9966	22575	0	20868,8	37113	57382,75	135105	0,0001
miR-429 $\uparrow$	0	35992	55022	70363	94970	2	75999,3	101978	147416	360635	0,0001
miR-4661-5p $\uparrow$	1	9966	9968	29566	56165	0	22298	35013,5	48827,75	98198	0,0001
miR-629-3p $\uparrow$	0	1	0	19600	25625	0	25625,8	28197,5	57641,5	139129	0,0001
mir-616-3p $\uparrow$	0	0	0	1	35256	0	2	9968	29898	922227	0,0001
mir-769-3p $\uparrow$	0	2	9966	15659	29564	0	9967	25625	47282,5	142010	0,0001
miR-9-5p $\uparrow$	3	111992	173805	207598	393219	39218	206985	292923	424697	1054572	0,0001
miR-143-3p $\downarrow$	186991	554934	600705	634837	698155	207043	380602	437562,5	497407,5	768001	0,0001
miR-144-5p $\downarrow$	1	157043	225813	250486	359566	1347	65956,8	99701,5	137166,5	298277	0,0001
miR-30a-5p $\downarrow$	144739	329089	411687	467594	578663	101842	217575	293903,5	353723,25	562749	0,0001
miR-378a-3p $\downarrow$	13968	51644	63484	81195	106077	5970	42513,3	54495,5	69197	117292	0,0036
miR-3912-3p $\downarrow$	0	1348	22575	27148	48199	0	1	9967	19599	58919	0,0009
miR-486-5p $\downarrow$	38744	92435	130364	187251	254847	8290	42994	63110	84818,25	323843	0,0001
miR-490-3p $\downarrow$	38744	99660	130364	29353	254847	0	0	63110,5	1	323843	0,0001

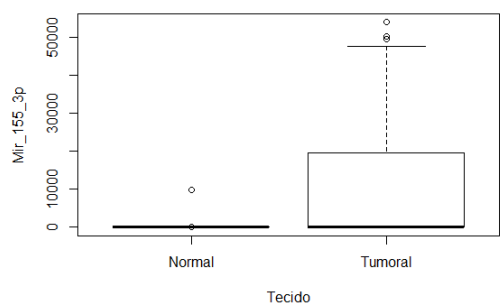
¹ $\uparrow$ e $\downarrow$ representam se o microRNA está ou superexpresso ou subexpresso respectivamente.



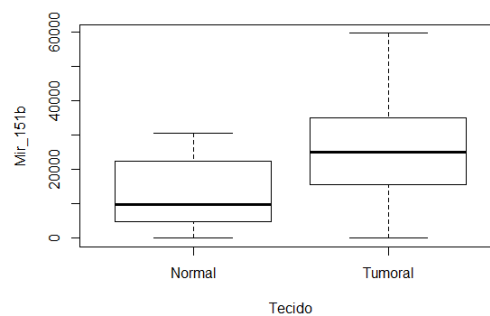
(a) miR-20a-3p.



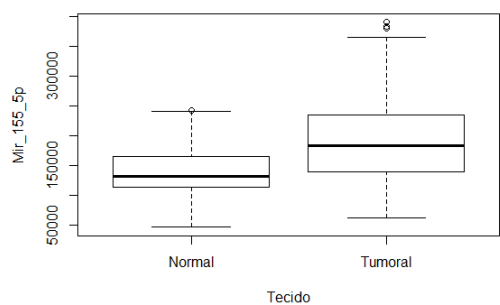
(b) miR-21-5p.



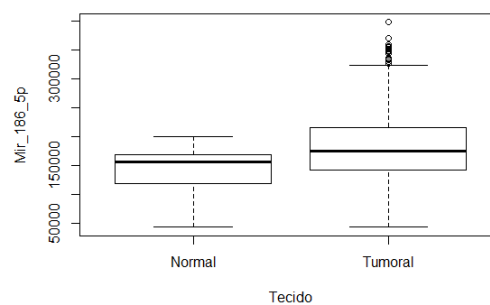
(c) miR-155-3p.



(d) miR-151b.

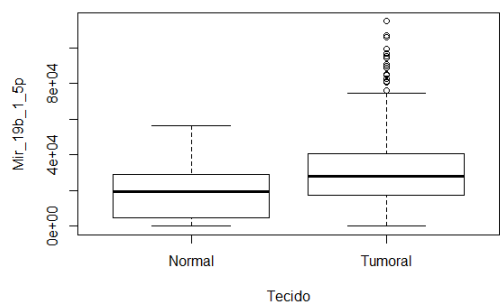


(e) miR-155-5p.

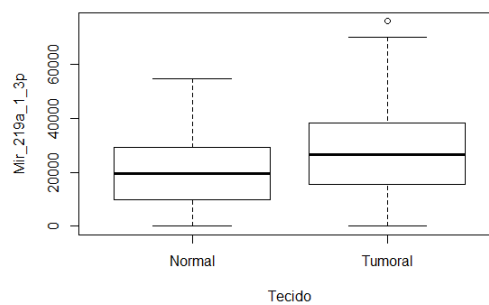


(f) miR-186-5p.

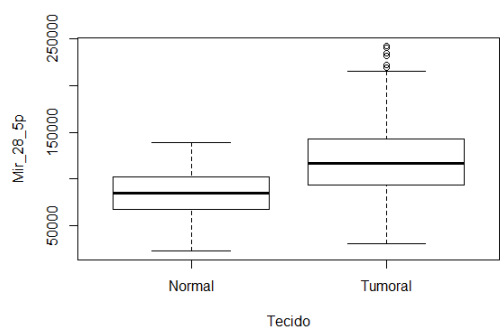
Figura 5: Boxplot da expressão dos microRNAs por tecido normal e tumoral para o teste de Wilcoxon-Mann-Whitney.



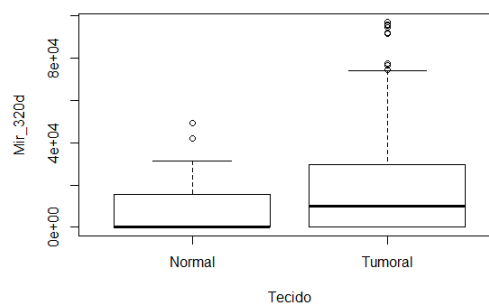
(a) miR-19b-1-5p.



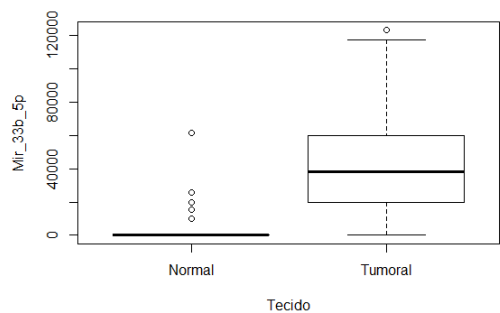
(b) miR-219a-1-3p.



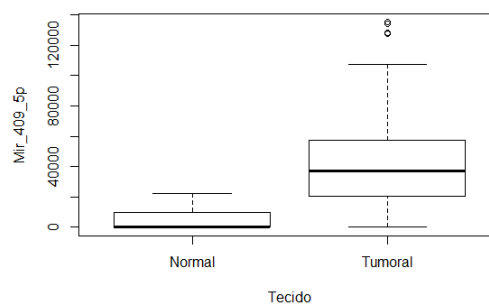
(c) mir-28-5p.



(d) miR-320d.

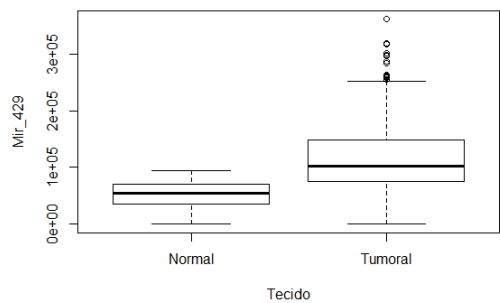


(e) miR-33b-5p.

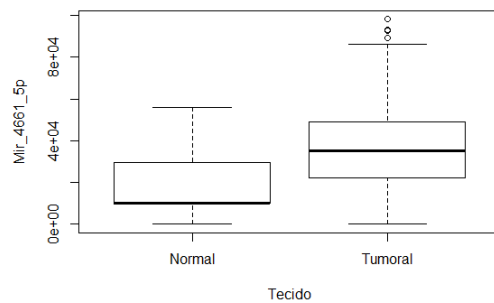


(f) miR-409-5p.

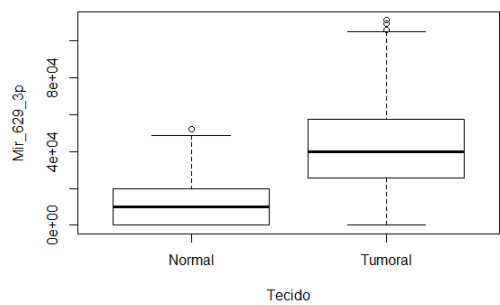
Figura 6: Boxplot da expressão dos microRNAs por tecido normal e tumoral para o teste de Wilcoxon-Mann-Whitney.



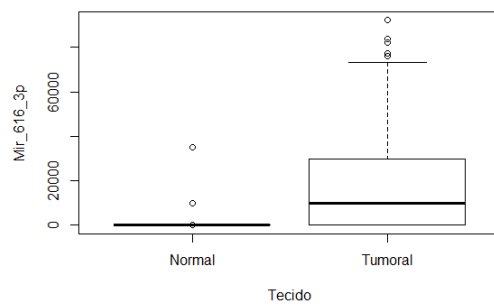
(a) miR-429.



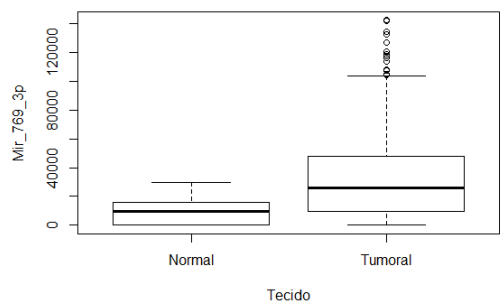
(b) miR-4661-5p.



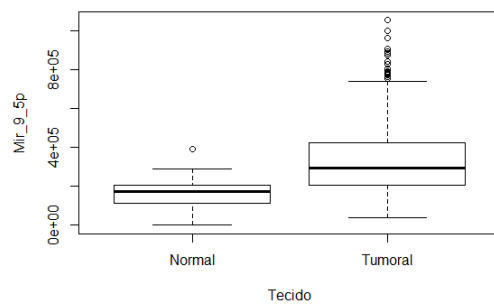
(c) miR-629-3p.



(d) miR-616-3p.

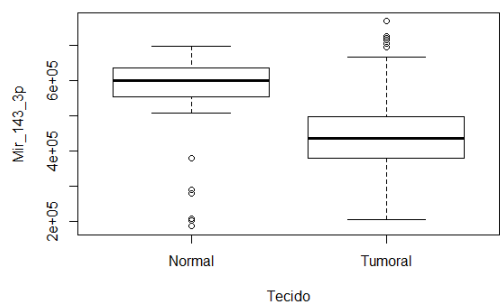


(e) miR-769-3p.

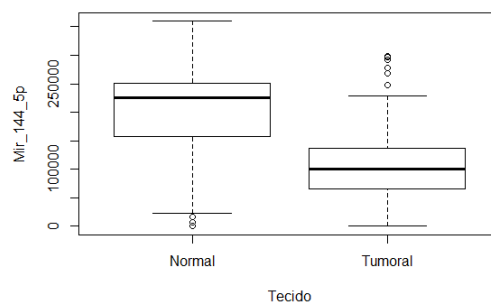


(f) miR-9-5p.

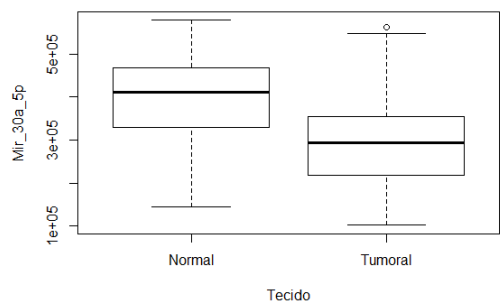
Figura 7: Boxplot da expressão dos microRNAs por tecido normal e tumoral para o teste de Wilcoxon-Mann-Whitney.



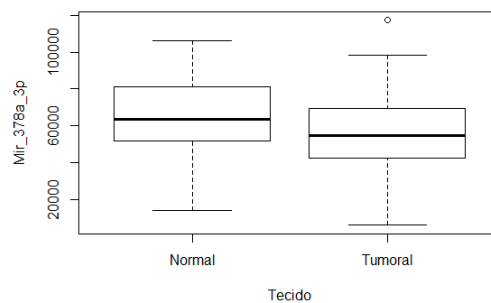
(a) miR-143-3p.



