



Universidade Estadual Paulista “Júlio de Mesquita Filho”
Instituto de Biociências – Câmpus de Botucatu
Programa de Pós-graduação em Biometria



Identificação de biomarcadores utilizando expressão de miRNA-seq e RNA-seq de carcinoma pulmonar de células não pequenas em estadiamento inicial

Bethina da Rocha Camargo

Botucatu
2024

Bethina da Rocha Camargo

Identificação de biomarcadores utilizando expressão de miRNA-seq e RNA-seq de carcinoma pulmonar de células não pequenas em estadiamento inicial

Tese de Doutorado apresentada ao Programa de Pós-graduação em Biometria da Universidade Estadual Paulista “Júlio de Mesquita Filho” como parte dos requisitos necessários para a obtenção do título de Doutora em Biometria.

Orientador: Prof. Dr. Rogério Antonio de Oliveira

Coorientadora: Profa. Dra. Patricia Pintor dos Reis

Botucatu
2024

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP

BIBLIOTECÁRIA RESPONSÁVEL: MARIA CAROLINA A. CRUZ E SANTOS-CRB 8/10188

Camargo, Bethina da Rocha.

Identificação de biomarcadores utilizando expressão de miRNA-seq e RNA-seq de carcinoma pulmonar de células não pequenas em estadiamento inicial / Bethina da Rocha Camargo. - Botucatu, 2024

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu

Orientador: Rogério Antônio de Oliveira

Coorientador: Patrícia Pintor dos Reis

Capes: 10203001

1. Árvores de decisão. 2. Aprendizado - Supervisão.
3. Bioinformática. 4. Algoritmo Florestas Aleatórias.
5. Transcriptoma.

Palavras-chave: Árvore de decisão; Aprendizado supervisionado; Bioinformática; Florestas aleatórias; Transcriptoma.

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

Aos 21 dias do mês de fevereiro do ano de 2024, às 14:00 horas, no(a) Laboratório Didático de Informática I (LDI I), realizou-se a defesa de TESE DE DOUTORADO de BETHINA DA ROCHA CAMARGO, intitulada **Identificação de biomarcadores utilizando expressão de miRNA-seq e RNA-seq de carcinoma pulmonar de células não pequenas em estadiamento inicial**. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. ROGERIO ANTONIO DE OLIVEIRA (Orientador(a) - Participação Presencial) do(a) Departamento de Biodiversidade e Bioestatística / Instituto de Biociências de Botucatu UNESP, Profa. Dra. SILVIA HELENA MODENESE GORLA DA SILVA (Participação Presencial) do(a) Departamento de Agronomia e Recursos Naturais / Faculdade de Ciências Agrárias do Vale do Ribeira - Câmpus de Registro - UNESP, Prof.^a Dr.^a MIRIAM HARUMI TSUNEMI (Participação Presencial) do(a) Departamento de Biodiversidade e Bioestatística / Instituto de Biociências de Botucatu UNESP, Prof. Dr. ROBSON FRANCISCO CARVALHO (Participação Presencial) do(a) Departamento de Biologia Estrutural e Funcional / Instituto de Biociências de Botucatu - Unesp, Profa. Dra. HILDETE PRISCO PINHEIRO (Participação Presencial) do(a) Departamento de Estatística / Instituto de Matemática, Estatística e Computação Científica. Após a exposição pela doutoranda e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, a discente recebeu o conceito final: APROVADA. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Prof. Dr. ROGERIO ANTONIO DE OLIVEIRA



Documento assinado digitalmente

ROGERIO ANTONIO DE OLIVEIRA

Data: 28/02/2024 09:45:25-0300

Verifique em <https://validar.iti.gov.br>

Dedico a Deus por ter me guiado e sustentado ao longo desta jornada.

Agradecimentos

Agradeço, em primeiro lugar, a Deus, por ser o arquiteto do meu destino, meu conselheiro e guia nesta jornada, além de ser meu amparo nos momentos desafiadores.

Meus amados pais, Jarbas e Marlene, me faltam palavras para expressar meu amor e admiração por vocês, desde minha infância, não mediram esforços para me oferecer amor, carinho, educação e qualidade de vida. Meu querido irmão, Neto, obrigada por toda afeição e companheirismo. Meu amor por vocês é imensurável.

Felipe, meu amado noivo, obrigada por todo seu amor, carinho, apoio, atenção, amizade e companheirismos, nesses anos juntos, com toda certeza, foram essenciais para o desenvolvimento desse trabalho. Te amo mil milhões.

Prezado orientador, prof. Dr. Rogério Antonio de Oliveira, expresso minha profunda gratidão por todos os ensinamentos, pelas conversas enriquecedoras, pelos valiosos conselhos, pelas oportunidades concedidas e, sobretudo, por acreditar no meu potencial desde o meu mestrado. A Profa. Dra. Patricia Pintor dos Reis gostaria de estender meus agradecimentos pela atenção e pela colaboração no desenvolvimento deste trabalho.

Em especial, manifesto meu reconhecimento à doutoranda Vanessa Das Graças Pereira De Souza pela colaboração, dedicação e atenção dedicada ao desenvolvimento desse trabalho.

Antecipadamente, expresso minha sincera gratidão à respeitável banca examinadora desta tese de doutorado, agradecendo pela disposição em participar e por todas as possíveis sugestões e contribuições.

Agradeço ao Programa de Pós-Graduação em Biometria e aos professores do Departamento pela disponibilidade e atenção dedicadas aos alunos. Aos Funcionários Arthur e Luiz por sempre estarem dispostos a ajudar. E as instalações oferecidas pelo departamento para a realização dessa pesquisa foram, sem dúvida, fundamentais para o sucesso do meu percurso acadêmico.

Com especial apreço, quero expressar minha sincera gratidão aos meus amigos: Lara Morena, Janielly Matos, Roniel Antonio, Leticia Godoi, Guilherme Rodrigues, Elizabeth Pinto, Juliana Gualberto e Mário Lucas por toda a companhia calorosa, momentos de risos e as valiosas trocas que compartilhamos.

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

Mas tu, quando orares, entra no teu aposento, e fechando a tua porta, ora a teu Pai que está em oculto; e teu Pai, que vê em oculto, te recompensará publicamente.

Mateus 6:6

Resumo

O adenocarcinoma pulmonar (LUAD) e o carcinoma de células escamosas pulmonar (LUSC) representam desafios significativos para a saúde pública global, dada a frequência de ocorrências, elevadas taxas de mortalidade e uma incidência em constante aumento. Diante desse cenário, torna-se urgente priorizar estudos que se concentrem na prevenção e detecção precoce dessas malignidades. O emprego de técnicas avançadas, como RNA-seq e miRNA-seq, emerge como uma contribuição substancial na identificação de biomarcadores potenciais associados a essas enfermidades. A integração de métodos de bioinformática, incluindo técnicas estatísticas e ferramentas biológicas, revela-se fundamental para investigações abrangentes em grandes conjuntos de dados transcriptômicos. O objetivo deste estudo foi identificar potenciais biomarcadores relacionados ao LUAD e LUSC em estadiamento inicial, utilizando técnicas de bioinformática para analisar extensos conjuntos de dados de miRNA-Seq e RNA-Seq, para isso foram realizadas três análises. Na primeira fase da análise deste estudo, ao comparar diferentes abordagens para a análise de expressão diferencial, os resultados apontaram que o teste de Wilcoxon-Mann-Whitney, com correção de Bonferroni, superou o EdgeR e o DESeq2 ao expandir as possibilidades de interpretações biológicas e apresentar maior acurácia na árvore de classificação. As análises II e III, integrando abordagens descritivas, biológicas e estatísticas, resultaram na identificação de importantes potenciais biomarcadores. A análise II foi realizada buscando discriminar o tecido tumoral (T) do tecido normal adjacente ao tumor (N) em LUAD e em LUSC. Para o LUAD o *TGFBR2* demonstrou ter grande importância na discriminação dos tecidos e em LUSC o destaque foram para as famílias de miR-29 e miR-30 junto ao miR-205-5p. Analisando o tempo de vida dos pacientes, as árvores de sobrevivência indicaram o miR-184 e o *FHL1* como importantes para LUAD e o miR-31-5p e o *LYSMD3* para LUSC. Na análise III buscando a discriminação dos subtipos LUSC vs. LUAD, o miR-944 demonstrou as maiores contribuições. Em suma, o presente estudo contribui de maneira significativa para a identificação de potenciais biomarcadores em adenocarcinoma e carcinoma de células escamosas pulmonares em estadiamento iniciais.

Palavras-chave: Transcriptoma, Bioinformática, Aprendizado Supervisionado, Árvore de Decisão, Florestas Aleatórias.

Abstract

Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) represent significant challenges to global public health, given the frequency of occurrences, high mortality rates and a constantly increasing incidence. Given this scenario, it is urgent to prioritize studies that focus on the prevention and early detection of these malignancies. The use of advanced techniques, such as RNA-seq and miRNA-seq, emerges as a substantial contribution to the identification of potential biomarkers associated with these diseases. The integration of bioinformatics methods, including statistical techniques and biological tools, proves to be fundamental for comprehensive investigations into large transcriptomic datasets. The aim of this study was to identify potential biomarkers related to LUAD and LUSC at early stage, using bioinformatics techniques to analyze extensive miRNA-Seq and RNA-Seq, for this three analyzes were carried out. In the first phase of the analysis of this study, when comparing different approaches for differential expression analysis, the results showed that the Wilcoxon-Mann-Whitney test, with Bonferroni correction, outperformed EdgeR and DESeq2 by expanding the possibilities of biological interpretations, and present greater accuracy in the classification tree. Analyzes II and III, integrating descriptive, biological and statistical approaches, resulted in the identification of important potential biomarkers. Analysis II was performed seeking to discriminate tumor tissue (T) from normal tissue adjacent to the tumor (N) in LUAD and LUSC. For LUAD, *TGFBR2* demonstrated great importance in tissue discrimination and in LUSC the emphasis was on the miR-29 and miR-30 families together with miR-205-5p. Analyzing the lifespan of patients, survival trees indicated miR-184 and *FHL1* as important for LUAD and miR-31-5p and *LYSMD3* for LUSC. In analysis III, seeking to discriminate the LUSC vs. LUAD, miR-944 demonstrated the greatest contributions. In summary, the present study contributes significantly to the identification of potential biomarkers in adenocarcinoma and lung squamous cell carcinoma in the early stages of the disease.

Keywords: Transcriptome, Bioinformatics, Supervised Learning, Decision Tree, Random Forests.

Lista de figuras

Figura 1 – Classificação nacional do câncer como causa de morte em idades < 70 anos em 2019. Fonte: OMS.	4
Figura 2 – Tipo mais comum de mortalidade por câncer por país em 2020 entre (A) homens e (B) mulheres. Fonte: GLOBOCAN 2020.	5
Figura 3 – Ilustração da localização do gene na célula, organizados em cadeias de DNA compactadas chamadas cromossomos. Criado pelo próprio autor em BioRender (2022).	13
Figura 4 – Tipos de RNA. Criado pelo próprio autor em BioRender (2022).	17
Figura 5 – Biogênese do miRNA. Adaptado de Inamura e Ishikawa (2016) pelo autor.	19
Figura 6 – A hierarquia do aprendizado. Adaptado de Faceli <i>et al.</i> (2011).	28
Figura 7 – Cronologia dos algoritmos de destaque em árvores de decisão. Criado no Lucidchart.	28
Figura 8 – Métodos de amostragem.	44
Figura 9 – Subamostragem e sobreamostragem para conjuntos de dados desequilibrados.	45
Figura 10 – Ilustração de uma árvore de decisão.	47
Figura 11 – Algoritmo de divisão do CART.	49
Figura 12 – Ilustração Florestas Aleatórias.	59
Figura 13 – Espaço ROC.	64
Figura 14 – Curva ROC.	65
Figura 15 – Fluxograma esquemático da metodologia utilizada e resultados da análise I.	66
Figura 16 – (A) Gráfico de barras com as medianas da expressão dos 10 principais DE-miRNAs em tecidos T vs. N. (B, C) Gráficos de barra indicando os valores de logFC dos 10 principais DE-miRNAs. (D) Diagrama de Venn mostrando a sobreposição entre os resultados obtidos. Retirado de Camargo <i>et al.</i> (2023).	69
Figura 17 – (A) As 10 vias com os menores valores de <i>p</i> identificados pela análise KEGG dos resultados das três metodologias. (B) Diagrama de Venn destacando a existência de sobreposição. Retirado de Camargo <i>et al.</i> (2023).	76
Figura 18 – Histogramas com as acurácias das simulações de árvore de classificação para o teste Wilcoxon-Mann-Whitney, Deseq2 e EdgeR. Retirado de Camargo <i>et al.</i> (2023).	77
Figura 19 – (A) Árvore de decisão com 100% de acurácia para o teste de Wilcoxon-Mann-Whitney. (B) Árvore de decisão com 99% de acurácia para Deseq2. (C) Árvore de decisão com 98% de acurácia para EdgeR. Retirado de Camargo <i>et al.</i> (2023).	78
Figura 20 – Fluxograma esquemático da metodologia utilizada nas análises II e III.	81
Figura 21 – Histograma da idade e dos dias do último acompanhamento - LUAD miRNA-seq.	82

Figura 22 – Histograma da idade e do número de dias até o último acompanhamento - LUAD RNA-seq.	82
Figura 23 – Diagrama de Venn entre os genes-alvos dos DE-miRNAs e os DEGs - LUAD.	88
Figura 24 – Rede regulatória entre DE-miRNA <i>up</i> e DEGs <i>down</i> - LUAD.	91
Figura 25 – Rede regulatória entre DE-miRNA <i>downregulated</i> e DEGs <i>upregulated</i> - LUAD.	92
Figura 26 – Rede de interação proteína-proteína - LUAD.	95
Figura 27 – Histograma das 100 mil acurácias para árvores de classificação dos DE-miRNAs - LUAD.	97
Figura 28 – Árvore de classificação e variáveis mais importantes para os DE-miRNAs - LUAD.	98
Figura 29 – Histograma das 100 mil acurácia para árvores de classificação dos DEGs - LUAD.	98
Figura 30 – Árvore de classificação e variáveis importantes para os DEGs - LUAD.	99
Figura 31 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DE-miRNA - LUAD.	100
Figura 32 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DEGs - LUAD.	100
Figura 33 – Árvore de sobrevivência pós-poda e variáveis mais importantes para dos DE-miRNAs - LUAD.	101
Figura 34 – Árvore de sobrevivência pós-poda e variáveis mais importantes para os DEGs - LUAD.	102
Figura 35 – Histograma da idade e dos dias do último acompanhamento - LUSC miRNA-seq.	108
Figura 36 – Histograma da idade e do número de dias do último acompanhamento - LUSC RNA-seq.	108
Figura 37 – Diagrama de Venn entre genes-alvos dos DE-miRNAs e DEGs - LUSC.	112
Figura 38 – Rede regulatória entre DE-miRNAs <i>up</i> e DEGs <i>down</i> - LUSC.	115
Figura 39 – Rede regulatória entre DE-miRNA <i>down</i> e DEGs <i>up</i> - LUSC.	118
Figura 40 – Rede de interação proteína-proteína - LUSC.	120
Figura 41 – Histograma das 100 mil acurácias para árvores de classificação dos DE-miRNAs - LUSC.	122
Figura 42 – Árvore de classificação e variáveis importantes para os DE-miRNAs - LUSC.	123
Figura 43 – Histograma das 100 mil acurácias para árvores de classificação dos DEGs - LUSC.	123
Figura 44 – Árvore de classificação e variáveis importantes para os DEGs - LUSC.	124
Figura 45 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DE-miRNAs - LUSC.	125
Figura 46 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DEGs - LUSC.	125

Figura 47 – Árvore de sobrevivência pós-poda e variáveis importantes para os DE-miRNAs - LUSC.	126
Figura 48 – Árvore de sobrevivência pós-poda e variáveis mais importantes para os RNAs - LUSC.	128
Figura 49 – Diagrama de Venn entre os genes-alvos dos DE-miRNAs e DEGs - LUAD- LUSC.	135
Figura 50 – Rede regulatória entre DE-miRNA <i>down</i> e DEGs <i>up</i> - LUAD-LUSC.	138
Figura 51 – Rede regulatória entre DE-miRNA <i>up</i> e DEGs <i>down</i>	139
Figura 52 – Rede de interação proteína-proteína - LUAD-LUSC.	140
Figura 53 – Histograma das 100 mil acurácias para árvores de classificação dos DE- miRNAs - LUAD-LUSC.	142
Figura 54 – Árvore de classificação pós-poda para os DE-miRNAs - LUAD-LUSC.	143
Figura 55 – Variáveis importantes para a árvore pós-poda de DE-miRNAs- LUAD-LUSC.	143
Figura 56 – Curva ROC para DE-miRNAs - LUAD-LUSC.	144
Figura 57 – Histograma das 100 mil acurácias para árvore de classificação dos DEGs -LUAD-LUSC.	144
Figura 58 – Árvore de classificação pós-poda para os DEGs - LUAD-LUSC.	145
Figura 59 – Variáveis importantes para os DEGs - LUAD-LUSC.	146
Figura 60 – Curva ROC para DEGs - LUAD-LUSC.	146
Figura 61 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DE-miRNA - LUAD-LUSC.	147
Figura 62 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DEGs - LUAD-LUSC.	148
Figura 63 – MiR-944 em LUSC vs. LUAD.	149

Lista de tabelas

Tabela 1 – Definições de T, N e M para o estadiamento inicial.	11
Tabela 2 – Descrição das variáveis demográficas, de diagnóstico e tratamento presentes nos bancos de dados.	33
Tabela 3 – Informações do banco de dados para a análise I.	34
Tabela 4 – Informações dos bancos de dados para a análise II.	34
Tabela 5 – Informações dos bancos de dados da análise III.	35
Tabela 6 – Tabela de contingência 2×2	36
Tabela 7 – Tabela de contingência obtida no tempo t_j	53
Tabela 8 – Hiperparâmetros dos algoritmos <i>rpart</i> e <i>randomForest</i>	62
Tabela 9 – Exemplo de uma matriz de confusão para um problema com duas classes. . .	63
Tabela 10 – DE-miRNAs identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD miRNA-seq precursores.	68
Tabela 11 – DE-miRNAs identificados com DESeq2 - LUAD miRNA-seq precursores. . .	70
Tabela 12 – DE-miRNAs identificados com EdgeR - LUAD miRNA-seq precursores. . .	70
Tabela 13 – MiRNAs precursores e maduros por metodologia.	71
Tabela 14 – Vias enriquecidas para cada método.	73
Tabela 15 – Lista de caminhos únicos por cada método.	76
Tabela 16 – Características demográficas, diagnóstico e tratamento das amostras dos pacientes dos bancos de dados LUAD miRNA-seq e RNA-seq.	83
Tabela 17 – DE-miRNAs <i>downregulated</i> (↓) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD.	84
Tabela 18 – DE-miRNAs <i>upregulated</i> (↑) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD.	85
Tabela 19 – DEGs <i>upregulated</i> (↑) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD.	86
Tabela 20 – DEGs <i>downregulated</i> (↓) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD.	87
Tabela 21 – As 50 maiores correlações negativas entre os pares - LUAD.	89
Tabela 22 – DE-miRNAs <i>up</i> e DEGs <i>down</i> validados experimentalmente - LUAD. . . .	90
Tabela 23 – Principais vias enriquecidas para os DE-miRNAs <i>up</i> - LUAD.	91
Tabela 24 – Vias enriquecidas para os DEGs <i>down</i> - LUAD.	92
Tabela 25 – DE-miRNAs <i>downregulated</i> e DEGs <i>upregulated</i> validados experimentalmente - LUAD.	93
Tabela 26 – Principais vias enriquecidas para DE-miRNAs <i>down</i> - LUAD.	94
Tabela 27 – Vias enriquecidas para os DEGs <i>up</i> - LUAD.	94
Tabela 28 – Processos biológicos relacionados a Rede PPI - LUAD.	96