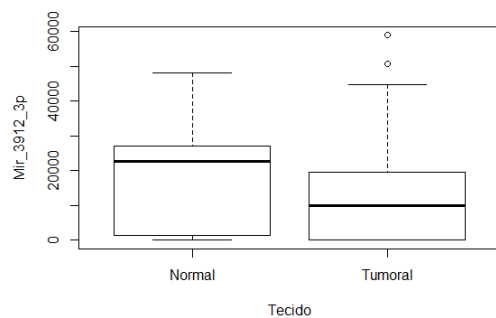
(b) miR-144-5p.



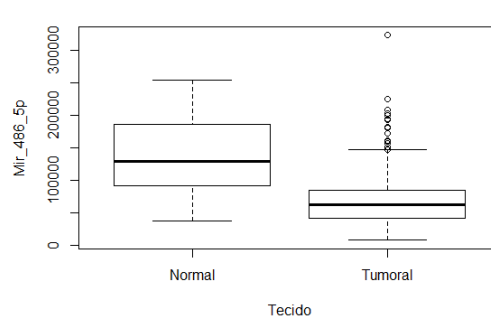
(c) miR-30a-5p.



(d) miR-378a-3p.

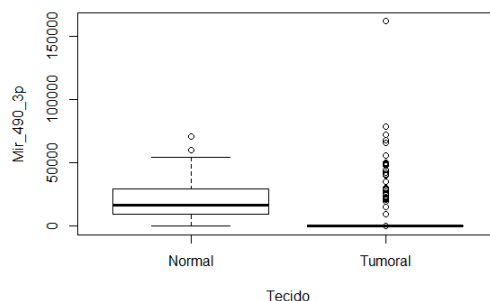


(e) miR-3912-3p.



(f) miR-486-5p.

Figura 8: Boxplot da expressão dos microRNAs por tecido normal e tumoral para o teste de Wilcoxon-Mann-Whitney.



(a) miR-490-3p.

Figura 9: Boxplot da expressão do microRNA por tecido normal e tumoral para o teste de Wilcoxon-Mann-Whitney.

4.3 Regressão logística

O modelo de regressão logística múltipla foi ajustado para os 709 microRNAs e as variáveis clínico-histopatológicas com o intuito de realizar o estudo do efeito das variáveis sobre a presença do câncer.

Encontraram-se 18 microRNAs, apresentados na Tabela 7, que explicam significativamente o aparecimento do tecido tumoral. Os microRNAs significativamente alterados não apresentaram relação com nenhuma variável clínico-histopatológico. Os parâmetros ajustados para o modelo de regressão logística estão apresentados na Tabela 7.

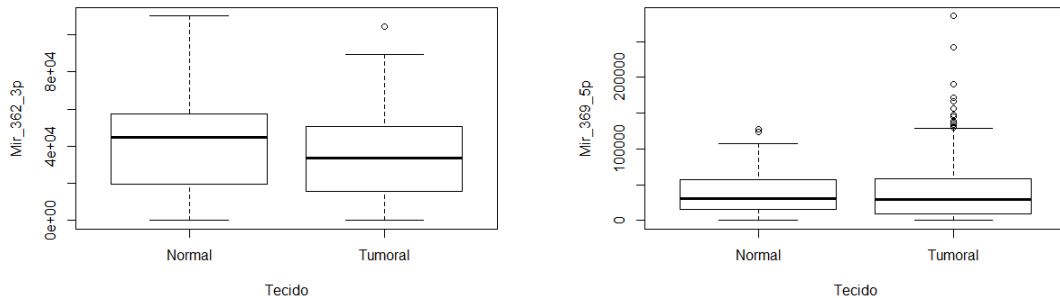
Pela Tabela 7, pode-se ver pelo sinal da estimativa do parâmetro do modelo de regressão logística que os microRNAs com valores positivos (miR-20a-3p, miR-21-5p, miR-28-5p, miR-320d, miR-33b-5p, miR-409-5p, miR-429, miR-493-5p, miR-577 e miR-629-3p) aumentam a chance do paciente apresentar câncer, para microRNAs com expressões elevadas. Por outro lado, o sinal negativo (miR-143-3p, miR-144-5p, miR-30a-5p, miR-362-3p, miR-369-5p, miR-486-5p, miR-490-3p e miR-509-3-5p) da estimativa mostra que expressões aumentadas para estes microRNAs podem diminuir as chances de apresentarem câncer pulmonar.

Tabela 7: Estimativas dos parâmetros do modelo de regressão logística.

microRNAs	Estimativa	Erro Padrão	Teste de Wald	valor p
miR-143-3p	-0,000030	0,000015	4,2606	0,039
miR-144-5p	-0,000050	0,000016	8,3782	0,004
miR-20a-3p	0,000185	0,000092	4,0705	0,044
miR-21-5p	0,000051	0,000017	9,0938	0,003
miR-28-5p	0,000169	0,000067	6,3862	0,012
miR-30a-5p	-0,000060	0,000023	5,9562	0,015
miR-320d	0,000212	0,000101	4,3716	0,037
miR-33b-5p	0,000360	0,000143	6,3143	0,012
miR-362-3p	-0,000250	0,000102	6,1206	0,013
miR-369-5p	-0,000080	0,000032	6,3043	0,012
miR-409-5p	0,000420	0,000183	5,2417	0,022
miR-429	0,000220	0,000085	6,6892	0,001
miR-486-5p	-0,000110	0,000030	13,0855	0,003
miR-490-3p	-0,000230	0,000075	9,4291	0,002
miR-493-5p	0,000157	0,000052	9,2389	0,002
miR-509-3-5p	-0,000170	0,000071	5,8666	0,015
miR-577	0,000669	0,000209	10,1961	0,001
miR-629-3p	0,000200	0,000055	13,0560	0,003

O teste qui-quadrado para a adequação de ajuste do modelo ao conjunto de dados não apresentou falta de ajuste com valor $p = 0,939$. A estatística C é igual a 99,9%, indicando a qualidade na separação dos tecidos em tumorais e normais.

Apresenta-se na Tabela 7, em negrito, cinco MicroRNAs (miR-362-3p, miR-369-5p, miR-493-5p, miR-509-3-5p e miR-577) e mais em relação a Tabela 6 do teste do Wilcoxon-Mann-Whitney, que estão ilustrados nas Figuras 10 e 11.



(a) miR-362-3p.

(b) miR-369-5p.

Figura 10: Boxplot da expressão dos microRNAs por tecido normal e tumoral para a regressão logística.

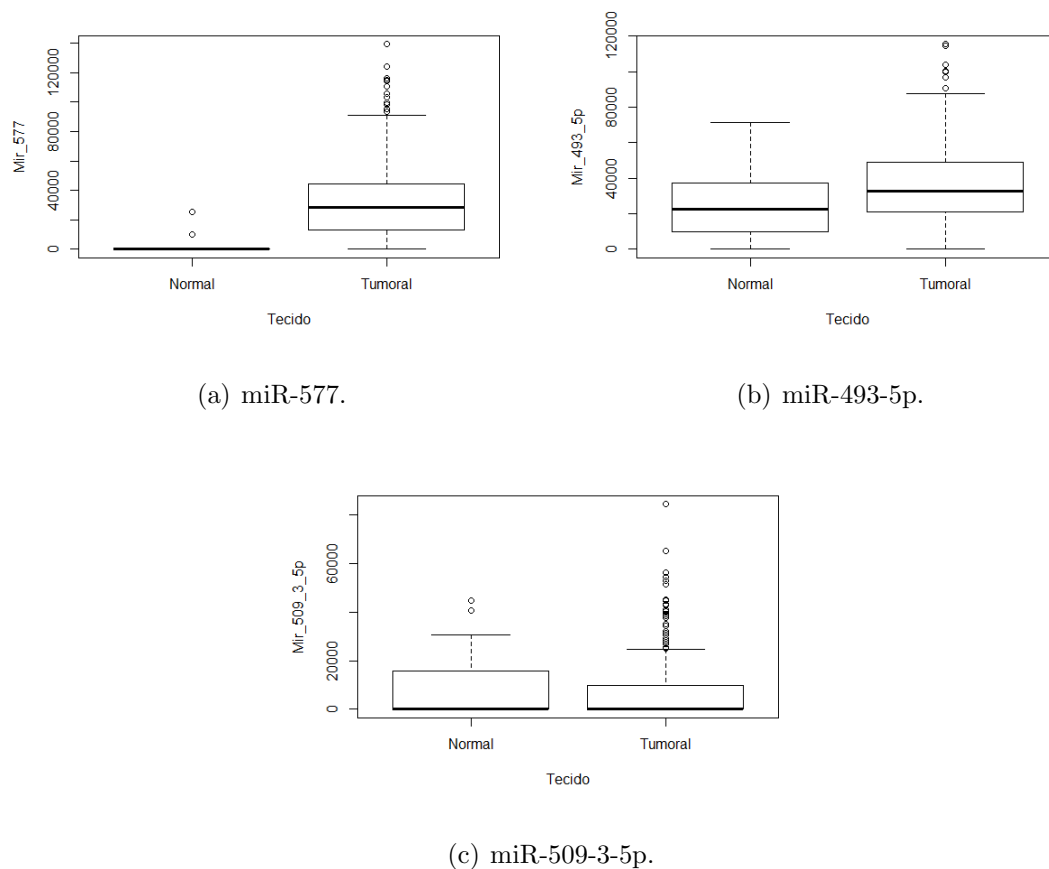


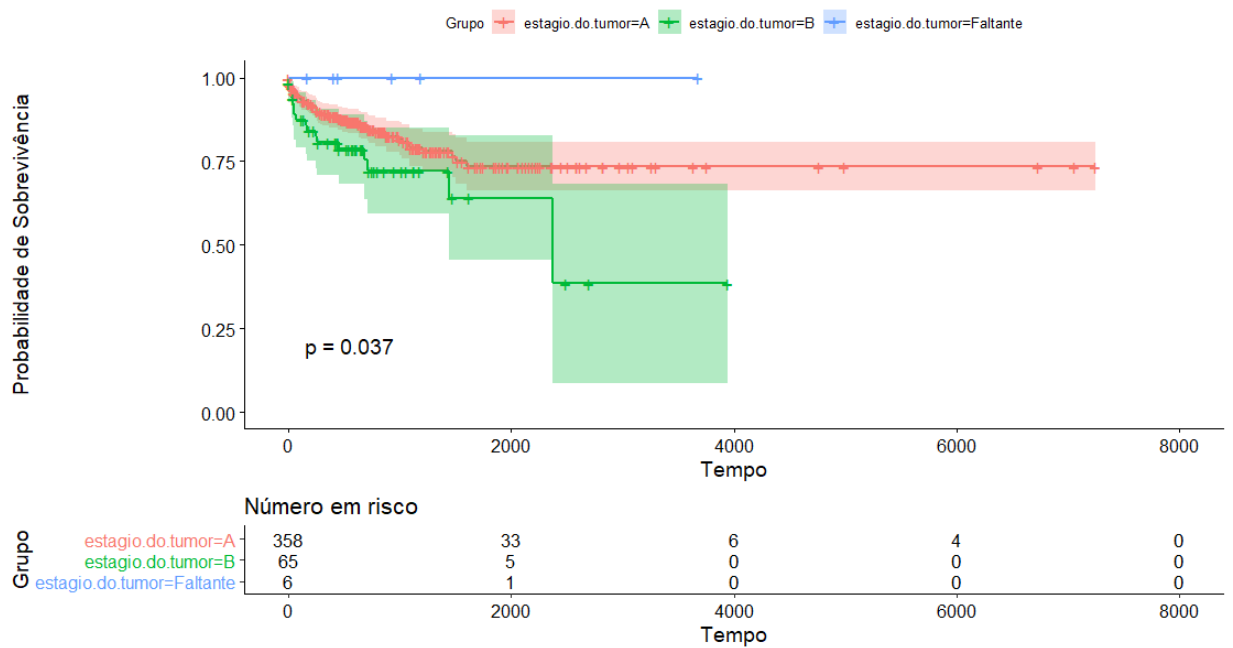
Figura 11: Boxplot da expressão dos microRNAs por tecido normal e tumoral para a regressão logística.

4.4 Análise do tempo de seguimento do paciente

A análise de sobrevivência foi realizada utilizando o software gratuito R Core Team (2018), por meio dos pacotes *survival* (Therneau, 2015) e *survminer* (Kassambara & Kosinski, 2019).