Tabela 29 – Resumo do modelo da árvore de sobrevivência para os DE-miRNAs - LUAD.	101
Tabela 30 – Resumo do modelo da árvore de sobrevivência para os DEGs - LUAD. . . .	102
Tabela 31 – Características demográficas, diagnóstico e tratamento das amostras dos pacientes dos bancos de dados LUSC miRNA-seq e RNA-seq.	107
Tabela 32 – DE-miRNAs <i>downregulated</i> (↓) identificados com o teste de Wilcoxon- Mann-Whitney com correção de Bonferroni - LUSC.	109
Tabela 33 – DE-miRNAs <i>upregulated</i> (↑) identificados com o teste de Wilcoxon-Mann- Whitney com correção de Bonferroni - LUSC.	110
Tabela 34 – 200 DEGs mais importantes identificados no teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUSC.	111
Tabela 35 – As 50 maiores correlações negativas entre os pares - LUSC.	113
Tabela 36 – DE-miRNAs <i>up</i> e DEGs <i>down</i> validados experimentalmente - LUSC. . . .	114
Tabela 37 – Principais vias enriquecidas para DE-miRNAs <i>up</i> - LUSC.	116
Tabela 38 – Vias enriquecidas para DEGs <i>down</i> - LUSC.	116
Tabela 39 – DE-miRNAs <i>down</i> e DEGs <i>up</i> validados experimentalmente - LUSC. . . .	117
Tabela 40 – Principais vias enriquecidas para DE-miRNAs <i>down</i> - LUSC.	119
Tabela 41 – Vias enriquecidas para DEGs <i>up</i> - LUSC.	119
Tabela 42 – Processos biológicos relacionados a Rede PPI - LUSC.	121
Tabela 43 – Resumo do modelo da árvore de sobrevivência para os DE-miRNAs - LUSC.	126
Tabela 44 – Resumo do modelo da árvore de sobrevivência para os DEGs - LUSC. . . .	127
Tabela 45 – DE-miRNAs <i>upregulated</i> (↑) identificados com o teste de Wilcoxon-Mann- Whitney com correção de Bonferroni - LUAD-LUSC.	131
Tabela 46 – DE-miRNAs <i>downregulated</i> (↓) identificados com o teste de Wilcoxon- Mann-Whitney com correção de Bonferroni - LUAD-LUSC.	132
Tabela 47 – DEGs <i>downregulated</i> (↓) identificados com o teste de Wilcoxon-Mann- Whitney com correção de Bonferroni - LUAD-LUSC.	133
Tabela 48 – DEGs <i>upregulated</i> (↑) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD-LUSC.	134
Tabela 49 – As 50 menores correlações - LUAD-LUSC.	136
Tabela 50 – DE-miRNAs <i>down</i> e DEGs <i>up</i> validados experimentalmente - LUAD-LUSC.	137
Tabela 51 – Principais Vias enriquecidas para os DE-miRNAs -LUAD-LUSC.	138
Tabela 52 – Vias enriquecidas para os DEGs - LUAD-LUSC	139
Tabela 53 – DE-miRNAs <i>up</i> e DEGs <i>downs</i> validados experimentalmente - LUAD-LUSC.	139
Tabela 54 – Principais vias enriquecidas para DE-miRNAs <i>up</i> e DEGs <i>down</i> - LUAD-LUSC.	140
Tabela 55 – Processos biológicos relacionados a Rede PPI - LUAD-LUSC.	141
Tabela 56 – Resumo do modelo de árvore de classificação para os DE-miRNAs - LUAD- LUSC.	142
Tabela 57 – Resumo do modelo da árvore de sobrevivência para os DEGs - LUAD-LUSC.	145
Tabela S1 – Maiores correlações negativas entre os pares - LUAD.	171

Tabela S2 – Vias enriquecidas para DE-miRNAs <i>up</i> - LUAD.	172
Tabela S3 – Vias enriquecidas para DE-miRNAs <i>down</i> - LUAD.	174
Tabela S4 – Maiores correlações negativas entre os pares - LUSC.	176
Tabela S5 – Vias enriquecidas para DE-miRNAs <i>up</i> - LUSC.	180
Tabela S6 – Vias enriquecidas para DE-miRNAs <i>down</i> - LUSC.	182
Tabela S7 – Processos biológicos relacionados a Rede PPI - LUSC.	185
Tabela S8 – Maiores correlações negativas entre os pares - LUAD-LUSC.	187
Tabela S9 – Vias enriquecidas para DE-miRNAs <i>down</i> -LUAD-LUSC.	189
Tabela S10–Vias enriquecidas para DE-miRNAs <i>up</i> - LUAD-LUSC.	192

Lista de abreviaturas e siglas

NSCLC	Câncer de pulmão de células não pequenas
LUAD	<i>Lung Adenocarcinoma</i>
LUSC	<i>Lung Squamous Cell Carcinoma</i>
RNA	Ácido ribonucleico
miRNA	MicroRNA
RNA-seq	Sequenciamento de RNA
miRNA-seq	Sequenciamento de miRNA
DEGs	Genes diferencialmente expresso
DE-miRNAs	MiRNAs diferencialmente expressos
T	Tecido tumoral
N	Tecido histológico normal adjacente ao tumor
vs.	Versus
SCLC	Câncer de pulmão de pequenas células
TNM	<i>Tumor–Node–Metastasis</i>
NGS	Sequenciamento de próxima geração
TCGA	O atlas do genoma do câncer
mRNA	RNA mensageiro
PPI	Rede de interação proteína-proteína
IA	Inteligência artificial
AM	Aprendizado de máquina
CART	<i>Classification and regression trees</i>
ROC	<i>Receiver Operating Characteristics</i>
FDR	<i>False Discovery Rate</i>

Sumário

	1 INTRODUÇÃO	1
1.1	Objetivos	2
1.2	Apresentação da tese	2
	2 REVISÃO DA LITERATURA	3
2.1	Conceitos biológicos	3
2.1.1	Câncer de pulmão	3
2.1.2	Carcinoma pulmonar de células não pequenas	7
2.1.3	Estadiamento inicial e tratamentos	10
2.1.4	Genoma e mutações	12
2.1.5	Transcriptoma e a busca por respostas no NSCLC	15
2.1.6	RNA e miRNA: Biogênese, Funções e Implicações	16
2.1.7	Genes e MiRNAs em câncer de pulmão	21
2.2	Bioinformática	23
2.2.1	Análise de expressão diferencial	23
2.2.2	Análises biológicas	26
2.2.3	Aplicações de Árvore de Decisão e Florestas Aleatórias	27
	3 BANCOS DE DADOS E MÉTODOS	32
3.1	Bancos de dados	32
3.2	Testes estatísticos para análise das variáveis categóricas	35
3.2.1	Estatística Qui-Quadrado de Pearson	35
3.2.2	Teste exato de Fisher	36
3.3	Análise de expressão diferencial	37
3.4	Análises biológicas	39
3.5	Modelos preditivos	41
3.5.1	Preparação de dados	42
3.5.2	Divisão de dados para treino e teste	43
3.5.3	Problema de desbalanceamento	44
3.5.4	Árvore de decisão	46
3.5.5	CART	47
3.5.6	Florestas Aleatórias	58
3.5.7	Hiperparâmetros	61
3.5.8	Avaliação de modelos preditivos	62
	4 RESULTADOS E DISCUSSÕES	66

4.1	Análise I: comparação de métodos de análise de expressão diferencial . .	66
4.2	Análises II e III	80
4.2.1	Análise de expressão diferencial, biológica e estatística de LUAD	81
4.2.2	Análise de expressão diferencial, biológica e estatística de LUSC	106
4.2.3	Análise de expressão diferencial, biológica e estatística de LUSC vs. LUAD	130
5	CONCLUSÃO	151
	Referências	153
6	MATERIAL SUPLEMENTAR	170

1 Introdução

O câncer de pulmão é uma preocupação global significativa, classificado como o segundo tipo de câncer mais diagnosticado e a principal causa de morte em todo o mundo (SUNG *et al.*, 2021). Este câncer, tradicionalmente associado ao tabagismo e à exposição passiva ao tabaco (ÖBERG *et al.*, 2011; YANG *et al.*, 2021), tem visto um aumento na incidência entre não fumantes (SUN; SCHILLER; GAZDAR, 2007; DUBIN; GRIFFIN, 2020). O diagnóstico tardio é comum entre os pacientes de câncer de pulmão, o que reduz suas opções de tratamento e, consequentemente, sua sobrevivência (SCHABATH; COTE, 2019).

O carcinoma pulmonar de células não pequenas (NSCLC - *Non Small Cell Lung Cancer*) é o principal tipo de câncer de pulmão e destaca-se por representar uma classe heterogênea de tumores que inclui diferentes subtipos histológicos (PIKOR *et al.*, 2013), sendo os mais comuns o adenocarcinoma (LUAD - *Lung Adenocarcinoma*) e o carcinoma de células escamosas (LUSC - *Lung Squamous Cell Carcinoma*). As doenças LUAD e LUSC diferem significativamente em termos de prognóstico e características moleculares, o que resulta em diferentes abordagens terapêuticas (CHEN *et al.*, 2017a).

Apesar das terapias atuais, as taxas de mortalidade relacionadas ao câncer de pulmão permanecem elevadas. Portanto, este estudo visa a identificação de potenciais biomarcadores específicos clinicamente aplicáveis para detecção precoce em LUAD e LUSC (KADARA *et al.*, 2021), visando melhorar a sobrevivência do paciente.

Os estudos do transcriptoma, incluindo RNAs (ácido ribonucleico) e os miRNAs (microRNAs), desempenham um papel crucial na pesquisa do câncer (RHODES; CHINNAIYAN, 2005). Essas moléculas têm potencial como biomarcadores diagnósticos, prognósticos e preditivos robustos. O sequenciamento de RNA (RNA-seq) e o sequenciamento de miRNA (miRNA-seq) são técnicas-chave que permitem a investigação aprofundada do transcriptoma.

Uma prática comum e fundamental, para entender a modulação da expressão gênica e suas implicações nas vias biológicas do câncer, envolve a identificação de genes diferencialmente expressos (DEGs - *differential gene expression*) e miRNAs diferencialmente expressos (DE-miRNAs - *differentially expressed miRNAs*) em condições específicas, como o tecido tumoral (T) em comparação com o tecido normal adjacente (N).

No contexto da explosão de dados genômicos, a bioinformática desempenha um papel

crucial para a realização de análises em extensos conjuntos de dados de sequenciamento de miRNA e RNA de pacientes com LUAD ou LUSC em estadiamento inicial. Este estudo emprega a identificação de miRNAs e RNAs diferencialmente expressos, juntamente com técnicas de bioinformática que incluem metodologias biológicas, estatística e de aprendizado de máquina.

O intuito é identificar miRNAs e RNAs que possam atuar como biomarcadores de detecção precoce. Os resultados apresentados enfatizam a relevância da bioinformática na análise abrangente de RNAs e miRNAs em grandes conjuntos de dados de pacientes com LUAD ou LUSC. Assim, esta tese visa contribuir para a compreensão dos subtipos LUAD e LUSC ao nível molecular precoce e melhorar as opções de diagnóstico e tratamento, com impacto positivo na qualidade de vida e na sobrevivência dos pacientes.

1.1 Objetivos

Objetivo geral

Identificar potenciais biomarcadores relacionados ao LUAD e LUSC em estadiamento inicial, utilizando técnicas de bioinformática para analisar extensos conjuntos de dados de miRNA-seq e RNA-seq.

Objetivos Específicos

- **Análise I:** Comparar as ferramentas EdgeR, Deseq2 e o teste de Wilcoxon-Mann-Whitney para a identificação de DE-miRNAs.
- **Análise II:** Identificar RNAs e miRNAs de relevância significativa na discriminação entre os tecidos T vs. N e na sobrevivência dos pacientes.
- **Análise III:** Identificar RNAs e miRNAs de importância significativa para distinguir os subtipos LUSC vs. LUAD.

1.2 Apresentação da tese

O Capítulo 1 contém a introdução e a descrição dos objetivos. Uma revisão bibliográfica é apresentada no Capítulo 2, que detalha todo o referencial teórico a respeito dos conceitos biológicos e de bioinformática. O Capítulo 3 traz as descrições dos bancos de dados, com as metodologias utilizadas. O Capítulo 4 apresenta e detalha todos os resultados encontrados. Por fim, têm-se as considerações finais no Capítulo 5, que resume as principais conclusões encontradas na pesquisa e o Capítulo 6 apresenta o material suplementar.

2 Revisão da Literatura

Neste capítulo será conduzida uma revisão abrangente da literatura, explorando os fundamentos biológicos relacionados ao câncer de pulmão e seus subtipos. Abordar-se-á a importância crucial do diagnóstico em estadiamento inicial, aprofundando-se na compreensão do genoma e transcriptoma associados ao câncer de pulmão. Além disso, será apresentada uma análise detalhada sobre o papel significativo dos RNAs e miRNA nesse contexto. A revisão incluirá também uma exploração minuciosa da bioinformática, destacando diversas técnicas e ferramentas, tanto biológicas quanto estatísticas, além das aplicações do aprendizado de máquina, essenciais para a análise de dados genômicos. Essa abordagem estabelecerá uma base sólida, integrando os aspectos biológicos com as ferramentas computacionais pertinentes, para compreender os processos moleculares envolvidos no câncer de pulmão e contextualizar a relevância dessas informações para o desenvolvimento de terapias personalizadas.

2.1 Conceitos biológicos

Nesta seção serão apresentados os principais aspectos relacionados ao câncer de pulmão, incluindo dados de incidência e fatores de risco. Em seguida, abordar-se-ão detalhes do NSCLC, com destaque para os subtipos LUAD e LUSC em estadiamento inicial. Adicionalmente, serão explorados os conceitos fundamentais relacionados ao genoma e transcriptoma, delineando a estrutura do DNA e sua interação com os processos de expressão gênica. Para contextualizar a análise subsequente, serão fornecidas definições e informações pertinentes sobre os diferentes tipos de RNAs, incluindo mensageiros (mRNA) e miRNAs. Esta base biológica é essencial para a compreensão integral das análises genômicas e de expressão gênica, abordadas nas seções subsequentes.

2.1.1 Câncer de pulmão

Estimativas da Organização Mundial da Saúde (OMS ou WHO - *World Health Organization*), revelam que o câncer ocupou o posto de primeira ou segunda principal causa de morte em pessoas com menos de 70 anos em 112 dos 183 países ao redor do mundo no ano de 2019 (OMS, 2020). Essa tendência é visualmente representada na Figura 1.

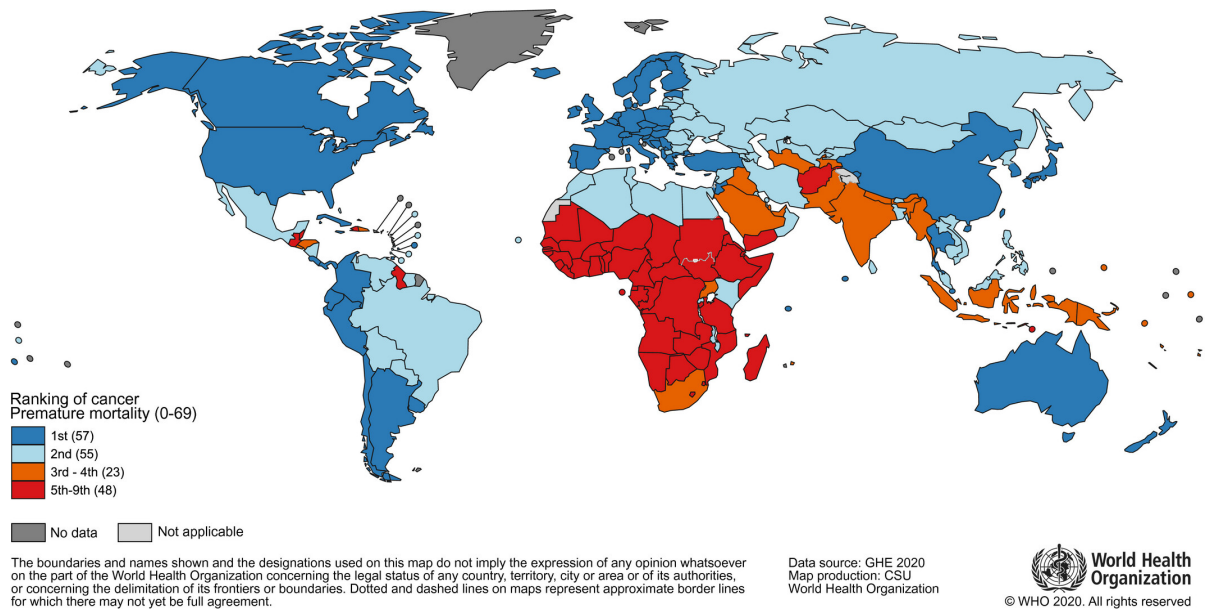


Figura 1 – Classificação nacional do câncer como causa de morte em idades < 70 anos em 2019.
Fonte: OMS.

O câncer de pulmão é uma neoplasia maligna que pode se originar nas células do epitélio brônquico, bronquíolos, alvéolos e glândulas mucosas brônquicas. Caracteriza-se pela tendência a recorrer após tratamento, pela capacidade de metastatizar e por abranger diversos tipos histológicos (ZAMAY *et al.*, 2017).

De acordo com dados recentes do Observatório Global do Câncer (GLOBOCAN - *Global Cancer Observatory*), o câncer de pulmão é o segundo tipo de câncer mais comumente diagnosticado em todo o mundo, com 2,2 milhões de novos casos registrados em 2020. No entanto, ele se destaca por ser a principal causa de óbito, respondendo por um em cada cinco falecimentos (18%). Quando consideramos a análise por gênero, o câncer de pulmão é a principal causa de morte em homens, enquanto nas mulheres ocupa a terceira posição, logo após o câncer de mama e colorretal (SUNG *et al.*, 2021). Esses dados são apresentados visualmente na Figura 2.