A variável tempo utilizada foi o tempo em dias até o último seguimento do paciente e a variável censura é o status de vida do paciente, se ele morreu devido ao câncer de pulmão ($\text{status} = 1$, com $n = 197$), e caso o paciente tenha sido perdido durante o estudo por outro motivo ($\text{status} = 0$, com $n = 356$), recordando que o total de amostras é $n = 553$.

De todas curvas de sobrevivência para as variáveis clínico-histopatológicas e dados demográficos o teste log-rank apresentou diferença significativa para o estágio do tumor (valor $p = 0,037$) e para a etnia (valor $p = 0,019$). Para o estágio do tumor pode-se observar na Figura 12 que o Grupo A apresentou maior sobrevivência, pois o câncer foi detectado precocemente e o Grupo B apresentou maior mortalidade. Para a etnia, observa-se na Figura 13 que os brancos apresentam menor sobrevivência que negros ou afro-americanos.



(a) Tempo de sobrevivência por estágio do tumor.

Figura 12: Curvas de sobrevivência estimadas por meio de Kaplan Meier para os grupos de estágio do tumor.

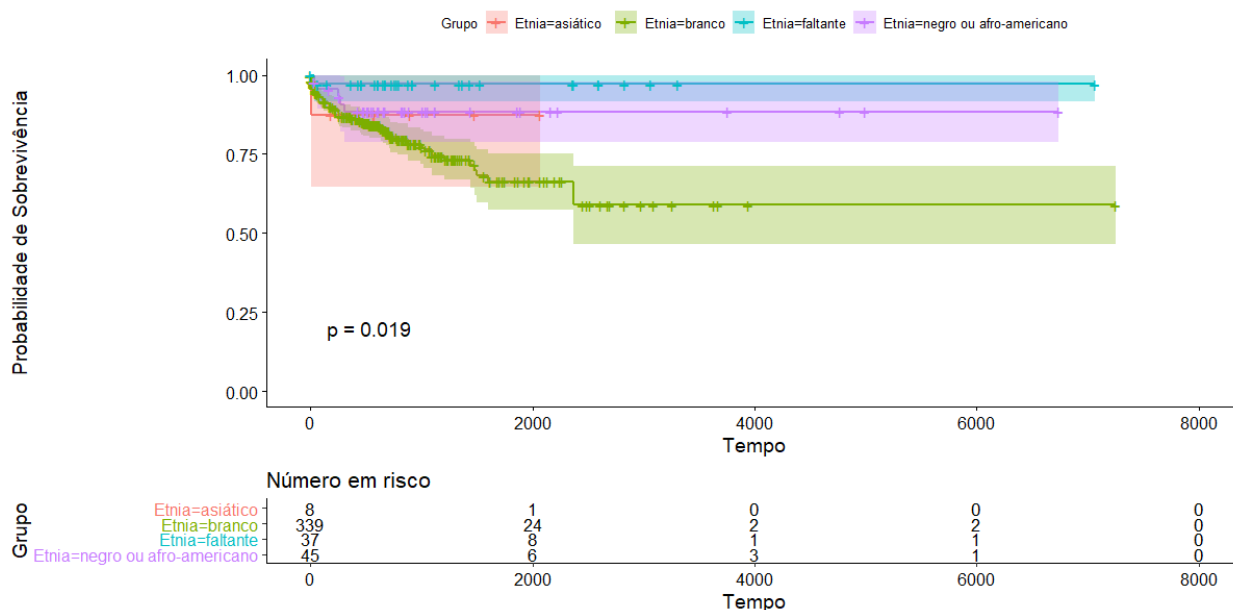
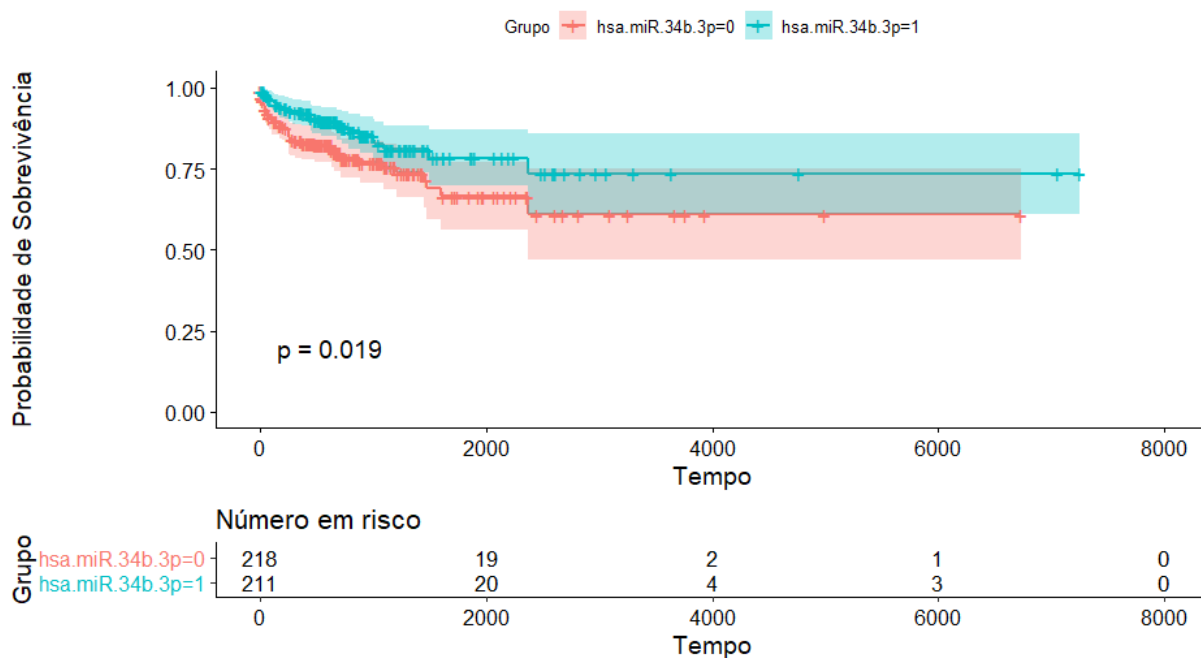


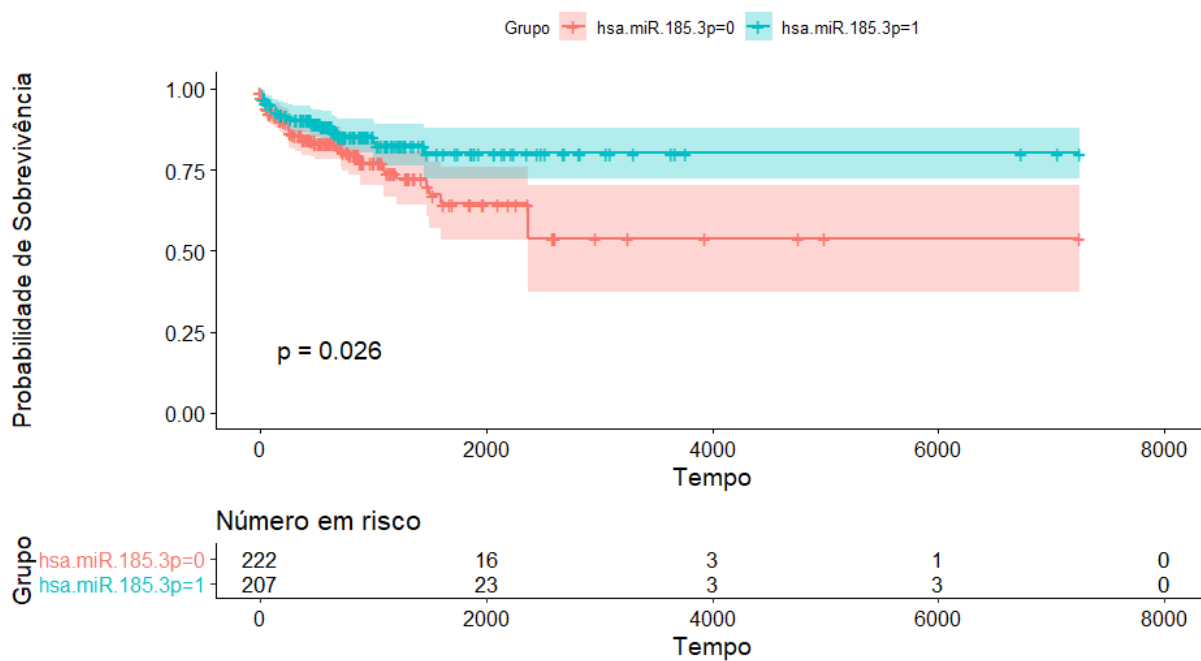
Figura 13: Curvas de sobrevivência estimadas por meio de Kaplan Meier para as etnias.

Para analisar se a expressão aumentada ou diminuída dos microRNAs interfeririam na sobrevivência, dividiu-se os dados de microRNAs em abaixo da mediana como Grupo 0 e, acima da mediana, como Grupo 1.

Os microRNAs miR-34b-3p, miR-185-3p, miR-1247-3p e miR-4661-5p apresentaram diferença estatística no teste log-rank com valores $p \leq 0,05$. A cor vermelha (Grupo 0) nas figuras seguintes indicam os valores abaixo da sua respectiva mediana e a cor azul (Grupo 1) indica os valores acima da mediana, com os respectivos intervalos de confiança. Pelas curvas de sobrevivências dos dois grupos, nota-se que o Grupo 1 com valores acima da mediana apresentaram maior sobrevivência em relação aos pacientes com valores abaixo da mediana para estes microRNAs. Observa-se que aconteceu isso para as Figuras 14 e 15.

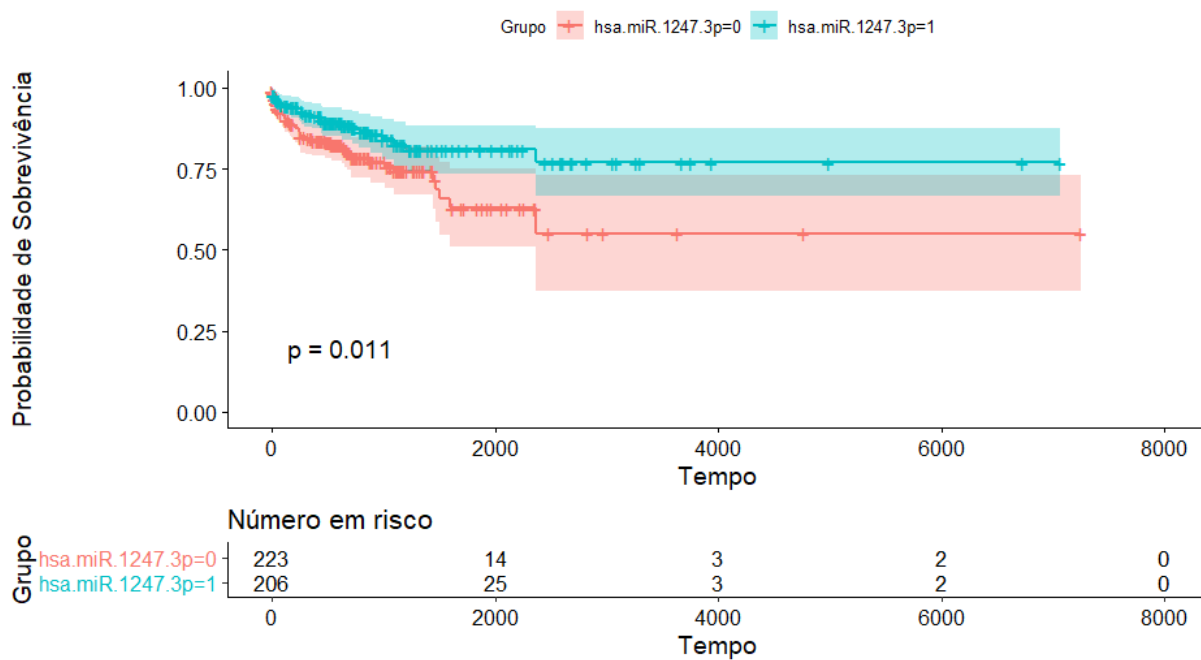


(a) Tempo de sobrevivência do miR-34b-3p.