Em nações desenvolvidas, como os Estados Unidos da América (EUA), foi relatado que aproximadamente 46.500 indivíduos de origem hispânica ou latina perderam suas vidas devido ao câncer. Notavelmente, o câncer de pulmão se destaca como a principal causa de morte, respondendo por 13% do total de óbitos. É importante ressaltar que cerca de 80% dos casos de câncer de pulmão neste grupo populacional estão diretamente relacionados ao consumo de tabaco (MILLER *et al.*, 2021).

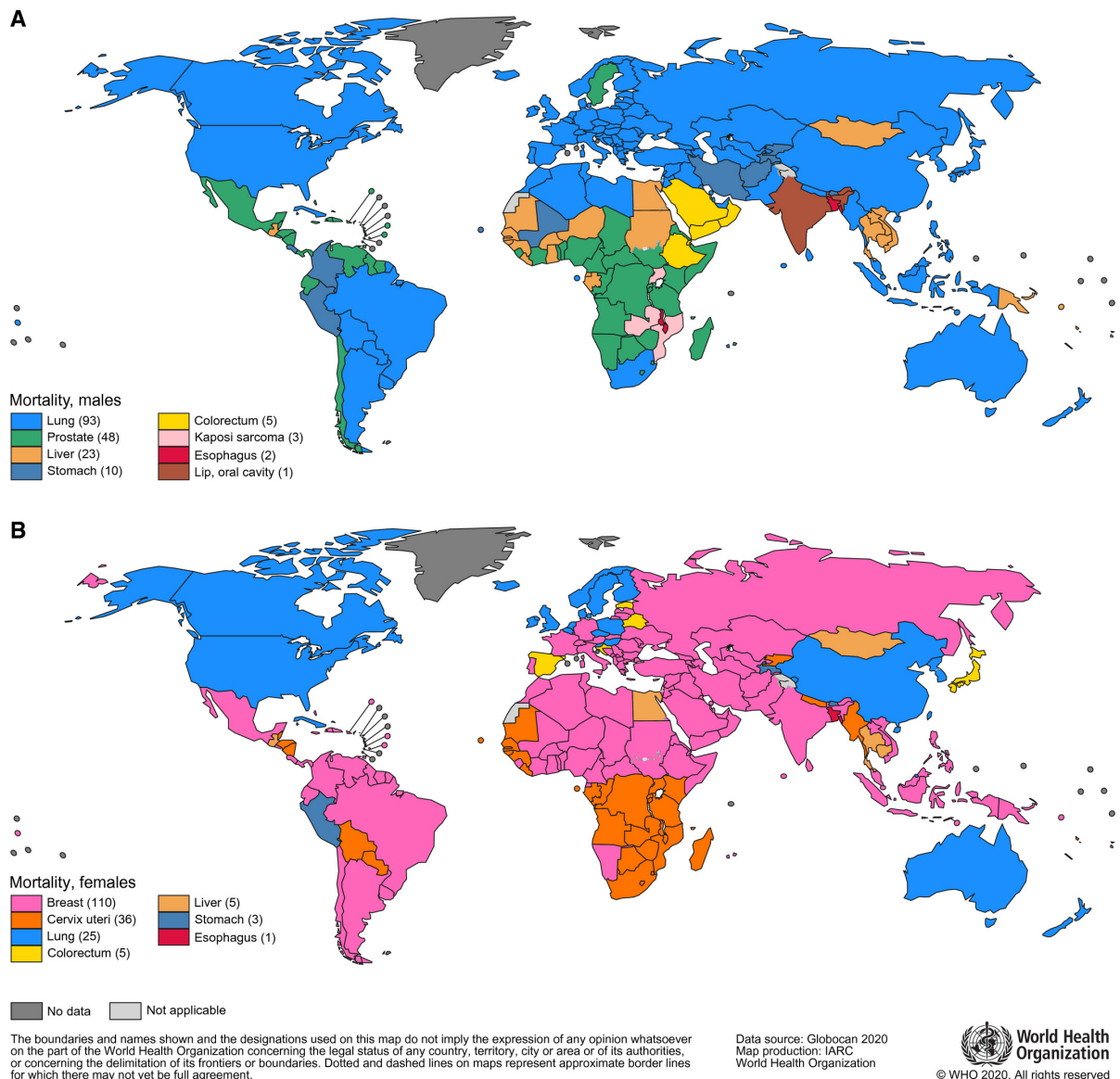


Figura 2 – Tipo mais comum de mortalidade por câncer por país em 2020 entre (A) homens e (B) mulheres. Fonte: GLOBOCAN 2020.

Segundo o Instituto Nacional de Câncer (INCA, 2022), o Brasil está previsto para enfrentar o diagnóstico de mais de 2 milhões de novos casos de câncer até o ano de 2025. Notavelmente, as regiões sul e sudeste do país concentram aproximadamente 70% dessa incidência. No que se refere ao câncer de pulmão, ele se posiciona como o quinto tipo mais comum de tumor maligno, representando cerca de 4,6% dos casos diagnosticados.

O diagnóstico do câncer de pulmão frequentemente se atrasa devido à ausência de sintomas nas fases iniciais da doença, resultando em um diagnóstico tardio. Esse atraso pode levar a um cenário onde, no momento do diagnóstico, o paciente já pode ter comprometimento regional dos linfonodos, com potencial para metástase ou até mesmo metástases a distância, reduzindo significativamente as opções terapêuticas disponíveis e, consequentemente, impactando negativamente a sobrevivência dos pacientes. Portanto, a detecção precoce desempenha um papel

crucial na melhoria das perspectivas de tratamento e na qualidade de vida dos pacientes com câncer de pulmão.

De acordo com [Islam et al. \(2015\)](#), os avanços na medicina têm contribuído para um aumento nas taxas de sobrevivência de pacientes com câncer de pulmão em um período de um ano. No entanto, é importante notar que a taxa de sobrevivência geral em um período de cinco anos permanece relativamente baixa.

O estudo realizado por [Cruz, Tanoue e Matthay \(2011\)](#) enfatiza a importância do diagnóstico precoce em pacientes com câncer de pulmão. Eles destacam que a taxa de sobrevivência em cinco anos para pacientes com câncer de pulmão no estadiamento I, quando tratado precocemente, é consideravelmente melhor do que para pacientes diagnosticados em estadiamentos mais avançados (II a IV). Isso ressalta a necessidade de detectar a doença em seus estadiamentos iniciais para melhorar as perspectivas de sobrevivência.

Comorbidades podem exercer um impacto significativo na sobrevivência de pacientes com câncer, conforme observado por [Lüchtenborg et al. \(2012\)](#). Entre as condições mais frequentemente associadas a pacientes diagnosticados com câncer de pulmão estão aquelas relacionadas ao envelhecimento e ao tabagismo, tais como doenças cardiovasculares, doença pulmonar obstrutiva crônica e outras neoplasias ([JANSSEN-HEIJNEN et al., 1998](#)).

É importante ressaltar que o principal fator de risco para o desenvolvimento do câncer de pulmão é o tabagismo ativo e a exposição passiva ao tabaco, como indicado por [Öberg et al. \(2011\)](#). Estima-se que aproximadamente 80-90% dos casos de câncer de pulmão no mundo tenham essa associação ([CORRALES et al., 2020](#)). Isso realça a preponderância do tabagismo como um fator contribuinte significativo para o câncer de pulmão em escala global. Estudos recentes, como o de [Thandra et al. \(2021\)](#), enfatizam que mais de 80% dos casos de câncer de pulmão na região ocidental do mundo estão associados ao tabagismo.

[Sun, Schiller e Gazdar \(2007\)](#) fornecem *insights* sobre o câncer de pulmão em não fumantes, destacando que aproximadamente 25% dos casos de câncer de pulmão globalmente não podem ser atribuídos ao uso de tabaco. Recentemente, [Dubin e Griffin \(2020\)](#) corroboraram essa informação. Além disso, [Boffetta et al. \(2001\)](#) sugerem um aumento nas taxas de câncer de pulmão em não fumantes. [Couraud et al. \(2012\)](#) ressaltam que o câncer de pulmão em nunca fumantes, definidos como aqueles que consumiram menos de 100 cigarros em toda a sua vida, é mais prevalente entre as mulheres.

Entretanto, as evidências quanto às taxas de mortalidade entre não fumantes e fumantes diagnosticados com câncer de pulmão apresentam resultados variados. [Kawaguchi et al. \(2010\)](#) sugerem taxas de mortalidade mais baixas na população não fumante. Por outro lado, o estudo de [Peto et al. \(2000\)](#) mostrou diferenças significativas, indicando um possível aumento da mortalidade em alguns casos.

Existem diversos fatores de risco associados ao câncer de pulmão em não fumantes,

embora até o momento nenhum deles seja reconhecido como causa direta da doença, conforme discutido por [Dubin e Griffin \(2020\)](#). Estes fatores incluem a idade ([THUN et al., 2006](#); [DIAS et al., 2017](#)), exposição ao fumo passivo ([NYBERG et al., 1998](#)), exposições ambientais ([YU et al., 2006](#)), exposição ao radônio ([SANDLER et al., 2006](#)), predisposição genética ([ALBERG; BROCK; SAMET, 2005](#)), doença pulmonar ([ALBERG; BROCK; SAMET, 2005](#)), presença de vírus oncogênicos ([GORLOVA et al., 2007](#)) e fatores hormonais, como os estrogênios ([PARKIN et al., 2005](#)).

Além disso, várias substâncias tóxicas são identificadas como tendo potencial carcinogênico para o câncer de pulmão, incluindo amianto, hidrocarbonetos aromáticos policíclicos, arsênico, compostos hexavalentes de cromo, entre outros ([SCHMID; KUWERT; DREXLER, 2010](#)).

Dado que os pacientes diagnosticados com câncer de pulmão podem ser fumantes, ex-fumantes ou mesmo não fumantes, é inegável que o câncer de pulmão representa um sério desafio em termos de saúde pública. Portanto, a prioridade deve ser dada ao desenvolvimento de estratégias eficazes para o controle, prevenção e diagnóstico precoce, bem como para o tratamento dessa doença.

2.1.2 Carcinoma pulmonar de células não pequenas

Os cânceres geralmente recebem nomes com base nos órgãos ou tecidos onde se originam, ou podem ser categorizados conforme o tipo de célula que deu origem a eles. Existem cinco grupos principais de câncer: carcinoma, sarcoma, leucemia, linfoma, mieloma, bem como cânceres que afetam o cérebro e a medula espinhal ([KAUR; KUMAR; KAUSHIK, 2022](#)).

O carcinoma é o tipo mais comum e pode se desenvolver a partir de diversos tipos de células epiteliais, que revestem tanto as superfícies internas quanto externas do corpo. Este grupo de cânceres possui subtipos específicos, sendo os mais comuns o adenocarcinoma, carcinoma basocelular, carcinoma epidermoide e carcinoma "*in situ*".

Quando abordamos a urgência de controlar, prevenir e tratar o câncer de pulmão, é fundamental começar pela compreensão de seus diferentes tipos. Existem dois principais tipos de câncer de pulmão: NSCLC, que representa a grande maioria dos casos (cerca de 80-85% de todos os cânceres de pulmão), e o carcinoma pulmonar de células pequenas (SCLC - *Small Cell Lung Cancer*), que corresponde a aproximadamente 15-20% dos casos de câncer de pulmão ([ETTINGER, 2004](#)).

A classificação patológica dos cânceres de pulmão pode ser detalhada no estudo realizado por [Travis, Brambilla e Riely \(2013\)](#). Compreender diferentes tipos de câncer de pulmão é essencial para o desenvolvimento de estratégias de diagnóstico e tratamento eficazes.

O SCLC é notório por sua agressividade, caracterizado como um carcinoma neuroendócrino que apresenta rápido crescimento, altas taxas de metástases locais e à distância em

estadiamentos iniciais. Este tipo de câncer está intrinsecamente ligado à exposição a carcinógenos do tabaco, sendo predominantemente diagnosticado em fumantes ou ex-fumantes, resultando em taxas de sobrevivência menores para os pacientes afetados ([RUDIN *et al.*, 2021](#)).

Por outro lado, o NSCLC se destaca como o tipo mais comum de câncer de pulmão, representando uma categoria diversificada de tumores. Embora o tabagismo permaneça como o principal fator de risco para o desenvolvimento desta doença, a exposição ao radônio e à poluição do ar também desempenham papéis significativos na sua etiologia ([GRIDELLI *et al.*, 2015](#)). Compreender essas distinções entre os tipos de câncer de pulmão é crucial para o diagnóstico e tratamento eficazes.

Pacientes com NSCLC comumente apresentam uma variedade de sinais e sintomas, incluindo tosse persistente, hemoptise, dor torácica, dificuldade respiratória, inchaço de gânglios linfáticos no pescoço ou clavícula, perda de peso não explicada, dor devido a metástases, fadiga e, ocasionalmente, febre ([XING *et al.*, 2019](#)). No entanto, é importante notar que, em muitos casos, especialmente nas fases iniciais da doença, o câncer de pulmão pode se manifestar de forma assintomática.

O NSCLC engloba diversos subtipos histológicos, como destacado por [Pikor *et al.* \(2013\)](#). A classificação dos subtipos histológicos do NSCLC é baseada principalmente em:

- **LUAD:** Este subtipo frequentemente se origina nas glândulas mucosas, é frequentemente encontrado nas áreas periféricas dos pulmões e representa cerca de 40% de todos os cânceres de pulmão. É notável por ser o tipo mais comum diagnosticado em indivíduos que nunca fumaram e em mulheres ([MYERS; WALLEN, 2022](#); [ZAPPA; MOUSA, 2016](#); [CHEN *et al.*, 2017a](#)).
- **LUSC:** Esse subtipo representa aproximadamente 30% de todos os cânceres de pulmão ([SIEGEL; NAISHADHAM; JEMAL, 2013](#)) e é frequentemente diagnosticado em estadiamentos avançados, com um prognóstico desfavorável ([GAO *et al.*, 2020](#)). Caracteriza-se por ser a segunda neoplasia pulmonar mais comum entre os fumantes, causando lesões nas vias aéreas, predominando nas partes centrais do pulmão ([ANUSEWICZ; ORZECOWSKA; BEDNAREK, 2020](#)) e é mais comum em homens ([CHEN *et al.*, 2017a](#)).

Além dos subtipos previamente mencionados, o NSCLC inclui uma variedade de outros subtipos histológicos, tais como o carcinoma de células grandes, o carcinoma adenoescamoso, o carcinoma sarcomatoide, o carcinoma de glândula salivar, o tumor carcinoide e o carcinoma não classificado.

As estratégias terapêuticas adotadas variam consideravelmente entre o LUAD e o LUSC, conforme destacado por ([RELLI *et al.*, 2019](#)). Isso se deve à natureza distinta desses subtipos, onde o LUAD é caracterizado por ser um carcinoma de células epiteliais alveolares com diferenciação adenoide, muitas vezes associado à produção de muco. Por outro lado, o LUSC é um carcinoma

basocelular epitelial que exibe queratinização e pontes intercelulares, conforme descrito por Travis (2020).

Além das diferenças histológicas, o LUAD e o LUSC apresentam distinções em termos de prognóstico e características moleculares (ANUSEWICZ; ORZECCHOWSKA; BEDNAREK, 2020; TIAN, 2017; FARUKI *et al.*, 2017; GENTLES *et al.*, 2015; CHARKIEWICZ *et al.*, 2017; WILKERSON *et al.*, 2012). Em geral, o LUAD tende a ter um prognóstico ligeiramente melhor do que o LUSC, mas o prognóstico específico pode variar com base em fatores individuais, como estadiamento do câncer e resposta ao tratamento. No nível molecular, esses subtipos diferem em várias características:

- **Assinaturas Mutacionais:** LUAD e LUSC exibem perfis de mutações genéticas distintos. Enquanto mutações no gene *EGFR* são mais comuns em LUAD, mutações em genes como *TP53* são mais prevalentes em LUSC.
- **Número de Cópias do DNA:** Alterações no número de cópias do DNA, como ganhos ou perdas genômicas, variam entre os subtipos, afetando a expressão gênica e a função celular de maneiras específicas.
- **Perfis de Expressão Gênica:** LUAD e LUSC demonstram padrões únicos de expressão gênica, refletindo suas diferenças biológicas. Esses perfis influenciam diretamente o comportamento celular e as respostas aos tratamentos.

Essas nuances moleculares são essenciais para orientar estratégias terapêuticas mais direcionadas e personalizadas, destacando a importância da compreensão das características específicas de cada subtipo de câncer de pulmão. A diferenciação entre esses subtipos geralmente requer uma avaliação visual minuciosa por um patologista experiente, sendo que, quando necessário, podem ser realizados testes moleculares para confirmar o diagnóstico.

O estudo conduzido por Hou *et al.* (2010) revela que os genes associados ao LUAD estão predominantemente envolvidos nas junções apertadas com moléculas de adesão celular, enquanto os genes relacionados ao LUSC estão mais associados à comunicação celular.

Além disso, Kleczko *et al.* (2019) enfatizam que duas notáveis evoluções marcaram a última década e oferecem promissoras perspectivas para o tratamento do NSCLC. A primeira é a identificação de múltiplos condutores oncogênicos, e a segunda diz respeito ao avanço das abordagens de imunoterapia. Essas inovações demonstraram eficácia sendo aprovadas para uso em subconjuntos específicos de pacientes com LUAD e LUSC.

O LUAD e o LUSC demonstram notáveis distinções em níveis moleculares, patológicos e clínicos. Devido a essas diferenças substanciais, há uma carência de procedimentos de diagnóstico precoce, triagem eficaz e intervenções personalizadas e específicas para esses subtipos, como discutido por Herbst e Jr (2003).

Consequentemente, persiste a necessidade de pesquisas contínuas com o propósito de identificar os mecanismos subjacentes tanto ao LUAD quanto ao LUSC. Essa busca visa o desenvolvimento de biomarcadores valiosos que aprimorem o diagnóstico e orientar intervenções terapêuticas de maneira mais precisa e eficaz.

2.1.3 Estadiamento inicial e tratamentos

A detecção precoce no câncer de pulmão é de suma importância, dadas as características desafiadoras e frequentemente fatais dessa doença. Os sintomas muitas vezes permanecem ocultos até que a doença alcance estadiamentos avançados. Portanto, a identificação precoce desempenha um papel crucial na luta contra o câncer de pulmão, proporcionando melhores perspectivas de tratamento e uma qualidade de vida prolongada para os pacientes.

A avaliação do estadiamento do câncer de pulmão segue uma abordagem sistemática, baseada em quatro fatores fundamentais: a localização do tumor primário, as dimensões e a extensão dos tumores (T), o envolvimento dos linfonodos (N) e a presença ou ausência de metástases à distância (M). Esse sistema, conhecido como sistema TNM (*Tumor Node Metastasis*), é desenvolvido e mantido pelo *American Joint Committee on Cancer* (AJCC) e pela *Union for International Cancer Control* (UICC).