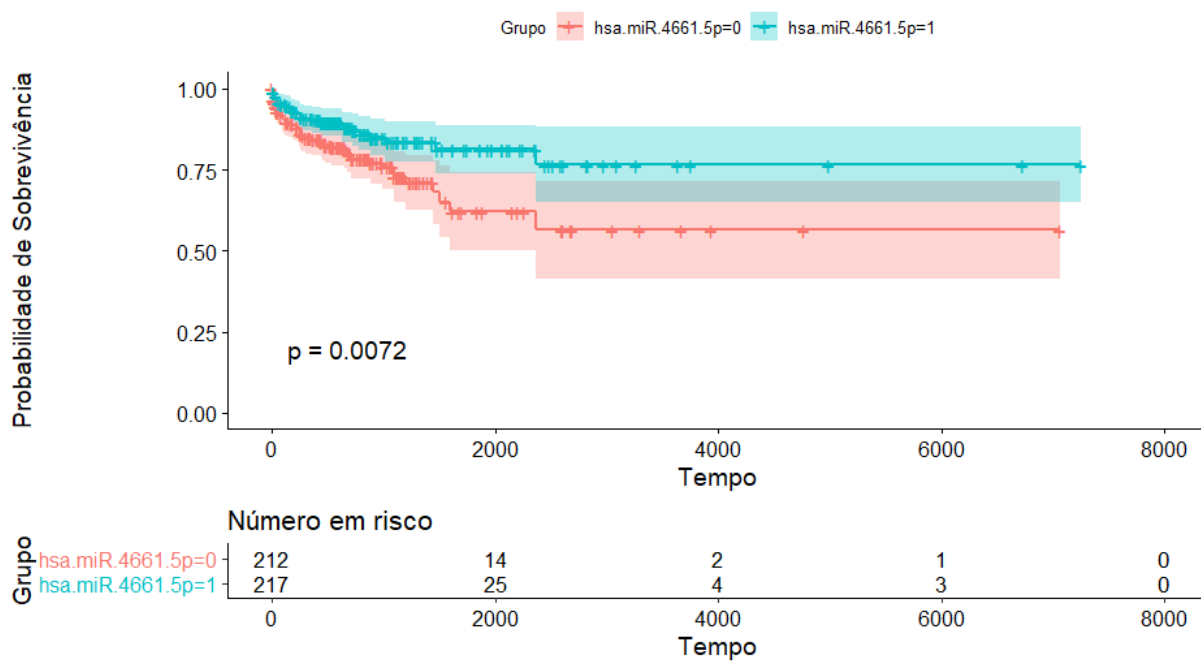


(b) Tempo de sobrevivência do miR-185-3p.

Figura 14: Curvas de sobrevivência por meio do estimador de Kaplan Meier para o miR-34b-3p e miR-185-3p.



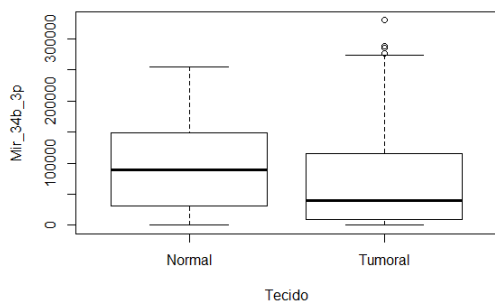
(a) Tempo de sobrevivência do miR-1247-3p.



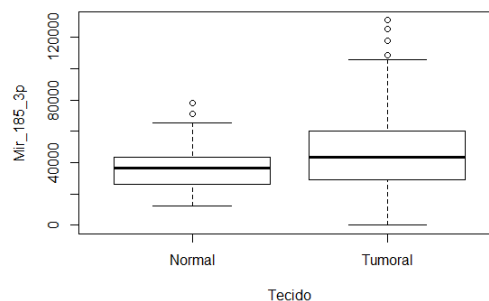
(b) Tempo de sobrevivência do miR-4661-5p.

Figura 15: Curvas de sobrevivência por meio do estimador de Kaplan Meier para o miR-1247-3p e miR-4661-5p.

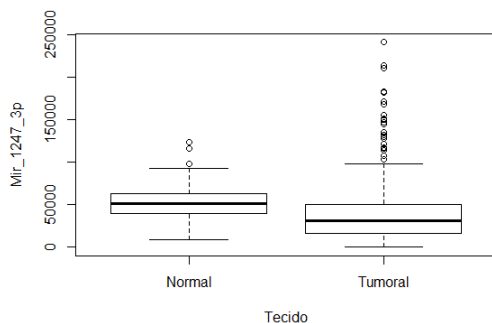
A Figura 16 apresenta o boxplot dos microRNAs que são estatisticamente significativos na análise de sobrevivência em relação ao tipo de tecido:



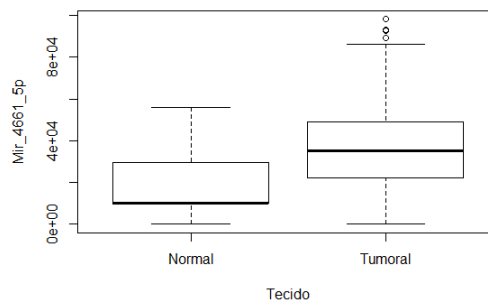
(a) miR-34b-3p.



(b) miR-185-3p.



(c) miR-1247-3p.



(d) miR-4661-5p.

Figura 16: Boxplot da expressão do microRNA por tecido normal e tumoral para a análise do tempo de seguimento do paciente..

4.5 Árvores de classificação e regressão

A seguir são apresentados os resultados encontrados nas construções das árvores de classificação e regressão.

4.5.1 Árvore de classificação

As árvores de classificação foram construídas no software estatístico R (R Core Team, 2018), utilizando os pacotes *rpart* (Therneau & Atkinson, 2018),

rpart.plot (Milborrow, 2019) e *partykit* (Hothorn & Zeileis, 2015).

Como descrito no Capítulo 3, a representação de uma árvore de decisão é composta por nós, ramificações e folhas que são construídas para explicar a relação que existe entre os dados. Na árvore, os nós constituem os pontos de maior separação e as variáveis são usadas para classificar os dados. Nas árvores de classificação, as folhas representam a fração de registros classificados corretamente pela árvore de decisão.

A Figura 17 da árvore completa (sem podas) construída para classificar o tecido histopatologicamente normal ou tumoral é composta por 5 microRNAs (miR-21-5p, miR-1287-3p, miR-210-3p, miR-133a-3p e let-7g-3p) e a variável sexo de todos os 709 microRNAs e as 9 variáveis clínicas. Pode-se destacar que a cor azul indica a probabilidade de a classificação ser “tumoral” e a cor cinza a probabilidade de a classificação ser “normal”, mas além das cores, no canto esquerdo dos nós finais também possui a classificação de cada um.

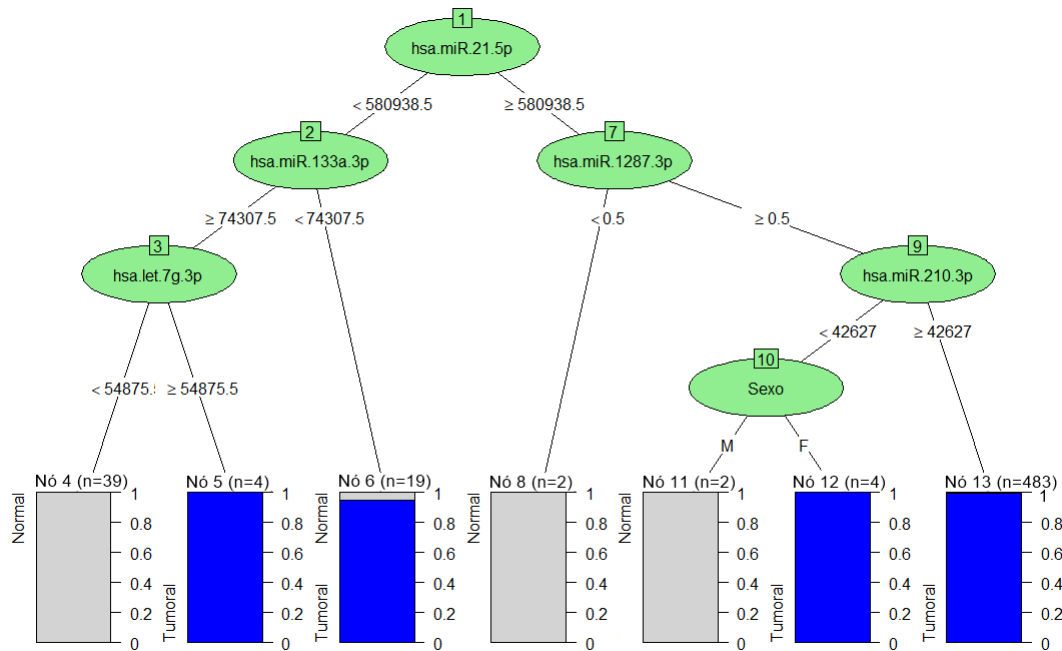


Figura 17: Árvore de classificação sem poda, em que, as cores azul e cinza representam tecido tumoral e normal, respectivamente.

Assim, por meio da Figura 17, é possível visualizar que existe uma boa classificação para os dados, já que as cores estão bem definidas em cada folha. Os nós finais 4, 8 e 11 foram classificados como tecidos normais e os nós 5, 12 e 13 foram classificados como tecidos tumorais. Apenas o nó 6 apresentou uma classificação de 0,95 para tumoral e 0,05 para normal.

O software R, apresentou o erro associado ao nó raiz (em inglês *root node error*), que é a porcentagem de registros classificados corretamente no primeiro nó de divisão (miR-21-5p). Na árvore da Figura 17, o primeiro nó classificou $45/553 = 0,0813$, ou seja, 45 amostras corretas de 553. Porém, como as cores da Figura 17 estão bem definidas entende-se que nos próximos nós houve uma melhora na classificação.

Mesmo que a árvore da Figura 17 não tenha ficado complexa, ou seja, com um número elevado de ramos, é possível melhorar esta árvore. A Tabela 8 apresenta o resumo do modelo para a construção da árvore de classificação.

Tabela 8: Resumo do modelo para a poda da árvore de classificação.

<i>cp</i>	<i>Nsplit</i>	Erro Relativo	Erro de Validação Cruzada	DP*
0,4000	0	1,0000	1,0000	0,1428
0,3778	1	0,6000	0,6667	0,1183
0,0889	2	0,2222	0,4000	0,0927
0,0444	3	0,1333	0,3111	0,0820
0,0222	4	0,0888	0,2889	0,0791
0,0100	6	0,0444	0,2889	0,0791

*Desvio Padrão.

O índice custo-complexidade (*cp*) indica os valores dos parâmetros de complexidade e *nsplit* é o número de divisões da árvore. O erro relativo é o erro para a previsão dos dados que foram usados para estimar o modelo, ou seja, é a fração de casos mal classificados na iteração em relação à fração de casos mal classificados no nó raiz. O erro de validação cruzada (gerado pela validação cruzada interna do rpart) e o desvio padrão dos dados validados cruzadamente inversamente proporcional ao número de indivíduos incluídos no processo. Cada linha da Tabela 8 é um índice da

árvore, isso ajuda a interpretar e tomar decisão sobre em que ponto podar a árvore.

Para realizar a poda da árvore apresentada na Figura 17 é necessário usar o cp . O valor ideal de cp é o da menor árvore que possua o menor erro de validação cruzada. Em outras palavras, refere-se ao tamanho da árvore e a taxa de erro para ajudar a evitar o ajuste excessivo. Assim, árvores grandes com uma baixa taxa de erro são penalizadas em favor de árvores menores. Pelas informações da Tabela 8, o melhor valor para realizar a poda é $cp = 0,0222$, o qual possui o menor erro de validação cruzada (0,2889).

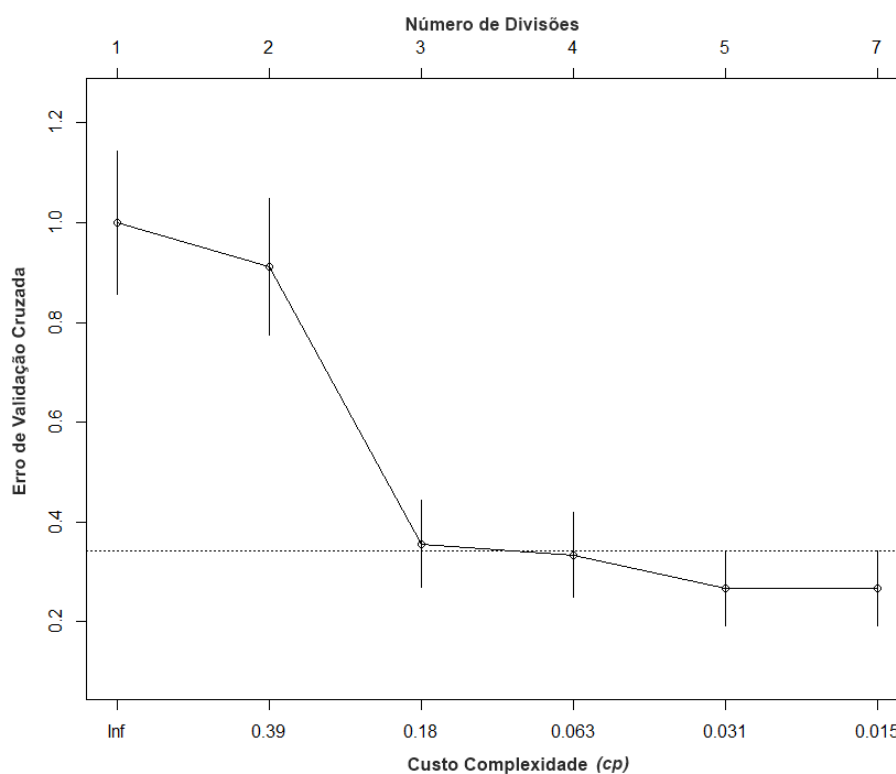


Figura 18: Gráfico do custo-complexidade, erro de validação cruzada e número de divisões para árvore de classificação.