A classificação do estadiamento visa fornecer uma nomenclatura globalmente aceita que descreve a extensão anatômica da doença, permitindo uma abordagem consistente em todo o mundo, para mais detalhes a respeito das classificações podem ser vistos em [Detterbeck et al. \(2017\)](#).

A Tabela 1 oferece descrições detalhadas das classificações do sistema TNM, conforme definido na 8ª edição ([DETTERBECK et al., 2017](#)), especificamente para os estadiamentos iniciais I, IA e IB, que são caracterizados por extensão do tumor entre T1 a T2a, sem envolvimento de gânglios linfáticos (N0) e sem metástase a distância (M0).

Tabela 1 – Definições de T, N e M para o estadiamento inicial.

Extensão do tumor	
T1	tumor $\leq$ 3cm
T1a(MI)	adenocarcinoma minimamente invasivo
T1a	tumor de disseminação superficial em vias aéreas centrais
T1a	tumor $\leq$ 1cm
T1b	1 < tumor $\leq$ 2cm
T1c	2 < tumor $\leq$ 3cm
T2	3 < tumor $\leq$ 5cm ou tumor envolvendo: pleura visceral, brônquio principal (não carina), atelectasia do hilo
T2a	3 < tumor $\leq$ 4cm
Linfonodos Regionais	
N0	sem metástase em linfonodos regionais
NX	linfonodos regionais não podem ser avaliados
Metástase a distância	
M0	sem metástase à distância
MX	metástase à distância não pode ser avaliada

A identificação precoce e a localização dos tumores no câncer de pulmão, especificamente no NSCLC, desempenham um papel crucial na determinação do prognóstico dos pacientes. Quando o câncer é descoberto em seus estadiamentos iniciais e continua localizado, a ressecção cirúrgica oferece um tratamento com taxas de sobrevivência notáveis, variando de 70% a 90% após 5 anos (NESBITT *et al.*, 1995; GOLDSTRAW *et al.*, 2016). Dado que a maioria dos pacientes com câncer de pulmão, especialmente NSCLC, é diagnosticada em estadiamentos avançados da doença (ZHANG; YANG; WANG, 2014), é evidente que o diagnóstico precoce é um componente indispensável para melhorar os resultados clínicos.

De acordo com Ruano-Raviña *et al.* (2020), cerca de 59% dos pacientes diagnosticados com câncer de pulmão no estadiamento I são assintomáticos, enquanto aproximadamente 40% apresentam pelo menos um sintoma. À medida que o câncer progride para estadiamentos mais avançados, como o estadiamento IV, os sintomas mais comuns incluem tosse e dispneia, com 33,9% e 26,7% dos casos, respectivamente.

O tratamento do câncer de pulmão varia dependendo do subtipo histológico e do estadiamento da doença. As opções terapêuticas mais comuns incluem cirurgia, radioterapia e quimioterapia, podendo ser utilizadas de forma isolada ou combinada com outras modalidades terapêuticas, conforme discutido por Huang *et al.* (2017).

Conforme observado por Mak, Li e Minchom (2019), as melhorias na cirurgia torácica, a identificação de alterações genômicas e moleculares, juntamente com os avanços em terapias orientadas por biomarcadores, têm proporcionado benefícios significativos para os pacientes com NSCLC. No entanto, devido à diversidade de tipos histológicos, o câncer de pulmão apresenta uma ampla variedade de subtipos moleculares e patológicos, cada um caracterizado por variações genéticas e epigenéticas¹ nas células, bem como uma combinação única de biomarcadores

¹ Mudanças hereditárias ou reversíveis na expressão gênica que não envolvem alterações na sequência do DNA.

proteicos, como enfatizado por [Zamay *et al.* \(2017\)](#).

Um marco importante no avanço do tratamento do NSCLC foi a identificação de mutações no receptor do fator de crescimento epidérmico (EGFR). Como documentado por [Kobayashi *et al.* \(2005\)](#), essas mutações foram identificadas em amostras de pacientes com NSCLC, representando um divisor de águas na compreensão desse câncer. Curiosamente, as mutações do EGFR são mais prevalentes em mulheres, asiáticos e não fumantes, conforme destacado por [Gazdar \(2009\)](#). Essas mutações estão fortemente associadas à capacidade de resposta positiva à terapia com inibidores de tirosina quinase (*TKI*), conforme evidenciado por [Lynch *et al.* \(2004\)](#). Esses avanços no entendimento das mutações do EGFR revolucionaram o tratamento do NSCLC, abrindo caminho para terapias mais eficazes e permitindo abordagens mais direcionadas.

Portanto, compreender os mecanismos biológicos subjacentes ao desenvolvimento do câncer de pulmão e a identificação de biomarcadores comuns e específicos para os diversos tipos histológicos são fundamentais para garantir diagnósticos precisos, tratamentos eficazes e o desenvolvimento contínuo de terapias inovadoras ([ZAMAY *et al.*, 2017](#)).

Em resumo, os estadiamentos iniciais do câncer de pulmão desempenham um papel crítico na determinação do prognóstico e no sucesso do tratamento. A detecção precoce, a classificação precisa e a compreensão dos mecanismos biológicos subjacentes a esta doença são essenciais para melhorar as perspectivas de tratamento e prolongar a qualidade de vida dos pacientes diagnosticados com câncer de pulmão. Portanto, é fundamental direcionar esforços para o estudo e diagnóstico precoces desses estadiamentos, a fim de reduzir o impacto dessa doença desafiadora e potencialmente letal.

2.1.4 Genoma e mutações

O genoma é considerado a enciclopédia da vida, representando a coleção completa de informações genéticas contidas em um organismo ([ALBERTS *et al.*, 2014](#)). Ele serve como um vasto repositório de dados, determinando todas as características e funções de um ser vivo. Esse conceito fundamental ultrapassa as barreiras das espécies e desempenha um papel vital em diversas áreas da biologia, desde a evolução até a medicina.

O genoma é organizado em estruturas distintas conhecidas como cromossomos ([LANDER; LINTON; BIRREN, 2001](#)). Cada cromossomo abriga genes, as unidades funcionais de informação genética. O componente principal do genoma é o ácido desoxirribonucleico, ou DNA ([WATSON; CRICK, 1953](#)). Essas longas e intrincadas moléculas consistem em quatro bases nitrogenadas: adenina (A), timina (T), citosina (C) e guanina (G), conforme ilustrado na Figura 3. O DNA forma a espinha dorsal do genoma, carregando informações cruciais para a síntese de proteínas e regulando processos celulares vitais. Dessa forma, o genoma não apenas contém genes, mas também abriga sequências não codificantes e elementos regulatórios, desempenhando um papel crucial na regulação da expressão gênica.

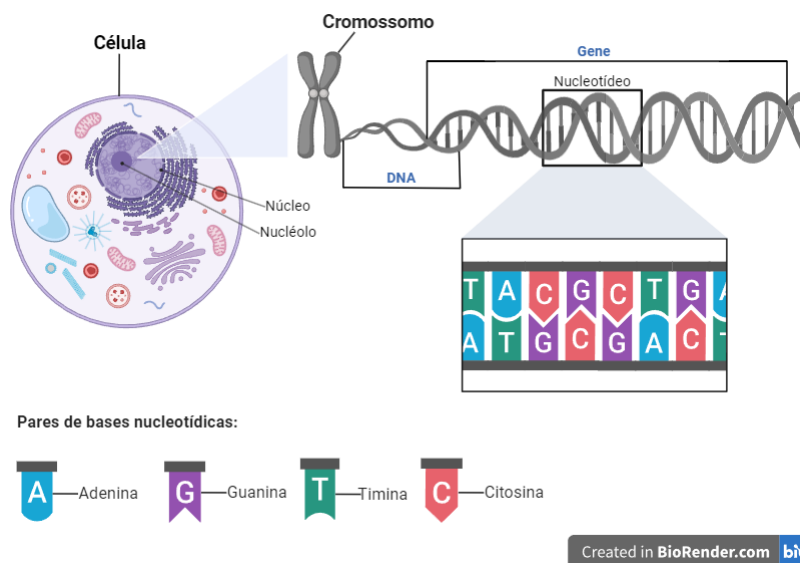


Figura 3 – Ilustração da localização do gene na célula, organizados em cadeias de DNA compactadas chamadas cromossomos. Criado pelo próprio autor em [BioRender](#) (2022).

A variabilidade é uma característica fundamental dos genomas, mesmo dentro da mesma espécie, os genomas podem diferir em termos de sequência de DNA ([SACHIDANANDAM et al., 2001](#)). Essa variabilidade inclui polimorfismos de nucleotídeo único (SNPs - *single nucleotide polymorphism*), variações em uma única base e alterações estruturais mais complexas, como deleções e duplicações. Essa variação genômica é a base da diversidade e evolução das espécies.

O sequenciamento do genoma, uma notável conquista da era genômica, é o processo de determinar a ordem exata das bases de DNA em um genoma ([SHENDURE et al., 2017](#)). A revolução do sequenciamento de próxima geração (NGS - *Next Generation Sequencing*) possibilitou o mapeamento rápido e acessível de genomas inteiros, possibilitando analisar os genomas de organismos de todos os tamanhos, desde microrganismos até seres humanos. A sequência dos nucleotídeos gera a informação genética, e as alterações nessa sequência são reconhecidas como mutações, que podem ou não ter consequências para o indivíduo.

As sequências que contêm informações genéticas transcritas em RNA, para a produção ou não de proteínas, ou seja, sequências codificantes do DNA, são conhecidas como genes. Os genes e as mutações, que eles podem sofrer, são estudados na oncologia, para identificar possíveis marcadores moleculares, que diferenciem uma célula normal de uma célula cancerígena, possibilitando a identificação de um tipo de doença ou estadiamento do câncer.

Estas mutações podem resultar de erros durante algum processo biológico devido a fatores como erro na divisão celular, danos causados no DNA por substâncias nocivas no ambiente ou herança genética. Existem diferentes tipos de mutações no DNA, incluindo mutações de ponto, que envolvem um único nucleotídeo, ou de sequências pequenas de nucleotídeos, que contêm deleções, duplicações ou inserções de novos nucleotídeos.

Um exemplo notável é o gene *EGFR*, em NSCLC, sendo amplamente estudado devido às

mutações de ponto, como L858R, L861Q e G719A/C/S, e deleções no éxon 19, associadas a altas taxas de resposta a inibidores de TKI. Por outro lado, a mudança de ponto T790M e inserção no éxon 20 indicam resistência aos TKIs de primeira e segunda geração (SANTOS; SHEPHERD; TSAO, 2011; PAEZ *et al.*, 2004).

Outras alterações no DNA envolvem trechos maiores e podem incluir rearranjos cromossômicos, deleções e duplicações de segmentos longos de DNA. Um exemplo notável é a inversão que pode ocorrer no cromossomo 2, resultante da fusão do gene Linfoma Anaplásico Quinase (*ALK*) com outros genes, associada à sensibilidade aos TKIs anti-*ALK* em NSCLC (SHAW; SOLOMON, 2011).

As modificações epigenéticas no DNA podem controlar a expressão de um gene, mas as mutações não estão propriamente na sequência do DNA. Por exemplo, a metilação (adição de um grupo metil -CH₃) da região promotora do gene *MGMT* faz com que essa região não seja reconhecida por fatores de transcrição, silenciando o gene (ESTELLER *et al.*, 1999). Esta modificação epigenética no gene de reparo *MGMT* está associada a um prognóstico favorável em pacientes adultos com glioblastoma que recebem temozolomida e outros agentes alquilantes (DONSON *et al.*, 2007). Todos esses tipos de mutações, incluindo mutações de ponto, podem resultar em câncer.

O Projeto Genoma Humano, que ocorreu de 1990 a 2003, trabalhou para identificar a sequência do genoma humano e os genes que ele contém. Estima-se que os humanos possuam entre 20 a 25 mil genes (MARIAN, 2014). Dado que cada célula humana contém aproximadamente 25 mil genes, é importante observar que uma única mutação em um gene geralmente não leva uma pessoa ao desenvolvimento de câncer. O câncer surge quando há um acúmulo de mutações em genes-chave. Esses genes-chave podem ser agrupados em três classes:

- **Proto-oncogenes:** esses genes estão relacionados com o crescimento e a divisão das células normais. No entanto, quando sofrem mutações ou tornam-se mais ativos do que o normal, podem se transformar em oncogenes, desencadeando o crescimento e sobrevivência anormais das células, como, por exemplo, o gene *EGFR*.
- **Supressores de tumor:** estes genes estão envolvidos no controle do crescimento e divisão celular normais. Quando células contêm mutações em genes supressores de tumor, podem se dividir de maneira descontrolada, dando origem ao câncer, como, por exemplo, o *TP53*.
- **Genes de reparo de DNA:** Estes genes desempenham um papel crítico na correção imediata de danos ao DNA. Células com mutações nesses genes têm maior probabilidade de desenvolver mutações adicionais em outros genes e alterações em seus cromossomos. Essas mudanças podem levar a células cancerosas. O gene *BRCA* é um exemplo representativo.

Este capítulo fornece uma base sólida para compreender a importância do genoma e das mutações genéticas, cruciais para o entendimento de como as alterações genéticas estão

relacionadas ao desenvolvimento e progressão do câncer. Essas informações servirão como fundamento para os capítulos subsequentes sobre o transcriptoma, RNA-seq e miRNA-seq, em que se exploram as técnicas e métodos para estudar a expressão gênica e as moléculas de RNA.

2.1.5 Transcriptoma e a busca por respostas no NSCLC

O transcriptoma é o conjunto de moléculas de RNA, que inclui não apenas os RNAs mensageiros, ribossomais e transportadores, mas também os miRNAs e outros tipos de RNA, todos produzidos por meio da transcrição do DNA. Essas moléculas desempenham um papel fundamental na regulação da expressão gênica, variando em diferentes estádios de desenvolvimento, condições fisiológicas e em resposta a estímulos externos (MENCK; SLUYS, 2017).

Duas técnicas de NGS se destacam na pesquisa do transcriptoma, cada uma com seus propósitos específicos:

- **RNA-seq:** Essa técnica é empregada para uma análise global do transcriptoma, abrangendo tanto os transcritos codificadores de proteínas (RNAm) quanto os transcritos não codificadores (RNAs não codificadores). O RNA-seq permite a quantificação e análise de todos os tipos de RNA presentes em uma amostra, tornando-se essencial para identificar transcritos expressos, medir níveis de expressão gênica, detectar variantes de *splicing* e investigar alterações na expressão gênica em diferentes contextos ou doenças (WANG; GERSTEIN; SNYDER, 2009).
- **miRNA-seq:** Esta técnica é projetada especificamente para analisar os miRNAs, que são pequenos RNAs não codificadores envolvidos na regulação pós-transcricional da expressão gênica. O miRNA-seq possibilita a identificação e quantificação de miRNAs em uma amostra, bem como a análise da expressão diferencial desses miRNAs em diferentes condições ou doenças. É fundamental para compreender o papel dos miRNAs na regulação gênica e sua relevância em processos biológicos e doenças.

Ambas as técnicas desempenham um papel crucial na pesquisa genômica e biológica. O RNA-seq é uma poderosa ferramenta que permite uma análise abrangente e quantitativa do transcriptoma. No contexto do NSCLC, o RNA-seq é fundamental para identificar genes que estão superexpressos (*upregulated*) ou subexpressos (*downregulated*) no tecido tumoral em comparação com o tecido saudável. Esta análise é essencial para compreender os mecanismos moleculares subjacentes à doença e identificar biomarcadores para diagnóstico e prognóstico.

Por outro lado, o miRNA-seq concentra-se na análise dos miRNAs, que desempenham um papel vital na regulação da expressão gênica e estão associados ao desenvolvimento e progressão do câncer. Estes pequenos RNAs conseguem silenciar genes específicos, desempenhando um

papel crítico na homeostase celular ([AMBROS, 2004](#)). No contexto do NSCLC, o miRNA-seq ajuda a identificar quais miRNAs estão envolvidos na doença e como afetam os processos moleculares subjacentes. A compreensão desses miRNAs é fundamental para desenvolver novas estratégias terapêuticas e identificar possíveis alvos para intervenção.

Essas técnicas não apenas ajudam a compreender os mecanismos moleculares subjacentes ao NSCLC, mas também podem levar à identificação de biomarcadores. Esses biomarcadores podem ser detectados em amostras clínicas, como sangue ou tecido tumoral, e têm o potencial de diagnosticar o câncer em estádios iniciais ou prever a resposta a tratamentos específicos.

Além disso, a pesquisa do transcriptoma desempenha um papel crucial na identificação de alvos terapêuticos para o NSCLC. A compreensão de quais genes estão superexpressos no câncer pode levar ao desenvolvimento de terapias personalizadas direcionadas a esses genes específicos, resultando em tratamentos mais eficazes e com menos efeitos colaterais.

No campo da biologia molecular, o estudo do transcriptoma proporciona uma visão profunda das redes de regulação gênica, essenciais para entender como os genes se comunicam entre si e como as vias metabólicas são orquestradas. Essas informações também são fundamentais para esclarecer os processos de diferenciação celular e desenvolvimento embrionário.

Em resumo, o transcriptoma é uma ferramenta inestimável na pesquisa do NSCLC e de outras doenças. O uso das técnicas de RNA-seq e miRNA-seq promovem avanços significativos na luta contra o câncer e outras doenças devastadoras. O estudo do transcriptoma é, portanto, um campo de pesquisa repleto de potencial e relevância.

2.1.6 RNA e miRNA: Biogênese, Funções e Implicações

O RNA (ácido ribonucleico) desempenha um papel fundamental na expressão genética e na regulação das atividades celulares. Como uma cópia fiel de uma das fitas do DNA, o RNA transporta informações genéticas essenciais para a síntese de proteínas, participando assim do fluxo de informações genéticas nas células.

A síntese de todos os tipos de RNAs celulares ocorre pelo processo de transcrição. Nesse processo, apenas pequenas seções da molécula de DNA, geralmente um único gene ou, no máximo, alguns, são transcritas para formar o RNA correspondente. Essa seletividade é vital porque nem todos os produtos gênicos são necessários ao mesmo tempo, em uma célula, e uma transcrição constante de todos os genes seria ineficiente e desnecessária. Para essa tarefa, existem três tipos distintos de RNA polimerase (I, II e III) em células eucarióticas, cada um responsável pela transcrição de um tipo específico de RNA. A RNA polimerase I transcreve o RNA ribossômico (rRNA), a RNA polimerase II é responsável pelos pré-mRNAs, snoRNAs, alguns miRNAs e alguns snRNAs, enquanto a RNA polimerase III transcreve moléculas de RNA pequenas, como tRNAs, rRNA pequenos, alguns miRNAs e alguns snRNAs. Essas três RNA polimerases desempenham papéis fundamentais na expressão genética em organismos eucarióticos ([GIRBIG;](#)

MISIASZEK; MÜLLER, 2022; BARBA-ALIAGA; ALEPUZ; PÉREZ-ORTÍN, 2021).