Os valores do parâmetro de complexidade são gerados em relação ao erro de validação cruzada calculado pelo algoritmo *rpart* na Figura 18. A linha tracejada representa o erro validado cruzado mais alto menos o erro validado cruzado mínimo,

mais o desvio padrão do erro nessa árvore. Uma escolha razoável de cp para poda geralmente é o valor mais à esquerda, em que a média é menor que a linha horizontal. Assim, tem-se outro método para escolher o melhor valor de cp e pode-se confirmar visualmente que o $cp = 0,0222$ é o melhor valor para a poda.

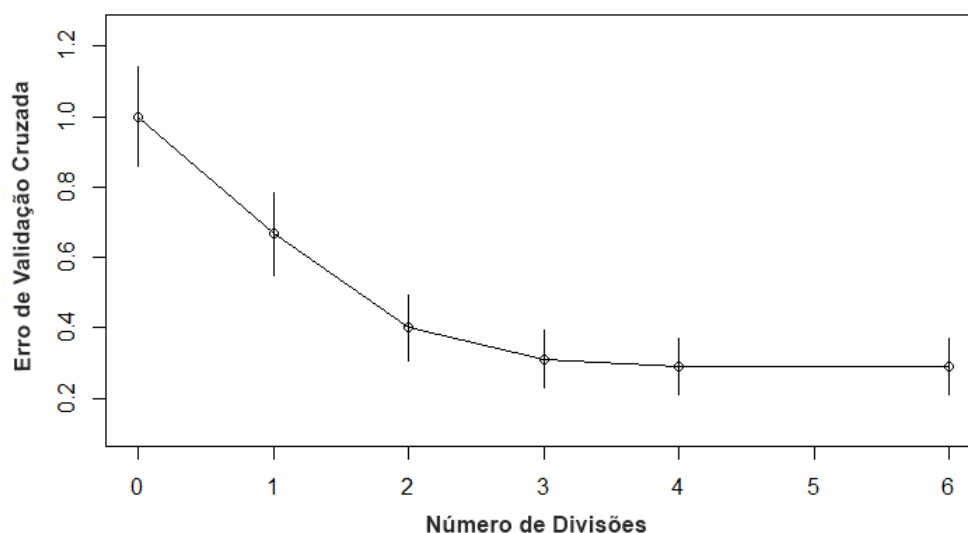


Figura 19: Gráfico do número de divisões pelo erro de validação cruzada para árvore de classificação.

A Figura 19 ilustra o número de divisões (*splits*) que acontece na árvore em relação ao erro de validação cruzada.

Após podar a árvore de classificação em tecido normal ou tumoral segundo o valor do custo-complexidade, obtém-se a Figura 20. Pode-se observar que a árvore continua a apresentar uma boa classificação mesmo com a retirada do microRNAs miR-210-3p e da variável sexo. O nó 6 continua a classificar o tecido da mesma forma da Figura 17. Nota-se que houve pouca mudança na árvore.

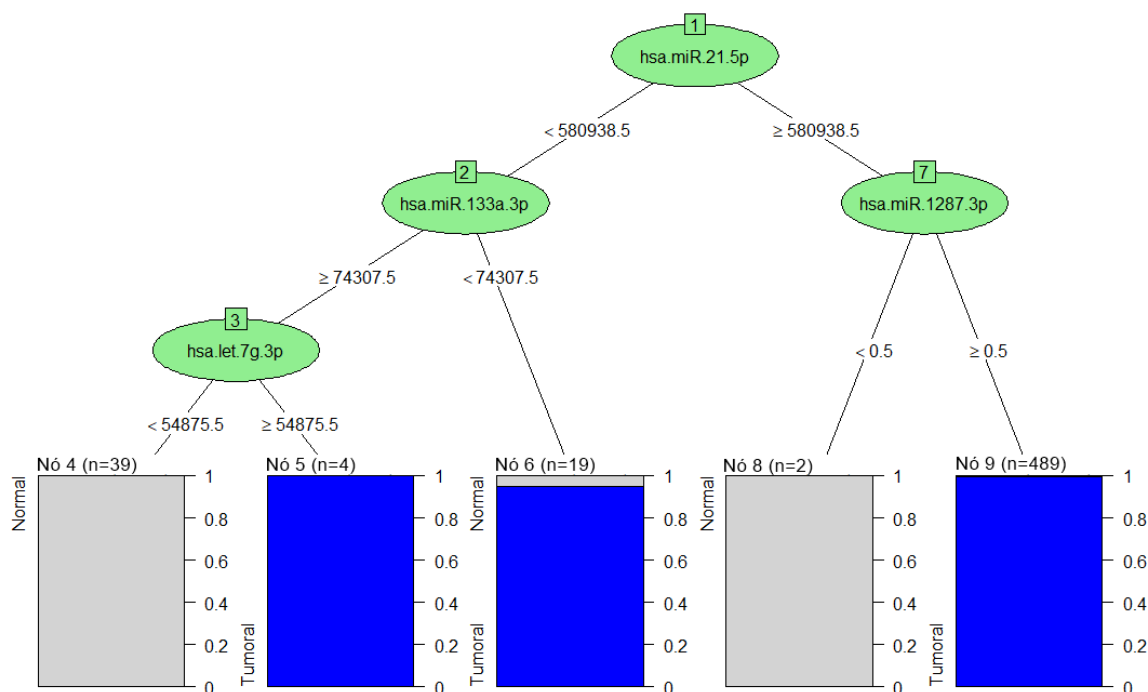
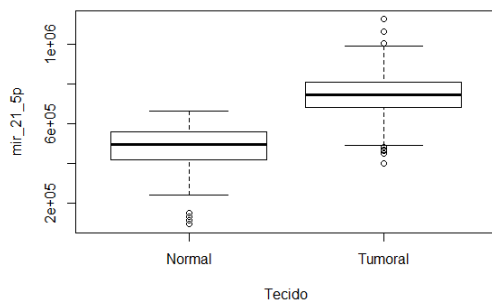


Figura 20: Árvore de classificação pós poda, em que, as cores azul e cinza representam tecido tumoral e normal, respectivamente.

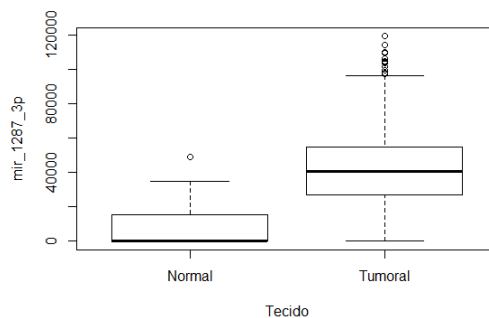
Assim, tem-se duas interpretações, como exemplo, para compreender melhor as informações da árvore:

- Se um microRNA apresentar a expressão do miR-21-5p maior que 580938,5 e a do miR-1287-3p menor que 0,5 então, é classificado como um tecido normal, mas se o último for maior que 0,5, então se classifica como tecido tumoral.
- Se um microRNA apresentar a expressão do miR-21-5p menor que 580938,5, uma possibilidade seria apresentar o miR-133a-3p maior que 74307,5 e assim ter o tecido com 95% de chance de ser tumoral e 5% de ser normal.

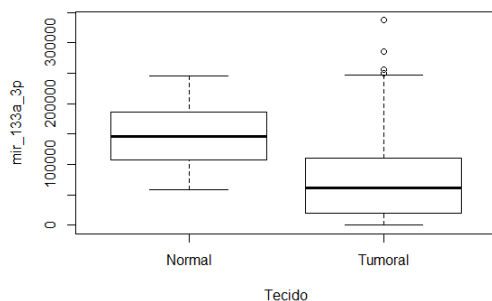
Na Figura 21, pode-se observar a distribuição dos microRNAs encontrados para a árvore podada. Nota-se que os quatro microRNAs possuem valores discrepantes, devido à grande variabilidade da expressão dos microRNAs.



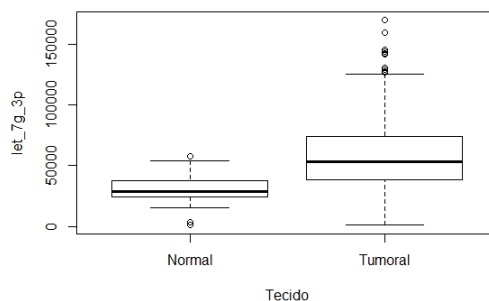
(a) miR-21-5p.



(b) miR-1287-3p.



(c) miR-133a-3p.



(d) let-7g-3p.

Figura 21: Boxplot da expressão dos microRNAs por tecido normal e tumoral para a árvore de classificação pós poda.

4.5.2 Árvore de sobrevivência

As árvores de sobrevivência, que foram construídas a partir das árvores de regressão, foram realizadas no software estatístico R e os pacotes são os mesmos da árvore de classificação, com adição do pacote *survival* (Therneau, 2015), para ajudar no estudo do tempo de vida dos pacientes.

Na árvore de regressão, os nós terminais representam a resposta prevista neste nó. Estes são calculados como a resposta média de todos os dados de treinamento, que foram classificados em cada nó do terminal.

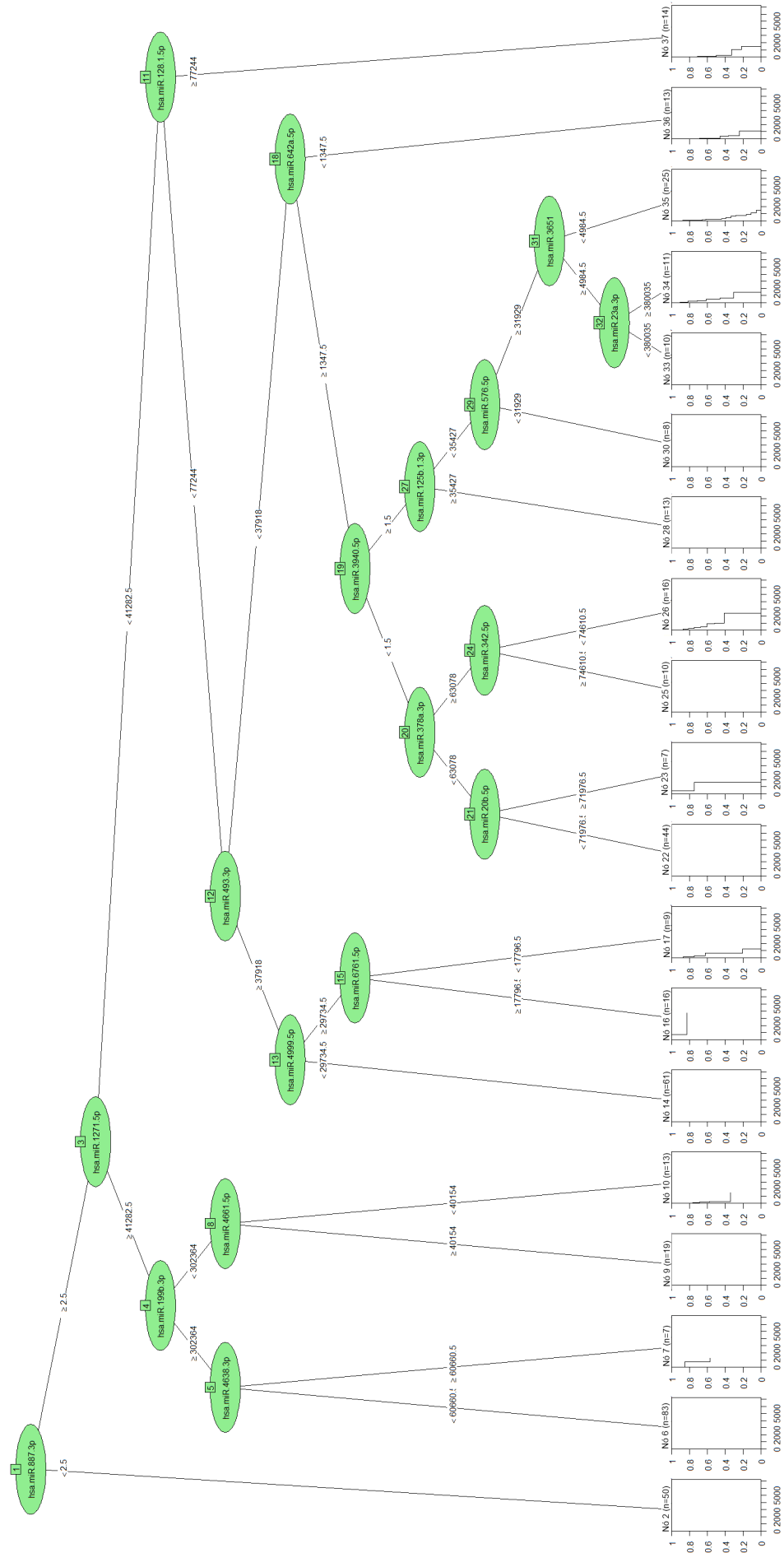


Figura 22: Árvore de sobrevivência sem poda.

A Figura 22 ilustra a árvore de regressão para o tempo de sobrevivência completa sem podas. Foram testadas todas as variáveis, os 709 microRNAs e as 9 variáveis clínicas histopatológicas. A árvore é composta por 18 microRNAs (miR-887-3p, miR-1271-5p, miR-199b-3p, miR-4638-3p, miR-4661-5p, miR-128-1-5p, miR-493-3p, miR-4999-5p, miR-6761-5p, miR-642a-5p, miR-3940-5p, miR-378a-3p, miR-20b-5p, miR-342-5p, miR-125b-1-3p, miR-576-5p, miR-3651 e miR-23a-3p).

A árvore de regressão para a sobrevivência ficou muito complexa, pois apresentou muitos galhos, ou seja, ficou muito extensa e pensando em utilizá-la na prática ficou inviável. Neste caso é necessário a realização da poda, conforme os resultados apresentados na Tabela 9.

Tabela 9: Resumo do modelo para a poda da árvore de sobrevivência.

<i>cp</i>	<i>Nsplit</i>	Erro Relativo	Erro de Validação Cruzada	DP*
0,0561	0	1,0000	1,0038	0,0714
0,0413	2	0,8877	0,9760	0,0739
0,0389	5	0,7636	1,1614	0,1019
0,0366	8	0,6470	1,1675	0,1094
0,0345	9	0,6104	1,1794	0,1118
0,0333	10	0,5758	1,1814	0,1124
0,0269	11	0,5425	1,2295	0,1200
0,0249	12	0,5155	1,2784	0,1254
0,0248	13	0,4906	1,2850	0,1268
0,0247	14	0,4658	1,2912	0,1274
0,0245	15	0,4410	1,2912	0,1274
0,0241	16	0,4164	1,2942	0,1274
0,0191	17	0,3923	1,3003	0,1281
0,010	18	0,3731	1,3051	0,1297

*Desvio Padrão.

O erro associado ao nó raiz foi de $385,18/429 = 0,8978$, ou seja, foram 385,18 amostras classificadas corretamente de 429 no primeiro nó (miR-887-3p).

Para a realização da poda por meio das informações geradas é possível encontrar o menor erro de validação cruzada (0,9760) e encontrar assim, o $cp = 0,0413$, cujo valor também pode ser confirmado na Figura 23. A Figura 24 apresenta o número

de divisões da árvore em relação ao erro de validação cruzada.

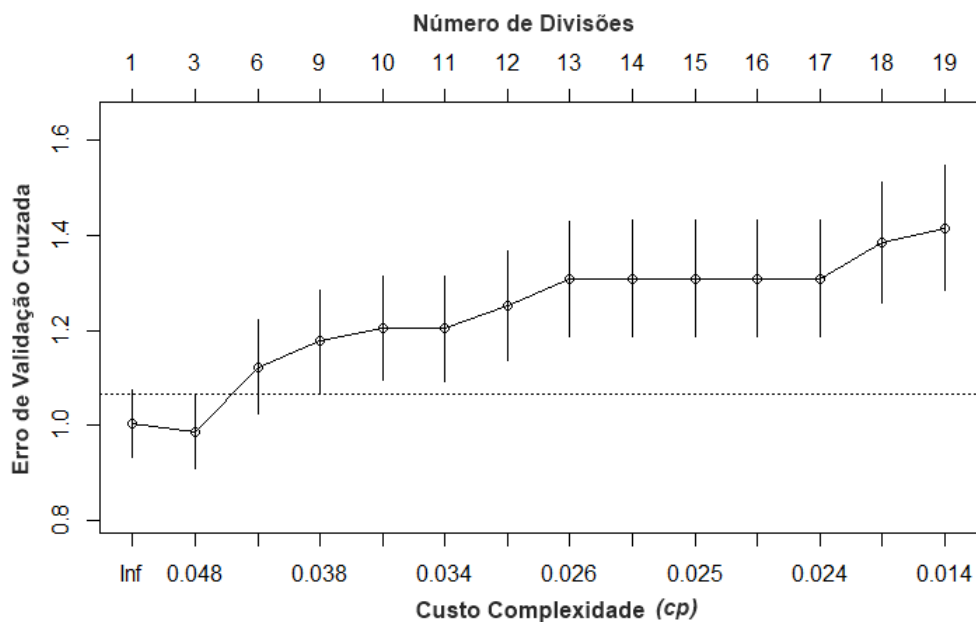


Figura 23: Gráfico do custo-complexidade, erro de validação cruzada e número de divisões para árvore de sobrevivência.

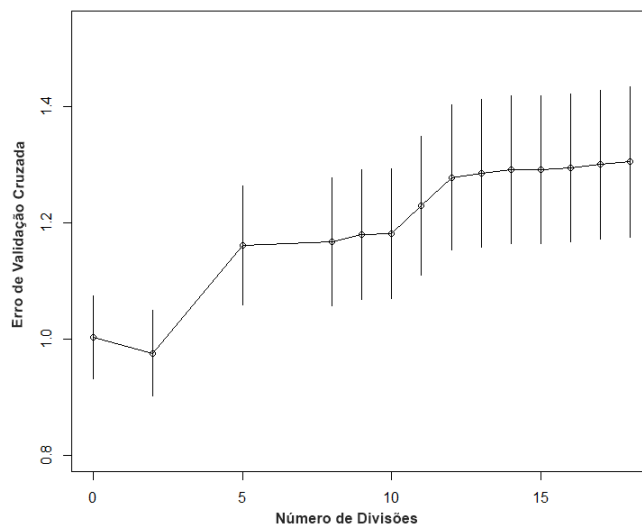


Figura 24: Gráfico do número de divisões pelo erro de validação cruzada para árvore de sobrevivência.

Após a poda da árvore de regressão para a sobrevivência, por meio do valor de custo-complexidade, obteve-se a Figura 25. Observa-se que a árvore teve uma diminuição de variáveis, de 18 para 5 microRNAs, os microRNAs miR-887-3p, miR-1271-5p, miR-128-1-5p, miR-493-3p e miR-4999-5p permaneceram na árvore.

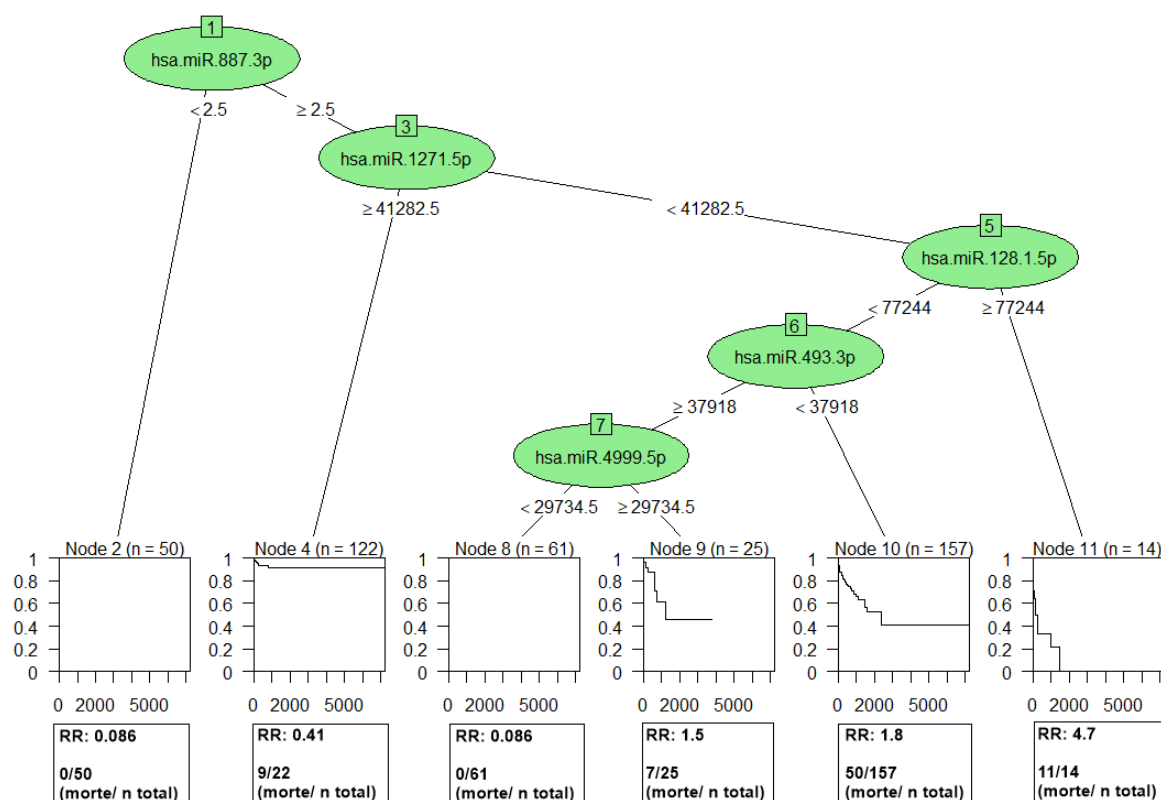


Figura 25: Árvore de sobrevivência pós poda. RR: Risco relativo.

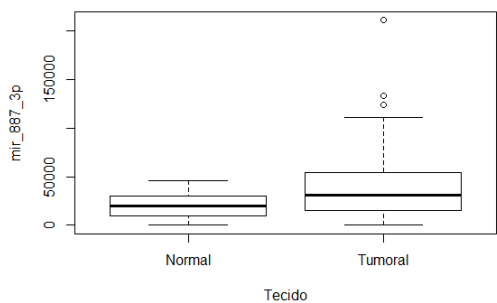
O nó terminal da árvore da Figura 25 exibe a curva de sobrevivência de Kaplan-Meier ajustada para as observações nesse nó.

Avaliando a árvore da Figura 25, pode-se observar que:

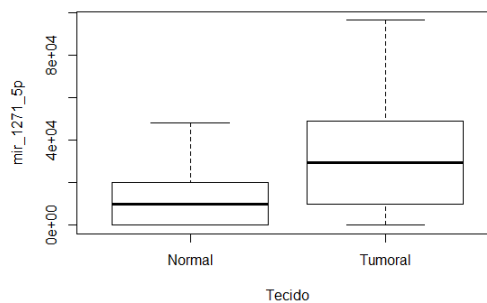
- Se o microRNA miR-887-3p apresentar expressão menor que 2,5 então têm-se um risco relativo de 0,086 (em relação a um risco estimado da linha de base) e classifica-se 0/50, ou seja, significa que nenhuma pessoa morreu e que 50 pessoas realmente sobreviveram.

- Se o microRNA miR-887-3p apresentar expressão maior que 2,5, a expressão do miR-1271-5p menor que 41282,5 e a expressão do miR-128-15p for maior que 77244, tem-se um risco relativo de 4,7 e classifica-se 11/14, ou seja, 11 pessoas realmente morreram e 3 pessoas sobreviveram.