A descoberta de um controle associado ao papel catalítico do RNA (SZYMANSKI; ERDMANN; BARCISZEWSKI, 2003) reconfigurou um novo e convincente conceito das moléculas de RNA como participantes ativos na regulação, catálise e controle de muitas reações que definem processos fundamentais celulares. Essa descoberta expandiu o entendimento anterior, no qual esses papéis eram na maioria atribuídos exclusivamente às proteínas (TAFT *et al.*, 2010).

Inicialmente foram descritas duas categorias de RNAs pequenos na regulação da expressão gênica, nomeados de RNAs de interferência (iRNA): os miRNAs, como reguladores endogênicos dos genes, e os siRNAs (sub-tipos de RNA de interferência) como defensores da integridade do genoma em resposta à invasão celular por ácidos nucleicos virais, transgênicos e de transposons (MEISTER; TUSCHL, 2004). Atualmente, são conhecidos muitos tipos/moléculas de RNAs que exercem diferentes funções. A Figura 4 ilustra os tipos de RNAs.

Tipos de RNA produzidos nas células

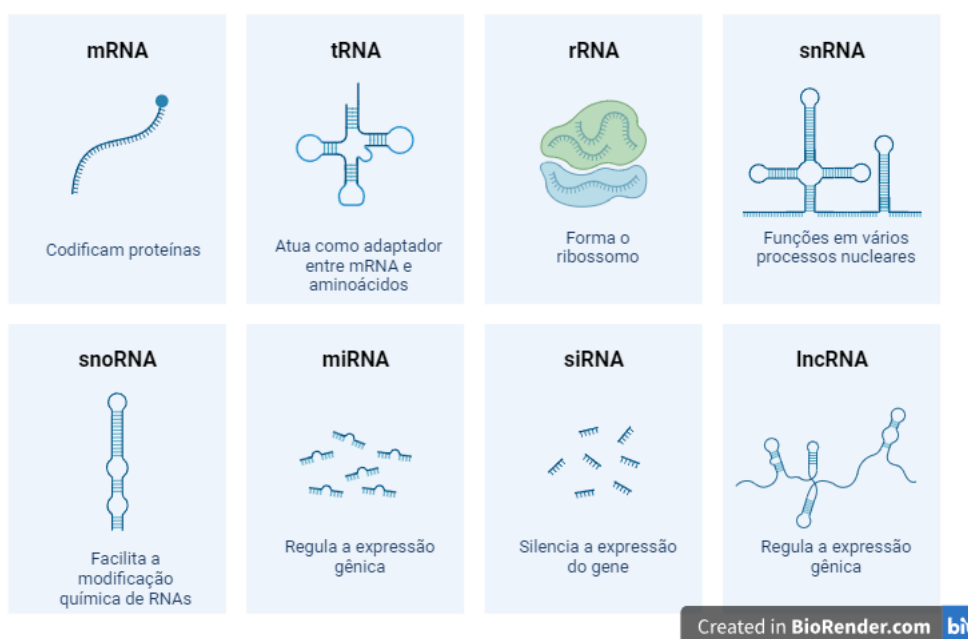


Figura 4 – Tipos de RNA. Criado pelo próprio autor em BioRender (2022).

Os miRNAs constituem uma classe de pequenos RNAs uni filamentos endógenos, geralmente com 18–22 nucleotídeos de comprimento, que desempenham um papel essencial na regulação gênica, atuando tanto ao nível transcricional, através da ligação direta com o DNA (IORIO; CROCE, 2012), quanto pós-transcricional (O'BRIEN *et al.*, 2018). Essas moléculas desempenham um papel de destaque como reguladores-chave da expressão gênica em células eucarióticas (KIM; NAM, 2006).

Um aspecto notável dos miRNAs é sua capacidade de direcionar múltiplos mRNAs, induzindo alterações na expressão de genes envolvidos em vias de interação funcional específicas

e temporais (LU; ROTHENBERG, 2018). Além disso, a combinação de vários miRNAs pode formar redes regulatórias altamente complexas.

Os miRNAs não apenas desempenham um papel fundamental na regulação gênica, mas também são associados a diversas doenças, incluindo doenças cardiovasculares, insuficiência cardíaca, doenças inflamatórias, neurodesenvolvimento, doenças autoimunes, hepáticas, musculoesqueléticas, dermatológicas e cânceres (KIM; NAM, 2006), mais detalhes sobre essas associações podem ser encontrados no trabalho de Ardekani e Naeini (2010).

A descoberta dos miRNAs remete a Lee, Feinbaum e Ambros (1993), que os identificou pela primeira vez em *Caenorhabditis elegans*. Essa descoberta ocorreu por meio de estudos em mutantes heterocrônicos que afetavam o desenvolvimento do organismo, modulando a proteína lin-14. Posteriormente, outros pesquisadores identificaram miRNAs em diversas espécies, cada um apresentando características únicas, incluindo os miRNAs em humanos (*Homo sapiens*).

Para padronizar a nomenclatura dos miRNAs, foi adotado um sistema no qual os miRNAs recebem um nome que identifica a espécie seguida do número em que foi descoberto. Dessa forma, para mencionar o miRNA 1 em *Homo sapiens*, utiliza-se a nomenclatura hsa-miR-1. Vale ressaltar que o "R" maiúsculo refere-se a miRNAs maduros e hsa-mir-1 com o "r" minúsculo, a miRNAs precursores, que serão explicados a seguir.

A biogênese dos miRNAs é um processo altamente regulado que envolve uma série de etapas ordenadas. Esse processo ocorre tanto no núcleo quanto no citoplasma celular, está ilustrado na Figura 5.

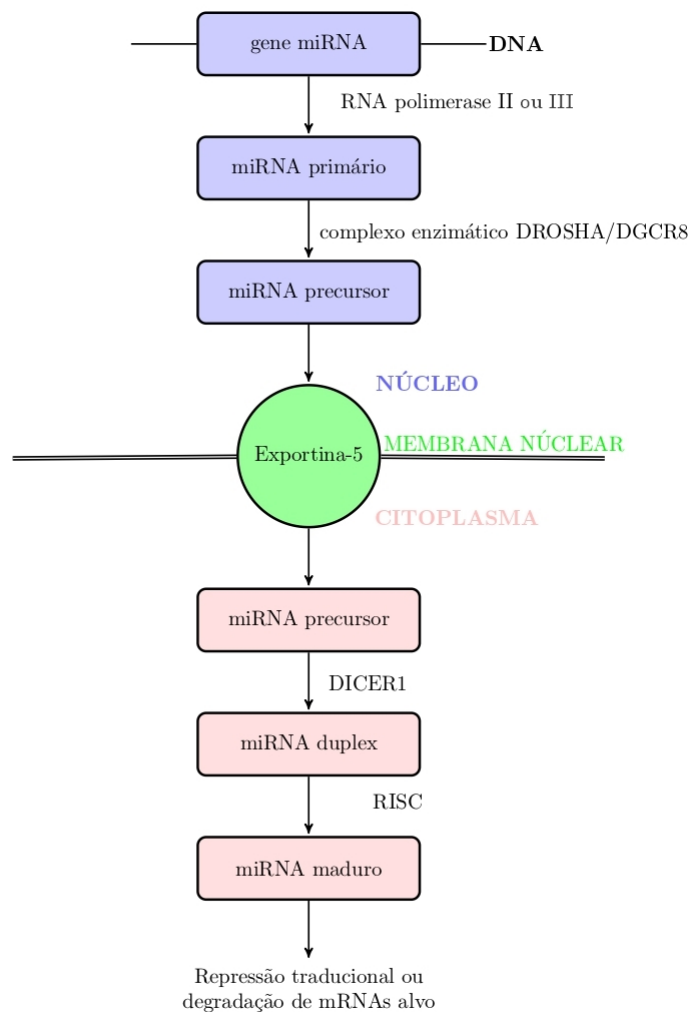


Figura 5 – Biogênese do miRNA. Adaptado de [Inamura e Ishikawa \(2016\)](#) pelo autor.

A seguir, descreve-se resumidamente as principais etapas da biogênese dos miRNAs:

1. A RNA polimerase II ou III inicia a síntese do pri-miRNA a partir da molécula de DNA. O pri-miRNA é caracterizado por conter uma estrutura em grampo (*hairpin*).
2. O complexo microprocessador Drosha-DGCR8 (Pasha) atua no núcleo, clivando o pri-miRNA e gerando o pré-miRNA, uma estrutura em forma de hairpin com cerca de 60-100 nucleotídeos.
3. O pré-miRNA é exportado do núcleo para o citoplasma pela proteína Exportina-5.
4. No citoplasma, o pré-miRNA é clivado pela RNase III endonuclease, chamada Dicer, em conjunto com a proteína de ligação de RNA de cadeia dupla TRBP. Isso resulta na formação de uma molécula de miRNA maduro, composta por uma fita guia funcional (miRNA) e uma fita complementar (miRNA*). Essas sequências são nomeadas conforme as fitas precursoras, por exemplo, miR-10a-3p e miR-10a-5p.

5. A fita do miRNA maduro com maior afinidade à proteína Argonauta é incorporada no complexo de silenciamento induzido por RNA (RISC) (INAMURA; ISHIKAWA, 2016), que orienta o miRNA até o mRNA alvo (BHAYANI; CALIN; LAI, 2012).

Normalmente, os miRNAs atuam inibindo a expressão dos genes-alvo. Quando um miRNA está *upregulated*, a expressão do gene-alvo é inibida, levando a uma *downregulation* do gene (AZIZI; OTHMAN; NAIDU, 2021; WU *et al.*, 2019). Wu *et al.* (2019) destacam que miRNAs conseguem regular a expressão gênica por degradação do mRNA ou por inibição da tradução por ligação ao complexo de silenciamento induzido por RNA (RISC).

Atualmente, já foram identificados mais de 2654 miRNAs maduros e 1.917 precursores no genoma humano, essas informações estão disponíveis em repositórios online, como o [miRBase](#), que também inclui abordagens de descoberta de miRNAs, tanto biológicas quanto bioinformáticas (KOZOMARA; BIRGAOANU; GRIFFITHS-JONES, 2019). Como o número de miRNAs é significativamente menor do que o de genes no genoma humano, estudos globais sobre miRNAs podem ser conduzidos com um número menor de amostras.

Além disso, é importante observar que existem casos de clusters de miRNAs que são transcritos primários policistrônicos. Acredita-se que esses miRNAs agrupados possam estar funcionalmente relacionados, direcionando genes similares ou pertencentes à mesma via, embora essa associação ainda não tenha sido totalmente validada (KIM; NAM, 2006).

Tanto por vias moleculares canônicas quanto não canônicas, os mecanismos de biogênese e processamento dos miRNAs exibem flexibilidade nos mecanismos de RNAs, destacando a variedade de funções realizadas por esses transcritos de RNA no contexto celular e nos mecanismos de oncogênese (KENT; MENDELL, 2006).

É importante ressaltar que estudos que buscam identificar biomarcadores e alvos terapêuticos estão cada vez mais reconhecendo o papel fundamental dos ácidos ribonucleicos não codificadores de proteínas, como os miRNAs, no surgimento, desenvolvimento e progressão de diversas neoplasias humanas.

Um biomarcador, como relatado por Seijo *et al.* (2019), é uma característica objetivamente medida e avaliada como um indicador de processos biológicos normais, patogênicos ou respostas farmacológicas a uma intervenção terapêutica. O trabalho de Strimbu e Tavel (2010) explica o conceito de biomarcadores e sua importância na pesquisa e diagnóstico de doenças.

A pesquisa sobre miRNAs e RNAs como biomarcadores está ganhando destaque devido à capacidade desses influenciarem a regulação gênica em diversas doenças, como câncer, doenças cardiovasculares, autoimunes e outras. Eles desempenham um papel fundamental na identificação de perfis moleculares que podem ser cruciais para o diagnóstico precoce, prognóstico e tratamento personalizado de pacientes. Além disso, a compreensão dos miRNAs e RNAs como reguladores da expressão gênica, tanto em níveis transcricionais quanto pós-transcricionais, está fornecendo valiosos *insights* sobre os mecanismos subjacentes a essas doenças.

Portanto, os miRNAs e os RNAs representam uma área de pesquisa em rápido crescimento, com implicações profundas nas ciências biomédicas. Eles contribuem desde o avanço no entendimento da biologia molecular até a busca por soluções para desafios clínicos e terapêuticos.

2.1.7 Genes e MiRNAs em câncer de pulmão

O câncer de pulmão é uma das principais causas de morbidade e mortalidade em todo o mundo, frequentemente associado ao tabagismo e a fatores ambientais. No entanto, as bases moleculares dessa doença revelaram que fatores genéticos desempenham um papel essencial em sua patogênese. Genes e miRNAs desregulados surgem como componentes fundamentais na formação e progressão do câncer de pulmão, sendo alvos de pesquisas intensas. Esses componentes desempenham papéis intrincados na regulação da expressão gênica e na rede de sinalização celular, influenciando a proliferação, sobrevivência, invasão e metástase das células cancerígenas.

A busca por compreender e identificar os genes desregulados associados ao câncer de pulmão iniciou no século XX, mas obteve avanços significativos nas décadas de 1970 e 1980 com o desenvolvimento de técnicas de biologia molecular e genética. Um dos marcos mais notáveis foi a descoberta do oncogene *RAS* em 1982 por [Barbacid \(1987\)](#), estabelecendo sua relação direta com o desenvolvimento do câncer. Outro gene crucial associado ao câncer é o gene supressor tumoral *TP53*, também conhecido como *p53*, que foi identificado e clonado em 1979, desempenhando um papel crítico na prevenção do câncer ([LANE; CRAWFORD, 1979](#)).

Com o avanço da tecnologia de sequenciamento genético e os avanços do Projeto Genoma Humano, inúmeros outros genes associados ao câncer foram identificados, incluindo *BRCA1* e *BRCA2*, que estão relacionados ao câncer de mama e ovário e cujas mutações hereditárias aumentam o risco de câncer ([WOOSTER et al., 1995](#)).

A identificação de drivers genéticos é fundamental para a genotipagem eficaz no diagnóstico e direcionamento do tratamento ([CINEGAGLIA et al., 2016](#)). Estudos como o realizado por [Kris et al. \(2014\)](#) a partir dos dados do *The Lung Cancer Mutation Consortium* revelaram mutações acionáveis em genes como *EGFR*, *K-RAS*, *N-RAS*, *ALK*, *ERBB2*, *BRAF*, *PIK3CA*, *AKT*, *MEK1* e amplificação de *MET* em mais de 60% em LUAD, os pacientes que receberam tratamentos personalizados com base na genotipagem obtiveram maior expectativa de vida.

A identificação de miRNAs desregulados associados ao câncer é uma área de pesquisa que se desenvolveu nas últimas duas décadas, mas sua associação começou a ser compreendida no início dos anos 2000. Um marco importante nesse campo foi a descoberta de que miR-15a e miR-16-1 estavam localizados na região genômica 13q14, frequentemente deletada na leucemia linfocítica crônica, e a identificação desses miRNAs como supressores tumorais em 2002 representou um avanço ([CALIN et al., 2002](#)). Estudos subsequentes demonstraram que miR-15 e miR-16 regulam negativamente a expressão do fator anti-apoptótico *BCL2*, o que sugere que sua regulação negativa pode resultar em níveis mais elevados de proteína *BCL2* e maior

atividade antiapoptótica (CIMMINO *et al.*, 2005).

O miR-21, considerado um miRNA oncogênico, é superexpresso em vários tipos de tumores sólidos e leucemias, desempenhando um papel essencial na regulação de processos celulares, incluindo proliferação, migração celular e inibição da apoptose (PFEFFER; YANG; PFEFFER, 2015; FENG; TSAO, 2016). Estudos subsequentes identificaram miR-15b e miR-27b capazes de distinguir pacientes com câncer de pulmão com alta sensibilidade e especificidade, enquanto a expressão de miR-1254 e miR-574-5p no soro permitiu discriminar pacientes com câncer de pulmão de não fumantes em estadiamento inicial (HENNESSEY *et al.*, 2012; FOSS *et al.*, 2011). Além disso, uma assinatura com o miR-16-5p, miR-92a-3p e miR-451a apresentou alta sensibilidade e especificidade na detecção de câncer de pulmão (REIS *et al.*, 2020).

O let-7, um dos miRNAs mais estudados, foi inicialmente identificado no estudo de Takamizawa *et al.* (2004), a pesquisa revelou que a expressão reduzida deste miRNA estava associada a uma notável redução na sobrevivência após ressecção potencialmente curativa. Estudos posteriores demonstraram que a superexpressão do let-7 estava relacionada à inibição do crescimento de células de câncer de pulmão. Outras pesquisas subsequentes, também contribuíram para a compreensão do papel do let-7 na regulação da expressão gênica e sua importância em processos relacionados ao câncer (JOHNSON *et al.*, 2005; INAMURA *et al.*, 2007).

Os miRNAs têm sido extensivamente estudados como biomarcadores para o estadiamento inicial do câncer de pulmão, com a detecção de miRNAs circulantes no plasma ou soro de pacientes com câncer de pulmão mostrando promissora aplicação como biomarcadores não invasivos (WANG *et al.*, 2020; VYKOUKAL *et al.*, 2022).

A alteração na expressão de miRNAs em câncer em comparação com seus respectivos tecidos normais é uma observação comum. Estas alterações podem fornecer *insights* valiosos sobre a patogênese do câncer, com o potencial de serem usadas como biomarcadores para diagnóstico, prognóstico e resposta ao tratamento. Os avanços recentes nas tecnologias de NGS aceleram a integração entre pesquisa científica e prática clínica, facilitando o acesso a dados valiosos por meio de repositórios públicos e ferramentas analíticas. Um exemplo notável é o Projeto Atlas do Genoma do Câncer (TCGA), uma iniciativa conjunta do Instituto Nacional do Câncer (NCI) e do Instituto Nacional de Pesquisa do Genoma Humano (NHGRI), que fornece informações cruciais para pesquisas clínicas (CHANG; LEE; HUANG, 2015). Segundo Chang, Lee e Huang (2015) os dados fornecidos pelo TCGA já colaboraram em muitas pesquisas e trabalhos para uma melhor compreensão dos mecanismos moleculares subjacentes ao LUAD e LUSC, ao esclarecimento das implicações clínicas, ao desenvolvimento de ferramentas apropriadas, algoritmos e à descoberta de novas terapias.

Sendo assim, a contínua investigação sobre genes e miRNAs e seu papel no câncer de pulmão desempenha um papel significativo na compreensão e tratamento dessa doença, prometendo diagnósticos mais precisos e terapias.