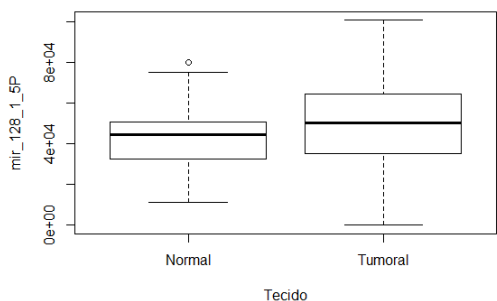
Nas Figuras 26 e 27 pode-se observar a distribuição dos microRNAs da árvore de sobrevivência pós poda, dividido por tipo de tecido. O tecido tumoral dos microRNAs miR-887-3p, miR-1271-5p, miR-128-15p e miR-493-3p (Figura 26) está superexpresso em relação ao tecido normal e o miR-4999-5p apresentou observações discrepantes.



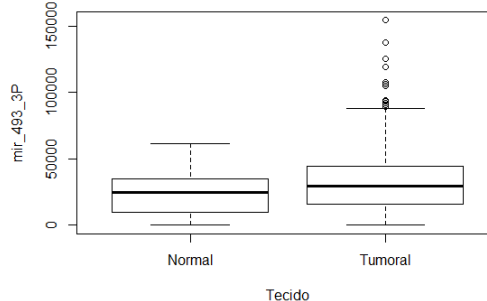
(a) miR-887-3p.



(b) miR-1271-5p.

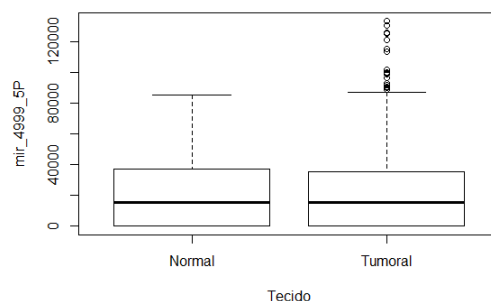


(c) miR-128-1-5p.



(d) miR-493-3p.

Figura 26: Boxplot da expressão dos microRNAs por tecido normal e tumoral para a árvore de sobrevivência pós poda.



(a) miR-4999-5p.

Figura 27: Boxplot da expressão do microRNA por tecido normal e tumoral para a árvore de sobrevivência pós poda.

4.6 Análise biológica

Os microRNAs identificados como significativamente desregulados, de acordo com os métodos de árvore de classificação e árvore de sobrevivência foram utilizados para análise de predição de genes-alvo potencialmente regulados pelos microRNAs. Essa análise foi realizada utilizando a ferramenta de bioinformática Portal de Integração de Dados de microRNAs (mirDIP, em inglês *microRNA Data Integration Portal*). O objetivo da análise é a geração de interpretação biológica dos resultados encontrados neste trabalho.

Dessa forma, os microRNAs identificados geraram listas de genes-alvo, os quais, por sua vez, foram utilizados para análise de anotação funcional, inseridos em redes de interação microRNAs-gene-alvo e de vias moleculares. Assim, foram encontradas as combinações de genes e as vias moleculares que estes regulam por meio da ferramenta de bioinformática *toppgene*. Os resultados dessas análises sugerem um papel para os microRNAs como moduladores de vias moleculares no câncer de pulmão de células não pequenas. Ambas as ferramentas de bioinformática estão disponíveis gratuitamente em sites.

- **Análise da árvore de classificação**

Para os microRNAs identificados na árvore de classificação pós poda (let-7g-3p, miR-1287-3p, miR-133a-3p, miR-21-5p) foram identificados 1889 genes-alvo. Destes, foram detectadas 82 combinações de genes que participam de vias moleculares enriquecidas (estatisticamente significativas). A Tabela 10 apresenta as primeiras vias moleculares com maior significância estatística.

Tabela 10: Vias moleculares identificadas a partir dos genes-alvo de microRNAs associados a árvore de classificação.

ID	Name	Valor p^1	Lista	Genoma
P00057	Via de sinalização Wnt	$2,89 \times 10^{-04}$	58	305
83105	Vias moleculares no câncer	$1,31 \times 10^{-03}$	68	395
137939	Efetores diretos do TP53	$6,05 \times 10^{-03}$	30	132
921162	Via de sinalização FoxO	$6,05 \times 10^{-03}$	30	132
1268855	Doenças de transdução de sinais	$1,22 \times 10^{-02}$	62	373
1269347	Redução dos níveis de Cálcio citosólicos	$1,27 \times 10^{-02}$	8	14
PW:0000166	Metabolismo celular do nicotinato e nicotinamida	$3,85 \times 10^{-02}$	7	12
P00012	Via de sinalização da Caderina	$4,18 \times 10^{-02}$	32	159
1270303	Orientação de axônios	$4,51 \times 10^{-02}$	82	554
PW:0000491	Sinalização da vasopressina	$6,13 \times 10^{-01}$	7	17
PW:0000713	Sinalização da vasopressina via receptor tipo 2	$7,65 \times 10^{-02}$	7	13
P00021	Sinalização do FGF	$8,14 \times 10^{-02}$	23	102
83048	Sinalização da MAPK	$9,07 \times 10^{-02}$	44	255
1269443	Sinalização doNGF	$1,05 \times 10^{-01}$	72	483
83064	Sinalização do TGF-beta	$1,05 \times 10^{-01}$	20	84
1269460	Sinalização do NGF via TRKA pela membrana plasmática	$1,64 \times 10^{-01}$	60	390
1272486	Resistência à insulina	$1,76 \times 10^{-01}$	23	107
989139	Sinalização AMPK	$1,82 \times 10^{-01}$	25	121

¹Valor p geral.

A Tabela 10 apresenta a identificação, nome da via molecular, o valor p que é gerado combinando os valores p de vários testes em um valor p geral, mas detalhes pode ser visto no trabalho de Chen et al. (2009), a contagem de ocorrências na lista de consultas e a contagem de ocorrências no genoma. Para este último item, por exemplo, a redução dos níveis de cálcio citosólicos é fundamental para o funcionamento e a sobrevivência da célula, com valor $p = 1,27 \times 10^{-01}$, teve oito genes que apareceram na lista dos genes encontrados neste trabalho comparado aos 14 que são necessários do genoma humano.

A via molecular de sinalização beta TGF está envolvidas em vários processos celulares, desde o embrião até o organismo adulto, incluindo o crescimento da célula, a diferenciação celular, a apoptose, homeostase e outras funções.

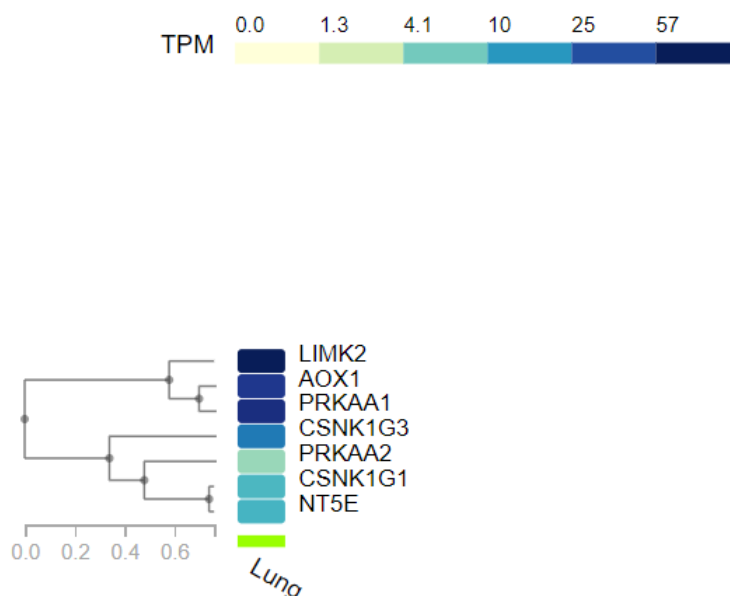


Figura 28: Genes para a via molecular metabolismo celular do nicotinato e nicotina-mida.

A Figura 28, gerada no GTEx Portal (site: <https://www.gtexportal.org/home/>), ilustra os genes que formaram a via molecular, o qual estão transcritos por milhão (TPM) e as combinações existentes entre os genes por meio dos traços em cinza que representam a expressão dos genes participantes dessa via molecular, em tecido pulmonar normal.

A importância dessa análise é para mostrar que os genes-alvo de microRNAs encontram-se expressos em tecido pulmonar normal. E que é importante verificar isso, pois se um ou mais genes-alvo de microRNAs não estão expressos normalmente, no tecido de interesse, não devem ser relevantes na doença, na maioria das vezes.

• Análise da árvore de sobrevivência

Para os microRNAs encontrados na árvore de sobrevivência pós poda (miR-

887-3p, miR-1271-5p, miR-128-1-5p, miR-493-3p e miR-4999-5p) foram identificados 1067 genes-alvo. Destes, foram detectadas 296 combinações de genes inseridos em vias moleculares. A Tabela 11 apresenta as mais significativas vias moleculares.

Tabela 11: Vias moleculares identificadas a partir dos genes-alvo de microRNAs associados a árvore de sobrevivência.

ID	Name	valor p^1	Lista	Genoma
1270303	Orientação de axônios	$1,08 \times 10^{-09}$	74	554
83065	Orientação de axônios	$1,27 \times 10^{-03}$	27	175
782000	Proteoglicanas no câncer	$2,33 \times 10^{-05}$	33	203
1269279	Fagocitose dependente do receptor Fcgamma (FCGR)	$4,05 \times 10^{-05}$	21	95
137915	Sinalização regulada pela tirosina quinase RET	$6,43 \times 10^{-05}$	13	38
1269484	Permeabilidade vascular mediada por VEGFR2	$1,76 \times 10^{-02}$	9	29
1269480	Sinalização pelo VEGF	$1,43 \times 10^{-04}$	44	341
799177	Via de sinalização do estrógeno	$3,41 \times 10^{-04}$	20	98
P00034	Via de sinalização da integrina	$4,93 \times 10^{-04}$	27	167
M3342	Via de sinalização da integrina	$5,62 \times 10^{-04}$	12	38
83067	Adesão focal	$5,48 \times 10^{-04}$	30	199
83048	Sinalização da MAPK	$6,39 \times 10^{-04}$	35	255
1269443	Sinalização pelo NGF	$6,45 \times 10^{-04}$	54	483
1270323	Interações L1CAM	$6,72 \times 10^{-04}$	20	102
83089	Regulação do citoesqueleto de actina	$7,03 \times 10^{-04}$	31	212
138047	Sinalização direta EPHB	$9,08 \times 10^{-04}$	11	33
169348	Eventos de sinalização mediados por FAK (quinase de adesão focal)	$9,60 \times 10^{-04}$	14	54

¹Valor p geral.

A Tabela 11 é formada da mesma forma a anterior. Apareceram vias moleculares iguais da árvore de classificação, como: orientação de axônicos e sinalização da MAPK.

A via proteoglicanas no câncer em microambiente tumoral contribui para o aparecimento de vários tipos de cânceres, incluindo da proliferação até a metastase (Nikitovic et al., 2018).

As verificações biológicas feita acima, sobre os microRNAs encontrados neste trabalho baseados nas árvores, teve como principal objetivo averiguar se as vias moleculares que estes microRNAs regulam estavam ligadas ao câncer ou não. Pode-se observar que nos três métodos as vias apresentam relação com o câncer de pulmão.

5 CONCLUSÕES

O banco de dados analisado neste trabalho continha dados demográficos, fatores de risco, características clínico-patológicas e 709 variáveis de microRNAs, para 508 amostras histopatológicas de tecidos normais e 45 tumorais.

O trabalho teve como objetivos específicos analisar o banco de dados por meio de quatro métodos estatísticos: teste de Mann-Whitney, regressão logística, análise de sobrevivência e árvores de classificação e regressão, o qual, foi realizado por meio do algoritmo CART. Depois de realizar as análises pode-se tirar algumas conclusões a respeito de cada método, como:

- Comparando os testes de regressão logística, que testa o efeito da probabilidade do tecido ser tumoral, e teste de Wilcoxon-Mann-Whitney, que compara as medianas em relação ao tecido normal as variáveis clinico-histopatológicas não apresentaram significância estatística. O teste Mann-Whitney apresentou 25 microRNAs estatisticamente significativos, a regressão logística apresentou 18 microRNAs, e destes tem-se 13 microRNAs concordaram (miR-143-3p, miR-144-5p, miR-20a-3p, miR-21-5p, miR-28-5p, miR-30a-5p, miR-320d, miR-33b-5p, miR-409-5p, miR-429, miR-486-5p, miR-490-3p e miR-629-3p).
- Na análise do tempo de seguimento do paciente encontrou-se as variáveis estagio do tumor, etnia, miR-34b-3p, miR-185-3p, miR-200b-5p, miR-331-5p, miR-1247-3p, miR-4661-5p e miR-5187-5p.
- Para a árvore de classificação restaram apenas 4 microRNAs: miR-21-5p, miR-133a-3p, miR-1287-3p e let-7g-3p. O miR-21-5p também aparece tanto no teste de Wilcoxon-Mann-Whitney quanto na regressão logística.

- A árvore de regressão para a sobrevivência pós poda apresentou 5 microRNAs: miR-887-3p, miR-1271-5p, miR-128-1-5p, miR-493-3p e miR-4999-5p. Na regressão logística apareceu o miR-493-5p, da família do miR-493.

Para árvore de sobrevivência sem poda apareceu o miR-4661-5p, que também apareceu no teste de Wilcoxon-Mann-Whitney e análise do tempo de seguimento, o miR-378a-3p, apareceu no teste de Wilcoxon-Mann-Whitney, e a família do miR-20, que apareceu o miR-20b-5p na árvore de sobrevivência sem poda e o miR-20a-3p no teste de Wilcoxon-Mann-Whitney e na regressão logística.

Nesse trabalho, foi encontrado-se o miR-21-5p superexpresso no tecido tumoral, da família dos miR-21. O miR-21 promove a mobilidade e a invasão de tecidos tumorais, ao se ligar e bloquear genes supressores (Lee & Dutta, 2009).

Takamizawa et al. (2004) mostram a influência das famílias do let-7 associase na diminuição deste microRNAs com um pior prognóstico. E a introdução dele indica a potencialidade de supressor tumoral. No trabalho, o let-7g-3p também foi estatisticamente significativo no processo de discriminação dos tecidos histologicamente normais ou tumorais dos pacientes com o adenocarcinoma pulmonar, conforme resultado obtido pela construção da árvore de classificação.

A análise biológica foi feita com a intenção de averiguar se as vias moleculares que estes microRNAs regulam, estavam ligadas ao surgimento do câncer, pode-se observar que há ligação, ou seja, os microRNAs encontrados neste trabalho possuem ligações e podem ser potenciais biomarcadores para o adenocarcinoma pulmonar.

Baseando-se nos resultados encontrados nas análises estatísticas e nas análises biológicas, realizadas com o intuito de responder o objetivo principal do trabalho, sobre os microRNAs relacionados ao câncer pulmonar, pode-se concluir que os microRNAs encontrados possuem potencial para serem biomarcadores do adenocarcinoma pulmonar e, após a investigação mais profundas destes microRNAs, estes poderiam ser futuramente utilizados em terapêuticas biológicas no tratamento do câncer pulmonar.

REFERÊNCIAS BIBLIOGRÁFICAS

ARGENTON, J. L. P.; ET AL. Árvore de regressão para dados censurados e correlacionados, 2013.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R.; STONE, C. Classification and regression trees. Wadsworth Int. **Group**, v.37, n.15, p.237–251, 1984.

CALIN, G. A.; DUMITRU, C. D.; SHIMIZU, M.; BICHI, R.; ZUPO; ET AL. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. **Proceedings of the national academy of sciences**, v.99, n.24, p.15524–15529, 2002.

CHAMORRO, A. C.; ALBUQUERQUE, E. P. A.; DELLA ROSA, V. A. MICRO-RNAs E CÂNCER: ABORDAGENS E PERSPECTIVAS/MICRORNAs AND CANCER: APPROACHES AND PERSPECTIVES. **Revista de Pesquisa em Saúde**, v.16, n.2, 2016.

CHANG, J. T.-H.; LEE, Y. M.; HUANG, R. S. The impact of the Cancer Genome Atlas on lung cancer. **Translational Research**, v.166, n.6, p.568–585, 2015.

CHEN, J.; BARDES, E. E.; ARONOW, B. J.; JEGGA, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. **Nucleic acids research**, v.37, n.suppl_2, p.W305–W311, 2009.

CIAMPI, A.; ZIGHED, D. A.; CLECH, J. Trees and induction graphs for multivariate response. In: Springer, 2000. **European Conference on Principles of Data Mining and Knowledge Discovery**; resumos. , 2000. 359–366.

CINEGAGLIA, N. D. C. MicroRNoma do carcinoma de pulmão de células não pequenas, 2015.

COLOSIMO, E.; GIOLO, S. Análise de Sobrevida Aplicada. 1ª edição. **São Paulo: Editora Edgard Blücher**, 2006.

CORREA, A. P. A.; QUEIROZ, E.; TREVISAN, N. Teste Qui-Quadrado. Universidade Federal do Paraná. Disponível em: http://www.leg.ufpr.br/lib/exe/fetch.php/disciplinas:ce001:teste_do_quiquadrado.pdf Acesso em, 2006.

DIAS FILHO, J. M.; CORRAR, L.; PAULO, E. Análise multivariada. **São Paulo: Atlas**, 2007.

FERLAY, J.; SOERJOMATARAM, I.; ERVIK, M.; DIKSHIT, R.; ESER, S.; MATHERS, C.; REBELO, M.; PARKIN, D. M.; FORMAN; ET AL. GLOBOCAN 2012 v1. 0, cancer incidence and mortality worldwide: IARC CancerBase No. 11. **Lyon, France: International agency for research on cancer**, v.2016, 2013.

FIGUEIRA, C. V. Modelos de regressão logística. Porto Alegre, 2006. 138p. Dissertação (Mestrado) - Universidade Federal do Rio Grande do Sul.

FONSECA, J. Indução de árvores de decisão, 1994. Tese (Doutorado) - Dissertação de Mestrado, Universidade Nova de Lisboa, Lisboa.

GARSON, D. PA 765 Statnotes: An online textbook. <http://www2.chass.ncsu.edu/garson/pa765/statnote.htm> [Accessed 28.2. 04], 2004.

GLOBOCAN. Lung Cancer. Estimated Incidence, Mortality and Prevalence Worldwide in 2012. <http://gco.iarc.fr/today/fact-sheets-populations>, 2018. Accessed: 2019-28-02.

GUO, C.; YO, S.; JANG, W. Evaluating predictive accuracy of survival models with PROC PHREG. In: SAS, 2018. ; resumos. , 2018.

HARRELL JR, F. The PHGLM procedure. SUGI Supplemental Library User's Guide. **SAS (Statistical Analysis System) Institute Inc., Version**, v.5, 1986.

HERBST, R. S.; HEYMACH, J. V.; LIPPMAN, S. M. Lung Cancer. **New England Journal of Medicine**, v.359, n.13, p.1367–1380, 2008, PMID: 18815398.

HOSMER JR, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. John Wiley & Sons, 2013. 398v.

HOTHORN, T.; ZEILEIS, A. partykit: A Modular Toolkit for Recursive Partytioning in R. **Journal of Machine Learning Research**, v.16, n.118, p.3905–3909, 2015.

HUNT, E. B.; MARIN, J.; STONE, P. J. Experiments in induction, 1966.

INCA. Instituto Nacional de Câncer José Alencar Gomes da Silva. Estimativa 2018: Incidência de Câncer no Brasil. Rio de Janeiro. <https://www.inca.gov.br/publicacoes/livros/estimativa-2018-incidencia-de-cancer-no-brasil>, 2018. Accessed: 2019-28-02.

IORIO, M. V.; CROCE, C. M. MicroRNAs in cancer: small molecules with a huge impact. **Journal of clinical oncology**, v.27, n.34, p.5848, 2009.

KALBFLEISCH, J. D.; PRENTICE, R. L. **The statistical analysis of failure time data**. John Wiley & Sons, 2011. 360v.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, v.53, p.457–481, 1958.

KASSAMBARA, A.; KOSINSKI, M. **survminer: Drawing Survival Curves using 'ggplot2'**, 2019. R package version 0.4.4.

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

LEE, Y. S.; DUTTA, A. MicroRNAs in cancer. **Annual Review of Pathological Mechanical Disease**, v.4, p.199–227, 2009.

LIU, B.; LI, J.; CAIRNS, M. J. Identifying miRNAs, targets and functions. **Briefings in bioinformatics**, v.15, n.1, p.1–19, 2012.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **The annals of mathematical statistics**, p.50–60, 1947.

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. **Cancer Chemother Rep**, v.50, p.163–170, 1966.

MEDINA, P. P.; SLACK, F. J. microRNAs and cancer: an overview. **Cell cycle**, v.7, n.16, p.2485–2492, 2008.

MESSENGER, R.; MANDELL, L. A modal search technique for predictive nominal scale multivariate analysis. **Journal of the American statistical association**, v.67, n.340, p.768–772, 1972.

MILBORROW, S. **rpart.plot: Plot “rpart”Models: An Enhanced Version of “plot.rpart”**, 2019. R package version 3.0.7.

MORGAN, J. N.; MESSENGER, R. C. THAID, a sequential analysis program for the analysis of nominal scale dependent variables, 1973.

MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. **Journal of the American statistical association**, v.58, 1963.

NAKANO, K.; MIKI, Y.; HATA, S.; EBATA, A.; TAKAGI, K.; MCNAMARA, K. M.; SAKURAI, M.; MASUDA, M.; HIRAKAWA, H.; ISHIDA, T.; ET AL. Identification of androgen-responsive microRNAs and androgen-related genes in breast cancer. **Anticancer research**, v.33, n.11, p.4811–4819, 2013.

NIKITOVIC, D.; BERDIAKI, A.; SPYRIDAKI, I.; KRASANAKIS, T.; TSATSAKIS, A.; TZANAKAKIS, G. N. Proteoglycans—biomarkers and targets in cancer therapy. **Frontiers in endocrinology**, v.9, p.69, 2018.

PEARSON, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v.50, 1900.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, p.81–106, 1986.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2018.

SCHMID, K.; KUWERT, T.; DREXLER, H. Radon in indoor spaces: an underestimated risk factor for lung cancer in environmental medicine. **Deutsches Arzteblatt international**, v.107, n.11, p.181, 2010.

SIEGEL, S.; CASTELLAN JR, N. J. **Estatística não-paramétrica para ciências do comportamento**. Artmed Editora, 1975.

SOBRAL, A. P. B. Previsao de carga horária-Uma nova abordagem por Arvore de Decisao, 2003. Tese (Doutorado) - PUC-Rio.

SONQUIST, J.; BAKER, E.; MORGAN, J. Searching for Structure (Institute for Social Research, University of Michigan). **Ann Arbor**, 1973.

SONQUIST, J. A. Multivariate model building, 1970.

SOUZA, C. D. P. Perfil de expressão de microRNAs e seus alvos moleculares em carcinoma pulmonar, 2016.

TAKAMIZAWA, J.; KONISHI, H.; YANAGISAWA, K.; TOMIDA, S.; OSADA, H.; ENDOH, H.; HARANO, T.; YATABE, Y.; NAGINO; ET AL. Reduced expression

of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. **Cancer research**, v.64, n.11, p.3753–3756, 2004.

THERNEAU, T.; ATKINSON, B. **rpart: Recursive Partitioning and Regression Trees**, 2018. R package version 4.1-13.

THERNEAU, T. M. **A Package for Survival Analysis in S**, 2015. Version 2.38.

TIMOFEEV, R. Classification and regression trees (CART). **Theory and Applications**, 2004.

TSUNODA, T.; TAKASHIMA, Y.; YOSHIDA, Y.; DOI, K.; TANAKA, Y.; FUJIMOTO, T.; MACHIDA, T.; OTA, T.; KOYANAGI, M.; KUROKI, M.; ET AL. Oncogenic KRAS regulates miR-200c and miR-221/222 in a 3D-specific manner in colorectal cancer cells. **Anticancer research**, v.31, n.7, p.2453–2459, 2011.

ULIVI, P.; FOSCHI, G.; MENGOZZI, M.; SCARPI, E.; SILVESTRINI, R.; AMADORI, D.; ZOLI, W. Peripheral blood miR-328 expression as a potential biomarker for the early diagnosis of NSCLC. **International journal of molecular sciences**, v.14, n.5, p.10332–10342, 2013.

VAN ECK, N. A. Statistical analysis and data management highlights of OSIRIS IV. **The American Statistician**, v.34, n.2, p.119–121, 1980.

WANG, J.; ZHANG, K.-Y.; LIU, S.-M.; SEN, S. Tumor-associated circulating microRNAs as biomarkers of cancer. **Molecules**, v.19, n.2, p.1912–1938, 2014.

WILCOXON, F. Individual comparisons by ranking methods. *Biom Bull*, 1945.

YIN, J.-Y.; DENG, Z.-Q.; LIU, F.-Q.; QIAN, J.; LIN; OTHER. Association between mir-24 and mir-378 in formalin-fixed paraffin-embedded tissues of breast cancer. **International journal of clinical and experimental pathology**, v.7, 2014.

YOUNG, L. On randomness in ordered sequences. **The Annals of Mathematical Statistics**, v.12, n.3, p.293–300, 1941.