2.2 Bioinformática

A bioinformática, um termo cunhado por Paulien Hogeweg e Ben Hesper na década de 1970 ([HOGEWEG, 2011](#)), se tornou um pilar essencial à medida que lidamos com o crescente volume de dados da revolução genômica ([HOOD; ROWEN, 2013](#)), uma vez que as ferramentas biológicas e estatísticas tradicionais se mostraram inadequadas diante do grande volume de informações.

Esta ciência interdisciplinar engloba biólogos, cientistas da vida molecular, especialistas em computação, estatísticos, matemáticos e muito mais. Assim, se tornou uma nova área da ciência que trabalha com avanço de novas técnicas computacionais aplicadas à biologia molecular e genômica. Atualmente, vem sendo muito utilizada em pesquisa com fins diagnósticos, preventivos e terapêuticos.

Com a bioinformática, a expressão de genes envolvidos nos diversos metabolismos pode ser quantificada por meio de métodos que avaliem o perfil transcricional celular. A técnica de *microarray* (ou microarranjos), por exemplo, permite comparar níveis de expressão gênica em larga escala em diferentes tecidos, tempos e condições.

A estatística também desempenha um papel essencial na análise de dados com a bioinformática, fornecendo interpretações significativas dos resultados. Em conjuntos de dados biológicos, a compreensão das aplicações de diferentes métodos estatísticos é crucial para a seleção da abordagem mais adequada, sendo essencial para evitar resultados inconsistentes e garantir a relevância biológica das descobertas.

Além disso, a combinação de bioinformática com estatísticas e aprendizado de máquina desempenhou um papel fundamental na análise de conjuntos de dados genômicos. Algoritmos de aprendizado de máquina permitem a análise de grandes volumes de dados genômicos e a previsão de genes de importância, oferecendo informações valiosas, por exemplo, na identificação de genes codificadores de proteínas em vírus que causam doenças ([SARUMI; LEUNG, 2022](#)).

Por fim, a bioinformática atualmente é uma aliada poderosa na pesquisa genômica, permitindo a análise, interpretação e integração de dados genéticos em um cenário de rápido avanço tecnológico. O seu papel é fundamental à medida que enfrentamos desafios cada vez mais complexos no campo da genômica e da biologia computacional.

2.2.1 Análise de expressão diferencial

A análise de expressão diferencial é uma metodologia utilizada em biologia molecular para identificar e quantificar as diferenças na expressão gênica entre dois ou mais conjuntos de amostras biológicas. Essa técnica é fundamental para entender como a atividade dos genes varia em diferentes condições, como em diferentes tecidos, estados de doença ou em resposta a tratamentos específicos. Essa abordagem fornece *insights* valiosos e contribui para a compreensão

das bases moleculares de doenças e fenômenos biológicos.

Essa análise desempenha um papel crucial na genômica e na pesquisa em biologia molecular, permitindo a identificação de transcritos codificadores e os não codificadores que desempenham papéis fundamentais em vias de sinalizações associadas à oncogênese. Estes elementos identificados podem emergir como valiosos biomarcadores, oferecendo perspectivas cruciais para diagnóstico, prognóstico e o desenvolvimento de terapias inovadoras no contexto da pesquisa oncológica (PENG; CROCE, 2016).

No entanto, a identificação de DE-miRNAs e DEGs é um desafio, devido as várias abordagens disponíveis na literatura. Até o momento, não existe um consenso estabelecido sobre a metodologia ideal que garanta resultados robustos, precisos e reproduzíveis em termos de análise de expressão diferencial. Métodos baseados em modelos binomiais negativos, como EdgeR (ROBINSON; MCCARTHY; SMYTH, 2010) e DESeq2 (LOVE; HUBER; ANDERS, 2014), são amplamente utilizados e empregam comumente o ajuste dos valores *p* com o método Benjamini-Hochberg (BENJAMINI; HOCHBERG, 1995), para controlar as *False Discovery Rate* (FDR).

No entanto, uma investigação recente de Li *et al.* (2022) destacou que esses algoritmos frequentemente apresentam altas FDR, podendo exceder 20% ao adotar um FDR de 5%. Além disso, muitos desses algoritmos foram desenvolvidos há vários anos, quando as bases de dados eram menos abrangentes em termos de informações disponíveis.

Uma abordagem alternativa e também popular é o Limma (RITCHIE *et al.*, 2015), que emprega modelos lineares para calcular diferenças significativas na expressão de genes e oferece flexibilidade ao analisar dados de expressão diferencial, podendo ser aplicado tanto a dados de *microarray*² quanto a dados de RNA-seq.

Tanto o EdgeR, DESeq2 quanto o Limma são ferramentas fundamentais que incorporam procedimentos de normalização durante a condução da análise de expressão diferencial. O EdgeR emprega o método de normalização TMM (*Trimmed Mean of M values*), enquanto o DESeq2 normaliza os dados por meio de fatores de escala globais, calculando a média geométrica desses fatores para cada amostra. Por sua vez, o Limma utiliza a técnica de normalização *Loess* para ajustar os dados.

A justificativa para essas etapas de normalização reside na necessidade de corrigir discrepâncias sistemáticas nas contagens de leituras entre as amostras, visando resultados mais precisos na identificação de genes diferencialmente expressos. Embora cada método possua suas particularidades, a escolha entre eles frequentemente depende das características específicas do experimento e da análise em questão.

Contudo, é importante ressaltar as ponderações apresentadas por Neve *et al.* (2013) em

² Um *microarray* é uma ferramenta de biologia molecular que permite a análise simultânea da expressão de milhares de genes ou a detecção de milhares de sequências de DNA, RNA ou proteínas em uma única experiência.

sua pesquisa. Destacando a influência significativa da normalização na interpretação e validade das etapas subsequentes da análise, o autor também alerta para a possível redução na variância e a introdução de dependência entre os resultados normalizados. Esses aspectos, embora muitas vezes negligenciados, merecem consideração cuidadosa. Adicionalmente, os autores apontam que a maioria das técnicas de normalização para expressão diferencial se concentra na média da expressão e é sensível a valores discrepantes.

Diante da complexidade inerente à análise de expressão diferencial, é evidente que a escolha da metodologia adequada é uma decisão crucial para a obtenção de resultados robustos e significativos. A diversidade de abordagens, tanto paramétricas quanto não paramétricas, reflete a busca contínua por ferramentas que atendam aos desafios específicos encontrados em diferentes conjuntos de dados e contextos experimentais. Nesse cenário, a compreensão dos métodos tradicionais, como EdgeR, DESeq2 e Limma, é essencial, mas também é imperativo explorar alternativas não paramétricas que oferecem flexibilidade diante de premissas menos rígidas. Assim, a análise de expressão diferencial se revela em constante evolução, em que a escolha criteriosa entre abordagens se torna uma peça-chave para a interpretação confiável dos resultados.

Abordagens não paramétricas, embora menos restritivas, geralmente demandam tamanhos de amostra substanciais para gerar resultados robustos. Entre as alternativas destacadas, temos o NOISeq ([TARAZONA et al., 2011](#)), *dearseq* ([GAUTHIER et al., 2020](#)) e o teste de soma de postos de Wilcoxon ([LI et al., 2022](#)).

O NOISeq, baseado na distribuição não paramétrica do número de leituras, apresenta-se como uma opção valiosa em cenários onde a normalidade das distribuições é incerta. Já o *dearseq*, focado na análise de experimentos de RNA-seq com réplicas em condições diferentes, oferece uma abordagem robusta para lidar com a variabilidade inerente nos dados.

No entanto, é oportuno ressaltar a recomendação de [Li et al. \(2022\)](#) para o teste de Wilcoxon-Mann-Whitney ([MANN; WHITNEY, 1947](#)) em estudos com grandes tamanhos de amostra ($n > 100$). Trabalhos anteriores, como os conduzidos por [Lim et al. \(2017\)](#) e [Iwamoto et al. \(2019\)](#), ilustram casos bem sucedidos de aplicação do teste de Wilcoxon-Mann-Whitney na identificação de expressão diferencial.

Essas abordagens não paramétricas, ao proporcionarem flexibilidade e robustez, são ferramentas valiosas na análise de expressão diferencial, especialmente quando as premissas das abordagens paramétricas podem não ser plenamente atendidas. Sua escolha dependerá das características específicas do experimento, e também de consideração cuidadosa dessas opções, para fornecer robustez das conclusões obtidas.

Em resumo, a análise de expressão diferencial é uma área complexa e em constante evolução. A escolha da técnica apropriada, seja ela paramétrica ou não paramétrica, depende das particularidades dos dados e dos objetivos específicos do estudo. Cada abordagem tem suas vantagens e limitações, tornando essencial considerar cuidadosamente as características dos

dados e os objetivos da análise ao selecionar a melhor estratégia para a análise de expressão diferencial.

2.2.2 Análises biológicas

Após a condução da análise de expressão diferencial, é comum na pesquisa e em estudos biológicos a utilização de uma variedade de ferramentas para análises mais aprofundadas. Essas análises desempenham um papel essencial no entendimento dos resultados e no avanço do estudo. Diversos estudos têm empregado essa prática para explorar as análises biológicas frequentemente realizadas após a análise de expressão diferencial, tais:

- **Análise de enriquecimento de vias** é um passo fundamental para identificar as vias biológicas e processos celulares significativamente enriquecidos com genes ou miRNAs diferencialmente expressos. Isso proporciona *insights* sobre os mecanismos subjacentes que podem estar envolvidos nas respostas biológicas observadas.
- **Análise de correlação entre mRNA e miRNA** é realizada para identificar relações de expressão entre os genes e os miRNA. Isso pode ajudar a revelar quais miRNAs podem estar regulando a expressão de genes-alvo e desempenhando um papel na regulação pós-transcricional.
- **Predição de genes-Alvo dos DE-miRNAs** é crucial para identificar quais genes podem ser regulados por esses miRNAs. Essa análise pode fornecer informações sobre possíveis alvos terapêuticos ou biomarcadores.
- **Identificação de DEGs e miRNAs comuns** pode revelar conjuntos de genes regulados por miRNAs específicos. Isso ajuda a entender as redes de regulação da expressão gênica.
- **Seleção de miRNA-mRNA validados experimentalmente** para obter resultados mais confiáveis, é fundamental identificar miRNA-mRNA pares que tenham sido validados experimentalmente, para esses pares terem maior probabilidade de serem biologicamente relevantes.
- **Identificação de genes expressos no tecido de interesse** é importante, especialmente em estudos envolvendo amostras de tecidos complexos. Isso ajuda a refinar os resultados e focar em genes relevantes para o contexto do estudo.
- **Rede regulatória entre DE-miRNA e DEGs** pode revelar interações complexas e hierarquias de regulação. Isso é fundamental para entender como os miRNAs afetam a expressão gênica.
- **Montagem e análise de redes de interação proteína-proteína (PPI)** são utilizadas para identificar proteínas-chave que desempenham funções centrais em uma rede de interações. Essa análise ajuda a identificar proteínas com importância biológica significativa.

- **Identificação de genes relacionados ao prognóstico** é essencial em estudos clínicos. Essa análise ajuda a identificar biomarcadores que podem estar ligados a desfechos clínicos, como sobrevivência ou progressão da doença.

Essas são apenas algumas das análises biológicas comuns realizadas após a análise de expressão diferencial, algumas dessas podem ser vistas no trabalho de [Souza et al. \(2023\)](#), por exemplo. A escolha das análises apropriadas depende dos objetivos específicos do estudo e das características dos dados. É importante destacar que a integração de múltiplas abordagens pode fornecer uma compreensão mais abrangente e detalhada dos mecanismos subjacentes aos resultados da expressão diferencial.

2.2.3 Aplicações de Árvore de Decisão e Florestas Aleatórias

Após a identificação de DE-miRNAs e DEGs e a condução de análises biológicas para identificar os principais miRNAs e genes relacionados aos objetivos da pesquisa, é altamente recomendável a realização de análises estatísticas, tais como classificação e predição, análise de sobrevivência, análise de clusterização, entre outras abordagens. Essas análises abrangentes buscam enriquecer a interpretação dos resultados, proporcionando uma contribuição significativa para uma compreensão mais profunda dos mecanismos subjacentes ao estudo.

Diante da alta dimensionalidade dos dados genômicos, resultante das complexidades inerentes à genômica, tem-se um desafio substancial para a bioinformática, análise de dados e estatística, uma vez que a estrutura de regressão clássica torna-se inviável ([CHEN; ISHWARAN, 2012](#)). Nesse contexto desafiador, o Aprendizado de Máquina (AM - *machine learning*), uma subárea da Inteligência Artificial (IA), surge como uma ferramenta valiosa. Por meio do AM, algoritmos são programados para aprender com experiências, oferecendo uma abordagem flexível e adaptável diante da complexidade inerente aos dados genômicos.

Na literatura há várias definições para AM, contudo [Mitchell \(1997\)](#) definiu como:

“A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência.”

De acordo com [Faceli et al. \(2011\)](#), o AM se divide principalmente em duas categorias:

- **Aprendizado Supervisionado:** algoritmos são aplicados a conjuntos de dados de treinamento rotulados para induzir um modelo preditivo capaz de prever os valores do atributo alvo. Técnicas amplamente reconhecidas incluem Florestas Aleatórias, K-NN (*k-nearest neighbors*), SVM (*support vector machine*), árvores de decisão, redes neurais, entre outras.
- **Aprendizado Não Supervisionado:** algoritmos extraem padrões dos valores de um conjunto de dados, sem a necessidade de rótulos. Técnicas notáveis abrangem agrupamento hierárquico e de k-médias, modelos de mistura gaussiana, entre outros.

Estas categorias representam abordagens fundamentais no campo do AM, cada uma com suas técnicas específicas que se adequam a diferentes contextos e objetivos. Algoritmos de IA, cada vez mais utilizados em tarefas preditivas ou descritivas, são desenvolvidos especialmente resolver problemas de alta dimensão (CHEN; ISHWARAN, 2012). A Figura 6 ilustra a hierarquia dessas categorias e as tarefas associadas.

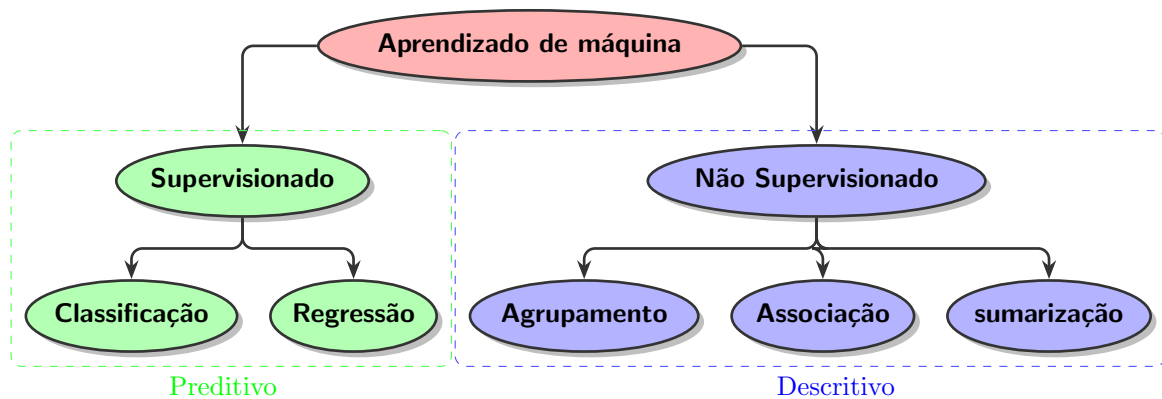


Figura 6 – A hierarquia do aprendizado. Adaptado de Faceli *et al.* (2011).

Uma técnica amplamente reconhecida e de fácil compreensão no aprendizado de máquina supervisionado é a árvore de decisão. Essa se destaca por sua eficiência, interpretação acessível das saídas, natureza não paramétrica (não supõe em relação ao erro dos modelos) e característica de caixa branca (*white box*) (FACELI *et al.*, 2011). A abordagem dessa técnica é particularmente eficaz para enfrentar problemas complexos, dividindo-os em subproblemas mais simples e aplicando recursivamente a técnica para unir esses subproblemas em uma árvore de decisão completa. Essa estratégia permite a resolução de problemas intrincados ao dividir o espaço de predição em subespaços, ajustando cada subespaço com modelos específicos (FACELI *et al.*, 2011).

O estudo realizado por Loh (2014) proporciona uma visão retrospectiva dos algoritmos mais influentes em árvores de decisão, apresentando-os de maneira cronológica na Figura 7 e descrevendo-os detalhadamente a seguir:

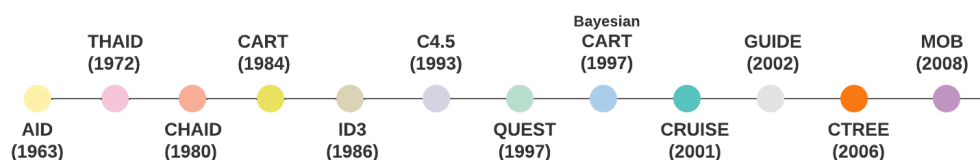


Figura 7 – Cronologia dos algoritmos de destaque em árvores de decisão. Criado no Lucidchart.

- **AID** (*Automatic Interaction Detection*) trabalha com uma covariável e uma variável resposta, ambas quantitativas, buscando particionar de forma binária o nó para minimizar a

soma de quadrados (MORGAN; SONQUIST, 1963).

- **THAID** (*Theta Automatic Interaction Detection*) apresenta uma adaptação do AID para variável resposta qualitativa (MESSENGER; MANDELL, 1972).
- **CHAID** (*Chisquared Automatic Interaction Detector*) não realiza particionamentos recursivos, mas junções recursivas fundindo dois ou mais nós com base em testes de hipótese estatísticos (KASS, 1980).
- **CART** (*Classification and regression trees*) segue a mesma abordagem de busca gananciosa como AID e THAID. No entanto, em vez de usar regras de parada, o CART deixa a árvore crescer e, em seguida, realiza poda com base na menor estimativa de erro de validação cruzada (BREIMAN *et al.*, 1984).
- **C4.5** é uma extensão do algoritmo de classificação ID3 (QUINLAN, 1986), podendo trabalhar com covariáveis tanto qualitativas como quantitativas. Apresenta uma nova medida relacionada ao ganho de informação, método de poda e trata eventuais valores faltantes (QUINLAN, 1993).
- **QUEST** (*Quick, Unbiased and Efficient Statistical Tree*) utiliza testes de hipóteses baseados na distribuição F de Snedecor, em conjunto com estatísticas de ordens das covariáveis (LOH; SHIH, 1997).
- **Bayesian CART** consiste do CART no contexto Bayesiano, utilizando distribuições a priori e a posteriori para modelar a distribuição da variável considerando os particionamentos (DENISON; MALLICK; SMITH, 1998).
- **CRUISE** (*Classification Rule with Unbiased Interaction Selection and Estimation*) utiliza particionamentos não-binários com teste de hipóteses baseados na distribuição qui-quadrado (KIM; LOH, 2001).
- **GUIDE** (*Generalized, Unbiased, Interaction Detection*) é um aprimoramento dos algoritmos QUEST e CRUISE, adotando seus pontos fortes e corrigindo suas fraquezas (CHAUDHURI; LOH, 2002).
- **CTREE** (*Conditional Inference Trees*) o algoritmo não foca apenas em maximizar uma medida, mas também utiliza o conceito de significância estatística das covariáveis do particionamento (HOTHORN; HORNIK; ZEILEIS, 2006).
- **MOB** (*Model-Based recursive partitioning*) realiza particionamento recursivo, ajusta mínimos quadrados, regressão logística e outros modelos, usando as funções de pontuação dos estimadores (ZEILEIS; HOTHORN; HORNIK, 2008).

O uso de métodos de aprendizado de máquina supervisionado para identificar o perfil molecular de tumores tem ganhado destaque em estudos recentes. [Kingsford e Salzberg \(2008\)](#) destacam que as árvores de decisão têm encontrado aplicação na biologia computacional e bioinformática, sendo valorizadas por sua capacidade de integrar diversos tipos de dados para realizar previsões precisas.

No trabalho de [Sherafatian e Arjmand \(2019\)](#), o *status* e subtipo do câncer de pulmão foram classificados utilizando miRNAs empíricos de controle negativo específicos para câncer de pulmão. Esses miRNAs foram empregados na normalização de conjuntos de dados de LUAD e LUSC do TCGA, utilizando árvores de decisão como ferramenta de classificação.

[Chen e Dhahbi \(2021\)](#) observam a escassez de biomarcadores para LUAD e LUSC, ressaltando que os mecanismos biológicos distintos desses biomarcadores ainda não foram totalmente elucidados. Poucos estudos exploraram o uso de aprendizado de máquina sobreposto para classificação de câncer e descoberta de biomarcadores.

Em um estudo de [Yang et al. \(2022\)](#), a precisão e os benefícios de três algoritmos de aprendizado de máquina, incluindo árvores de decisão, foram comparados para LUAD e LUSC. Os resultados indicaram que a utilização desses algoritmos apresenta potencial para auxiliar os médicos na tomada de decisões clínicas.

Outra metodologia amplamente utilizada é conhecida como Florestas Aleatórias (*Random Forests*) concebida por [Breiman \(2001a\)](#), influenciada por trabalhos anteriores de [Amit e Geman \(1997\)](#), [Ho \(1998\)](#) e [Dietterich \(2000\)](#), entre outros.

Este algoritmo tem aplicações diversificadas em AM e representa uma abordagem de conjunto baseada em árvores não paramétricas que combina conceitos de vizinhos mais próximos adaptativos com o método de ensacamento (*bagging*) de [Breiman \(1996\)](#) para uma inferência adaptativa de dados eficaz. A capacidade do algoritmo de fornecer medidas de importância das variáveis, seu desempenho robusto diante de inúmeros variáveis e a versatilidade para enfrentar problemas de grande escala destacam-se na literatura ([BIAU; SCORNET, 2016](#)).

Além disso, as Florestas Aleatórias conseguem selecionar e classificar variáveis com base em medidas de importância variável ([CHEN; ISHWARAN, 2012](#)). Contudo, é importante notar que a análise de Florestas Aleatórias pode ser desafiadora devido à sua natureza considerada "caixa-preta", o que significa que é possível examinar suas entradas e saídas, mas não há compreensão explícita de seu funcionamento interno.

A natureza gananciosa das divisões nodais passo a passo nas Árvores de Decisão e em Florestas Aleatórias permite a imposição de regularização, tornando-as eficazes em problemas com "grande número de variáveis e pequeno número de observações" ([HUYNH; NGUYEN; DO, 2020](#); [ISHWARAN et al., 2010](#)). As árvores de decisão e Florestas Aleatórias lidam de maneira hábil com correlações e interações entre variáveis.

O uso de árvores de decisão e Florestas Aleatórias tornou-se proeminente como ferramenta

essencial na análise de dados genômicos e na pesquisa em bioinformática. A disponibilidade de funções implementadas em linguagens de programação como R ([R Core Team, 2020](#)) e Python ([ROSSUM; JR; L, 2006](#)) amplia o acesso a essas técnicas, tornando-as acessíveis para análise de problemas de classificação e regressão em saídas qualitativas e quantitativas, respectivamente.

Além disso, a aplicação de algoritmos de AM para a detecção de potenciais biomarcadores emerge como um tópico promissor e inovador, com o potencial de contribuir significativamente para a prevenção e detecção aprimorada do câncer de pulmão. A integração dessas metodologias inovadoras na pesquisa biológica reflete o avanço constante na aplicação de técnicas computacionais para a compreensão e enfrentamento de desafios complexos em genômica e saúde.

3 Bancos de Dados e Métodos

Neste capítulo são apresentados e detalhados todos os bancos de dado utilizados. A seguir, descrevem-se os métodos e técnicas de bioinformática empregados na pesquisa, abrangendo análises tanto biológicas quanto estatísticas. Essas abordagens são detalhadas conforme as etapas do estudo, com o propósito de contribuir para a exploração de potenciais biomarcadores associados a LUAD ou LUSC nos estágios iniciais da doença. Todas as análises e plotagens relacionadas à obtenção dos dados, pré-processamento, filtragem, expressão diferencial, tratamento de desequilíbrio de classe, aplicação e avaliação de algoritmos de aprendizado de máquina, entre outras, foram conduzidas no *software* R versão 4.1.1 (R Core Team, 2020). Essas operações foram executadas em um computador equipado com processador Ryzen 5 5600G, 16GB de memória RAM e Placa de Vídeo GTX 1060 de 3GB.

3.1 Bancos de dados

Neste trabalho utilizam-se amostras de expressão de miRNA (*miRNA Expression Quantification*) e expressão gênica (*RNA Expression Quantification*) de pacientes com LUAD ou LUSC, que estão publicamente disponíveis no TCGA. Os bancos de dados de expressão foram previamente gerados por sequenciamento de próxima geração (seq) no TCGA Research Network (<http://cancergenome.nih.gov/>) e filtros buscando por amostras de pacientes nos estágios I, IA e IB, conforme o sistema de estadiamento do AJCC (4^a a 7^a edição).

Os arquivos de amostras de dados de quantificação da expressão de isoforma¹ miRNA-seq e RNA-seq disponíveis no portal *Genomic Data Commons* (GDC - <https://portal.gdc.cancer.gov/>) e as informações clínicas das amostras dos pacientes com LUAD ou LUSC foram baixados, todas separadamente, usando a Ferramenta de Transferência de Dados GDC (*GDC Data Transfer Tool*), no Laboratório de Oncologia Molecular e Translacional da UNIPEX (FMB/UNESP), com colaboração da doutoranda Vanessa das Graças Pereira de Souza utilizando técnicas de bioinformática em julho de 2022.

Após baixar o banco de dados de miRNA-seq e RNA-seq para as doenças LUAD e LUSC, estes foram unidos, por meio do ID, com os respectivos dados clínicos das amostras presentes

¹ Formas diferentes de uma proteína que pode ser produzida a partir de genes diferentes, ou a partir do mesmo gene por processo alternativo.

nos dados, resultando nos bancos de dados:

- I. **LUAD**: dados clínicos e miRNAs-seq;
- II. **LUAD**: dados clínicos e RNAs-seq;
- III. **LUSC**: dados clínicos e miRNAs-seq;
- IV. **LUSC**: dados clínicos e RNAs-seq.

Para todos esses bancos de dados, foi necessário realizar uma limpeza, em que amostras que não possuíam dados clínicos, colunas totalmente vazias ou que continham mais de 80% dos dados de contagens brutas com zeros, foram removidas do banco. Amostras de pacientes que eram duplicadas/triplicadas/quadruplicadas foram substituídas pela respectiva média.

Assim, as informações clínicas das amostras baixadas ficaram compostas por 12 variáveis, que incluem informações demográficas, de diagnóstico e tratamento dos pacientes, e estão apresentadas na Tabela 2 com suas respectivas definições.

Tabela 2 – Descrição das variáveis demográficas, de diagnóstico e tratamento presentes nos bancos de dados.

Variáveis	Descrição
Tipo de amostra	tumor primário ou tecido sólido normal.
Demográficos	Caracterização do paciente por meio da segmentação da população
Sexo	identifica o sexo do paciente.
Raça	agrupamento social e cultural auto-descrito de um indivíduo.
Estado Vital	estado de sobrevivência do paciente.
Idade	Idade do paciente (em anos).
Diagnóstico	Investigação, análise e reconhecimento da presença e natureza da doença, condição lesão a partir ou de sinais e sintomas expressos
Estágio clínico	grupo de estágio determinado a partir de informações clínicas sobre o tumor, linfonodo regional e metástases e pelo agrupamento de casos com prognóstico semelhante para câncer.
Tamanho do tumor	extensão do câncer primário com base em evidências obtidas de parâmetros de avaliação clínica determinados antes do tratamento.
Diagnóstico primário	termo usado para descrever o diagnóstico histológico do paciente, conforme descrito pela Classificação Internacional de Doenças para Oncologia (CID-O) da OMS.
Malignidade anterior	usado para descrever a história do paciente de diagnóstico prévio de câncer.
Tecido ou órgão de origem	utilizado para descrever o local anatômico de origem, da doença maligna do paciente, conforme descrito pela Classificação Internacional de Doenças para Oncologia (CID-O) da OMS.
Dias para o último acompanhamento	intervalo de tempo desde a data do último acompanhamento até a data do diagnóstico patológico inicial, representado como um número calculado de dias.
Tratamento	Registro da administração e intenção de agentes terapêuticos fornecidos a um paciente
Tratamento ou terapia	aplicável relacionado à administração de agentes terapêuticos recebidos.

Fonte: TCGA.

A seguir, são apresentados os bancos de dados empregados na realização das três análises de interesse conduzidas neste estudo:

Análise I: utilizou-se ao banco de dados de expressões de miRNA-seq precursores do banco de dados I LUAD. Para as análises realizadas nesse banco, as informações clínicas dos pacientes não foram utilizadas, uma vez que o objetivo nessa etapa do trabalho era a comparação de ferramentas de análise de expressão diferencial. A Tabela 3 apresenta as informações do banco de dados.

Tabela 3 – Informações do banco de dados para a análise I.

Informações	Descrição
Banco de dados	LUAD miRNA-seq precursores ($n = 300$)
Informações clínicas	Não
Técnica	miRNA-seq (864 miRNAs)
Variáveis	
Resposta	T vs. N
Explicativas	Expressões dos miRNAs

Observação: Em julho de 2022 foram obtidas amostras nos bancos de dados I LUAD e III LUSC de miRNA-seq precursores, ou seja, as expressões de sequenciamento dos miRNAs encontrados no TCGA não continham informações sobre o braço -3p ou -5p. Nos resultados da análise I do trabalho, houve a necessidade de transformar esses miRNAs em maduros para conseguir realizar as análises biológicas desejadas. Por isso, antes de iniciar as partes II e III e sabendo que este mesmo problema já tinha sido relatado previamente, [Kuo et al. \(2015\)](#) desenvolveram um *script* em Python para interrogar essas informações. A ideia adotada por eles foi aplicada neste trabalho usando *bash* e *software* R. Assim, para as análises II e III os miRNAs são maduros e possuem as informações a respeito dos braços -3p e -5p.

Análise II: utilizaram-se 4 bancos de dados completos de LUAD e LUSC (I, II, III e IV). A Tabela 4 apresenta as informações mais detalhadas dos bancos de dados.

Tabela 4 – Informações dos bancos de dados para a análise II.

Informações	Descrições			
Bancos de dados	LUAD miRNA-seq ($n = 299$)	LUAD RNA-seq ($n = 300$)	LUSC miRNA-seq ($n = 249$)	LUSC RNA-seq ($n = 266$)
Informações clínicas	Sim	Sim	Sim	Sim
Técnica	miRNA-seq (1.057 miRNAs)	RNA-seq (40.599 RNAs)	miRNA-seq (1.061 miRNAs)	RNA-seq (40.918 RNAs)
Variáveis				
Resposta	T vs. N			
Explicativas	Variáveis clínicas e expressões dos miRNAs ou RNAs			

O número de RNAs é alto em LUAD e LUSC RNA-seq porque o TCGA usa o Genecode para fazer as anotações, isto inclui, por exemplo, codificação de proteínas, *long non-coding* RNA (lncRNA) e pseudogene. O trabalho de [Frankish et al. \(2021\)](#) explica com detalhes o Genecode.

Análise III: a fim de identificar os miRNAs e RNAs que diferenciam as doenças LUAD e LUSC em estadiamento inicial, os bancos de dados I e III, LUAD e LUSC miRNA-seq foram combinados. Isso foi feito selecionando os miRNAs comuns e excluindo os miRNAs não comuns, além de excluir as amostras de tecido normal. O mesmo procedimento foi repetido para os bancos de dados II e IV de LUAD e LUSC RNA-seq. É importante ressaltar que as características clínicas dos pacientes não foram consideradas neste estudo. Como resultado, foram geradas dois novos bancos de dados, cujos detalhes são apresentados na Tabela 5.

Tabela 5 – Informações dos bancos de dados da análise III.

Informações	Descrições	
Bancos de dados	LUAD-LUSC miRNA-seq ($n = 496$)	LUAD-LUSC RNA-seq ($n = 511$)
Informações clínicas	Não	Não
Técnica	miRNA-seq (988 miRNAs)	RNA-seq (39.275 RNAs)
Variáveis		
Resposta	LUSC vs. LUAD	LUSC vs. LUAD
Explicativas	Expressões dos miRNA-seq	Expressões dos RNA-seq

3.2 Testes estatísticos para análise das variáveis categóricas

Testes estatísticos, testes de hipóteses ou testes de significâncias são ferramentas estatísticas que permitem decidir entre duas ou mais hipóteses, utilizando dados observados de um determinado experimento.

3.2.1 Estatística Qui-Quadrado de Pearson

O teste qui-quadrado de Pearson (Q_p) foi proposto por [Pearson \(1900\)](#). A estatística do teste compara se a frequência de um evento específico na amostra difere significativamente da frequência esperada. Neste trabalho, utiliza-se Q_p para avaliar dados categóricos quanto à associação entre variáveis qualitativas dispostas em tabelas de contingência 2×2 ou $s \times 2$. Antes de realizar o teste, é crucial observar algumas condições: amostras independentes, seleção aleatória e observações como contagens ou frequências pertencentes a uma única categoria.

As hipóteses são:

$$\begin{cases} H_0 : \text{Ausência de associação entre as variáveis X e Y;} \\ H_1 : \text{Presença de associação.} \end{cases}$$

Sob as condições adequadas, pode-se utilizar a estatística Q_p , com a fórmula geral para tabelas $s \times r$ apresentada no livro de [Giolo \(2017\)](#):

$$Q_p = \sum_{i=1}^s \sum_{j=1}^r \frac{(n_{ij} - e_{ij})^2}{e_{ij}},$$

considere s e r o número de categorias das variáveis X e Y de uma mesma unidade experimental, em que $e_{ij} = \frac{(n_{i+})(n_{+j})}{n}$, $i = 1, \dots, s$ e $j = 1, \dots, r$ representam os valores esperados e n_{ij}

os valores observados, para as categorias de i e j e N_{i+} e N_{+j} os totais para as categorias i e j . Quando n_{ij} é suficientemente grande, a distribuição associada à estatística Q_p pode ser aproximada pela distribuição qui-quadrado com $(s - 1)(r - 1)$ graus de liberdade.

É crucial obter χ^2 tabelado, dependente do grau de liberdade e do nível de significância adotado. Se $Q_p \geq \chi^2$ tabelado, rejeitamos H_0 . [Agresti \(2018\)](#) destaca algumas limitações do Q_p , incluindo a necessidade de frequências esperadas serem pelo menos iguais a 5 para uma aproximação válida, e que o teste apenas indica o grau de evidência para uma associação, sem descrever a intensidade dessa associação.

A ausência de associação entre a variável explicativa X e a resposta Y implica na homogeneidade dos grupos. Quando a associação é estabelecida, pode haver interesse em descrever sua intensidade. Para mais detalhes, consulte o livro de [Giolo \(2017\)](#). O teste de Qui-Quadrado de Pearson foi executado utilizando a função `chisq.test()`, com $p < 0,05$ considerado estatisticamente significativo.

3.2.2 Teste exato de Fisher

O teste exato de Fisher, proposto por Ronald Fisher e amplamente discutido por [Fisher \(1922\)](#), [Fisher \(1954\)](#), [Agresti \(1992\)](#), é uma ferramenta estatística empregada para determinar a existência de associação não aleatória entre duas variáveis categóricas. Em situações em que as frequências esperadas em uma tabela de contingência $r \times s$; $r \geq 2$ e $s \geq 2$ (Tabela 6, para tabela 2×2) são menores que 5, comprometendo a aproximação pela distribuição qui-quadrado, recorre-se a métodos exatos.

Tabela 6 – Tabela de contingência 2×2 .

Categorias da variável X	Categorias da variável resposta Y		Totais
	$j = 1$	$j = 2$	
$i = 1$	n_{11}	n_{12}	n_{1+}
$i = 2$	n_{21}	n_{22}	n_{2+}
Totais	n_{+1}	n_{+2}	$n_{++} = n$

Diante disso, aplicam-se métodos exatos para testar as hipóteses:

$$\begin{cases} H_0 : p_{ij} = p_{i+} \times p_{+j}; \forall i, j (\text{Ausência de associação}); \\ H_1 : p_{ij} \neq p_{i+} \times p_{+j}; \text{para algum } (i, j) (\text{Presença de associação}); \end{cases}$$

O teste exato de Fisher se baseia na distribuição hipergeométrica. A probabilidade de obter qualquer particular arranjo de frequências n_{11} , n_{12} , n_{21} , n_{22} , quando os totais marginais-linha e

coluna são fixos, é dada por:

$$p = \frac{\binom{n_{1+}}{n_{11}} \binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}} = \frac{(n_{1+})! (n_{2+})! (n_{+1})! (n_{+2})!}{(n)! (n_{11})! (n_{12})! (n_{21})! (n_{22})!},$$

em que $(k)! = k \times (k-1) \times (k-2) \dots 3 \times 2 \times 1 \times 0!$, sendo $0! = 1$.

Para obtenção da probabilidade associada ao arranjo observado das frequências, o teste utiliza as probabilidades dos arranjos, que forneçam tanta ou mais evidência de associação, e o arranjo observado. A soma dessas probabilidades é comparada com um nível de significância α . Valores dessa soma menores do que α conduzem a rejeição da hipótese nula [Giolo \(2017\)](#).

Em tabela de contingência $s \times r$ é ainda mais comum ocorrer diversos valores esperados menores do que 5. Nesses casos, as probabilidades de interesse são calculadas a partir da distribuição hipergeométrica multivariada, isto é,

$$p = \frac{\prod_{i=1}^s (n_{i+})! \prod_{j=1}^r (n_{+j})!}{n! \prod_{i=1}^s \prod_{j=1}^r (n_{ij})!}.$$

O teste exato de Fisher foi conduzido utilizando a função `fisher.test()`, considerando um valor $p < 0,05$ como estatisticamente significativo.

3.3 Análise de expressão diferencial

O teste de Wilcoxon-Mann-Whitney, também conhecido como teste da soma dos postos de Wilcoxon, foi desenvolvido por [Wilcoxon \(1945\)](#) para comparação de amostras independentes com tamanhos iguais. Posteriormente, [Mann e Whitney \(1947\)](#) ampliaram sua aplicação para amostras de tamanhos diferentes.

Este é um teste não paramétrico, ou seja, menos sensível às suposições sobre a distribuição dos dados e mais robustos em relação a dados que não seguem uma distribuição normal. As observações podem ser variáveis ordinais ou contínuas, e o teste é utilizado para avaliar se as populações das duas amostras podem ser consideradas semelhantes ou não.

Embora seja frequentemente considerado um teste de medianas populacionais, visto que a média não é uma representação adequada dos dados neste contexto, [Hart \(2001\)](#) argumenta em seu artigo, que o tratar exclusivamente como um teste de medianas pode levar a uma análise inadequada. Ele destaca que o teste avalia não apenas a posição central dos dados, mas também é sensível às diferenças na localização e forma das distribuições, exigindo uma abordagem mais abrangente para uma interpretação adequada.

O método para realização do cálculo para o teste é:

- Considere duas amostras de populações independentes;
- Junte as amostras e ordene os valores crescentemente, independentemente do fato de qual população provém;
- Determine a posição de cada observação (postos), em que o posto pode variar de 1 a N , sendo $N = n_1 + n_2$. Em caso de amostras com os mesmos valores atribuem-se a média dos postos correspondentes. Posteriormente, separam-se novamente as amostras originais;
- Calcule a estatística U :

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \text{ e } U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2,$$

em que, n_1 e n_2 são os tamanhos de cada amostra e R_1 e R_2 a soma dos postos de cada amostra, para testar as hipóteses:

$$\begin{cases} H_0 : \text{não existe diferença nas medianas dos grupos (por exemplo, ao comparar T vs. N);} \\ H_1 : \text{existe diferença nas medianas dos grupos.} \end{cases}$$

Se o menor valor de $U_{calc} < U_{tab}$, rejeita-se H_0 . Caso haja muito empates e ambas as amostras tenham tamanhos iguais ou superiores a 10, pode-se fazer aproximação:

$$Z = \frac{U - \mu_U}{\sqrt{\sigma_U^2}} \sim \mathbf{N}(0, 1),$$

em que, $U = \max(U_1, U_2)$, $\mu_U = \frac{n_1 \cdot n_2}{2}$ e $\sigma_U^2 = \frac{n_1 \cdot n_2 (n_1 + n_2 + 1)}{12}$. Porém, no caso de empates no número de ordens deve-se fazer uma correção no cálculo de σ_U^2 :

$$\sigma_U^2 = \frac{n_1 \cdot n_2}{12} \times \frac{N^3 - N - \sum (u_i^3 - u_i)}{N^2 - N},$$

considere u_i o número de observações empatadas, para o i -ésimo valor.

O teste de Wilcoxon-Mann-Whitney foi realizado por meio da função `wilcoxon.test()`, considerando valor $p < 0,05$ como estatisticamente significativo.

O presente estudo envolve a realização de múltiplos testes independentes consecutivos, demandando a implementação de uma correção no erro do tipo I, ou seja, α é utilizando para comparar o conjunto de todas as possíveis comparações e não para as comparações individuais.

Diversos métodos de correção são encontrados na literatura, destacando-se o método de Benjamini-Hochberg (BENJAMINI; HOCHBERG, 1995), Bonferroni (WEISSTEIN, 2004; PERNEGER, 1998), e Holm-Bonferroni (HOLM, 1979).

Neste estudo, optou-se pela aplicação da correção de Bonferroni. Este método é uma técnica de ajuste para comparações múltiplas, sendo especialmente útil quando diversos testes estatísticos, dependentes ou independentes, são realizados simultaneamente. O objetivo é mitigar o risco de ocorrência de falsos positivos.

Assim, o nível de significância original do teste, representado por α , necessita ser reduzido considerando o número total de comparações realizadas. Esse novo nível de significância, denominado α_{novo} , é calculado pela fórmula:

$$\alpha_{novo} = \frac{\alpha_{original}}{n},$$

em que, α_{novo} é o novo nível de significância, $\alpha_{original}$ é o valor original para rejeição da hipótese nula (tipicamente, $p < 0,05$), e n é o número total de comparações ou testes efetuados. Essa abordagem torna o teste mais rigoroso na rejeição da hipótese nula, conferindo-lhe um caráter mais conservador.

3.4 Análises biológicas

Realizaram-se análises biológicas com o intuito de entender biologicamente os DEGs e DE-miRNAs encontrados na análise de expressão diferencial. A seguir, são destacados, por análise, os procedimentos realizados e as respectivas ferramentas empregadas nesse processo.

Análise I consiste em realizar as etapas:

1. **Conversão de Sequências Precursoras de miRNA em Maduras:** Utilizou-se a ferramenta **miEAA2.0** (KERN *et al.*, 2020) (<<https://ccb-compute2.cs.uni-saarland.de/mieaa2/>>).
2. **Análise de Enriquecimento de Vias:** Realizou-se esta análise por meio das ferramentas **DIANA-miRPath** (v.3.0) (VLACHOS *et al.*, 2015) (<<https://dianalab.e-ce.uth.gr/html/mirpathv3/index.php?r=mirpath>>) e utilizando dados da **Enciclopédia de Genes e Genomas de Quioto** (KEGG - *Kyoto Encyclopedia of Genes and Genomes*) (KANEHISA; GOTO, 2000), um recurso de banco de dados para compreensão das funções e utilidades de alto nível do sistema biológico a partir de informações de nível molecular.
3. **Geração de Diagramas de Venn:** Esta etapa foi realizada por meio do pacote Venny (v.2.1) (OLIVEROS, 2007).

Análises II e III englobam as etapas:

1. **Predição de Genes-Alvo dos DE-miRNAs:** Utilizou-se a versão mais recente do Portal de Integração de Dados de miRNAs - **mirDIP 5.3.0.1** (*microRNA Data Integration Portal*) (<<http://ophid.utoronto.ca/mirDIP/>>).

O mirDIP é uma ferramenta bioinformática que oferece uma abordagem global para a previsão de genes direcionados por miRNAs, integrando diversas fontes de dados e utilizando o *Integrated Score* para priorizar as interações previstas dos DE-miRNAs/genes-alvo. Neste estudo, foi adotado o *Score Class Very High*.

2. **Comparação dos Genes-Alvo dos DE-miRNAs com os DEGs:** Identificaram-se os DEGs candidatos por meio da sobreposição entre DEGs e genes-alvo preditos dos DE-miRNAs. Esses genes foram destacados e apresentados em um diagrama de Venn utilizando o pacote VennDiagram (CHEN, 2022).
3. **Análise de Correlação mRNA-miRNA:** Realizou-se uma análise de correlação para os DEGs candidatos cuja expressão se opõe aos DE-miRNAs.

A correlação de Spearman, ou ρ (rho) de Spearman, é uma medida não paramétrica de correlação de postos. Os coeficientes de correlação de Spearman entre as taxas de expressão de DE-miRNA e DEGs candidatos foram calculados usando a função `rcorr()` do pacote Hmisc (Harrell Jr, 2023). Valores de $p < 0,05$ foram considerados estatisticamente significativos.

4. **Seleção de miRNA-mRNA Validados Experimentalmente:** Utilizou-se o banco de dados miRTarBase (<https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/index.php>) para selecionar interações entre DE-miRNAs e DEGs candidatos validados experimentalmente.
5. **Rede Regulatória entre DE-miRNA e DEGs:** Construiu-se uma rede regulatória utilizando o software *Cytoscape* (SHANNON *et al.*, 2003), considerando apenas as interações com mais de uma conexão.
6. **Enriquecimento de Vias:** Realizou-se uma análise para identificar vias enriquecidas nos alvos gênicos, utilizando dados da KEGG (KANEHISA; GOTO, 2000).

- Para miRNA: Utilizou-se a ferramenta **DIANA-miRPath** (v.3.0) (<<https://dianalab.e-ce.uth.gr/html/mirpathv3/index.php?r=mirpath>>) e foram selecionadas as vias KEGG mais enriquecidas, considerando as taxas de descoberta falsas (FDR) $p < 0,05$ ajustado por Benjamini-Hochberg (BENJAMINI; HOCHBERG, 1995).
- Para RNA: Utilizou-se a ferramenta **Enrichr** (<<https://maayanlab.cloud/Enrichr/>>), considerando valor $p < 0,05$ ajustado também por Benjamini-Hochberg.

7. **Montagem e Análise de Rede de Interação Proteína-Proteína (PPI):** Realizou-se a montagem de redes PPI entre alvos gênicos de miRNAs desregulados, utilizando a ferramenta String (<<https://string-db.org/>>). Os resultados foram visualizados com o software *CytoScape*.

A rede é montada a fim de entender a relação funcional dos DEGs candidatos e os fenômenos biológicos envolvidos. Esta ferramenta permite alcançar uma rede global abrangente e objetiva, incluindo interações protéicas diretas (físicas) e indiretas (funcionais). A rede possui várias informações importantes, como:

- Os nós da rede representam proteínas (cada nó representa todas as proteínas produzidas por um único locus gênico codificador de proteínas) e as arestas representam associações proteína-proteína.
- O coeficiente de agrupamento é uma medida de quão conectados os nós da rede estão. Redes altamente conectadas possuem valores próximos de 1.
- O grau médio de um nó é o número de quantas interações uma proteína tem em média na rede e indica a relevância regulatória desta proteína.
- A pontuação combinada é um indicador de confiança, ou seja, a probabilidade de o STRING considerar uma interação verdadeira, conforme as evidências disponíveis. Todas as pontuações são classificadas de 0 a 1, sendo 1 a maior confiança possível.

3.5 Modelos preditivos

Usualmente, métodos estatísticos empregam inferência em pesquisas para compreender como determinados fatores influenciam o comportamento da população em estudo. No entanto, os métodos de AM têm o propósito de prever a variável de interesse por meio de variáveis explicativas, conforme relatado e descrito por [Breiman \(2001b\)](#) em seu artigo. Ele destaca duas abordagens distintas:

- **Cultura de Modelagem de Dados (*Data Modeling Culture*):** Enfatiza a escolha de um modelo correto, visando a interpretabilidade e a validação por meio da qualidade do ajuste e análises de resíduos.
- **Cultura de Modelagem Algorítmica (*Algorithmic Modeling Culture*):** Prioriza a escolha do modelo com maior precisão na validação preditiva, sem considerar necessariamente a explicabilidade do modelo.

Os modelos preditivos constroem um estimador a partir de um conjunto de exemplos rotulados por meio de uma função. Os rótulos assumem valores em um domínio conhecido. Se o domínio for um conjunto nominal, há um problema de classificação, enquanto se for um conjunto de valores, haverá um problema de regressão ([FACELI et al., 2011](#)).

[Lee, Bang e Kim \(2016\)](#) destacam que modelos preditivos clínicos são utilizados para estudar a relação entre resultados futuros ou desconhecidos e o estado de saúde de pacientes com condições específicas.

Um classificador ou regressor é uma função que, dado um exemplo sem rótulos, atribui esse exemplo a uma das possíveis classes ([DIETTERICH, 1998](#)). Em AM, as variáveis do conjunto de dados são chamadas de atributos. O atributo de saída, também conhecido como atributo alvo ou meta, representa os valores ou classes a serem previstos utilizando os demais atributos, denominados atributos de entrada ou atributos previsores ([FACELI et al., 2011](#)).

Segundo [Faceli et al. \(2011\)](#) dado um conjunto de observações de pares $\mathbf{D} = \{(\mathbf{x}_i, f(\mathbf{x}_i)), i = 1, \dots, n\}$, em que f representa uma função desconhecida e x_i é um vetor de dados. Um algoritmo preditivo de classificação aprende uma aproximação $\hat{f}$ e permite estimar a classe de f para novas observações de $\mathbf{x}$, assim:

- **Classificação:** $y_i = f(\mathbf{x}_i) \in \{c_1, \dots, c_m\}$, ou seja, $f(\mathbf{x}_1)$ assume valores em um conjunto discreto, não ordenado.

Dessa forma, cada linha do banco de dados corresponde a uma observação, onde a coluna com a variável resposta apresenta o valor da função f para essa observação, que corresponde ao atributo alvo. Para problemas de classificação, ela é designada como classe, e as colunas restantes são atributos de entrada, utilizados como entrada do algoritmo na realização da predição ([FACELI et al., 2011](#)).

A estimação da qualidade de um modelo preditivo é medida pelo custo associado às previsões do modelo. Para os problemas de classificação, é comum usar a função de custo varia de $0 - 1$, ou seja, o custo que para a previsão incorreta ($\hat{f}(\mathbf{x}) \neq f(\mathbf{x})$) é 1 e o custo para a previsão correta $\hat{f}(\mathbf{x}) = f(\mathbf{x})$ é 0.

A função $\hat{f}$ pode assumir diferentes formas, como combinações lineares e expressões lógicas. Vale ressaltar que $\hat{f}$ é uma aproximação de uma função desconhecida f , ou seja, podem existir $\mathbf{x}$ para os quais $\hat{f}(\mathbf{x}) \neq f(\mathbf{x})$.

3.5.1 Preparação de dados

Normalmente, a aplicação direta de algoritmos de AM nos dados pode ser inviável, requerendo a utilização de técnicas de pré-processamento para corrigir problemas e tornar o conjunto de dados mais adequado para o algoritmo.

Diversas tarefas úteis e/ou necessárias incluem:

- Integração de dados;
- Eliminação manual de atributos;
- Amostragem de dados;
- Balanceamento de dados;
- Limpeza de dados;
- Redução da dimensionalidade;
- Transformação de dados.

Essas atividades estão vinculadas a um processo conhecido como Mineração de Dados (MD) ([WITTEN; FRANK; MARK, 2016](#)), cujo objetivo é extrair conhecimento novo, útil e relevante do conjunto de dados.

3.5.2 Divisão de dados para treino e teste

O ideal e esperado seria treinar os dados no modelo e depois testá-los em um conjunto de dados que o algoritmo nunca aprendeu. Contudo, esperar por novos dados não é sempre viável em termos de tempo e custos ([MUELLER; MASSARON, 2021](#)).

Na maioria das vezes tem-se apenas um conjunto com n objetos, o qual deve ser empregado na indução do preditor e também na sua validação. Calcular o desempenho preditivo do modelo nos mesmo objetos empregados em seu treinamento produz estimativas otimistas ([FACELI *et al.*, 2011](#)). O uso do mesmo conjunto de exemplos no treinamento e validação do preditor é conhecido como ress substituição ([TOUSSAINT, 1974](#)), o erro ou acerto obtido neste tipo de avaliação é denominado aparente.

Para obter estimativas de desempenho preditivo mais confiáveis, é recomendável o uso de métodos de amostragem alternativos, envolvendo subconjuntos de treinamento e teste. Dessa forma, os dados de treinamento são utilizados para induzir e ajustar o modelo, enquanto os dados de teste simulam objetos novos ao preditor, não vistos durante a indução. Os subconjuntos são disjuntos para garantir que as medidas de desempenho sejam obtidas de um conjunto de exemplos diferente daquele usado no aprendizado.

Uma solução para este problema é dividir aleatoriamente seus dados em conjuntos de treinamento e de teste. A divisão comum é entre 70% a 80% para treino e de 30% a 20% para teste, respectivamente. ([IZBICKI; SANTOS, 2020](#); [MUELLER; MASSARON, 2021](#)).

[Baranauskas e Monard \(2000\)](#) destacam que quando um classificador utiliza todo o conjunto de dados, ele terá, em média, um desempenho melhor do que um classificador com apenas uma parte dos dados, assim, esse método tende a superestimar a taxa de erro real. Porém [Michie *et al.* \(1995\)](#) relatam que para conjunto de dados grande isso pode não ser um problema, embora não seja definido quando o conjunto de dados é considerado grande.

O processo de divisão ou amostragem dos dados pode ser realizado por vários métodos, os mais simples e amplamente utilizados são o *Holdout* e a amostragem aleatória, representados nas Figuras [8a](#) e [8b](#), respectivamente.

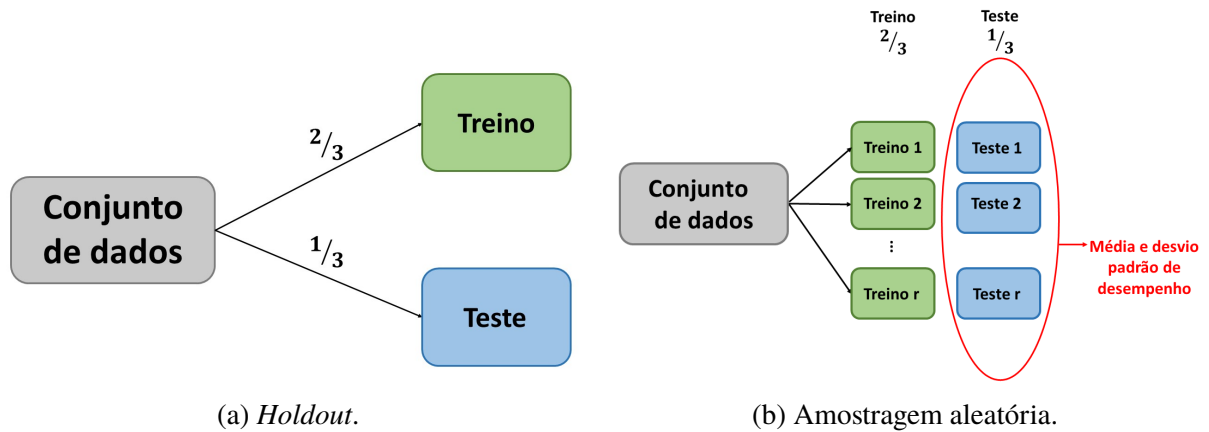


Figura 8 – Métodos de amostragem.

- **Holdout**: o conjunto de dados é dividido em dois conjuntos mutuamente exclusivos: um conjunto de treinamento e um conjunto de teste. Divide-se o conjunto de dados em uma proporção fixa de p para treinamento e $(1 - p)$ para teste. Normalmente considera-se $p > \frac{1}{2}$. Os valores tipicamente utilizados são $p = \frac{2}{3}$ e $1 - p = \frac{1}{3}$. Sua implementação é simples e computacionalmente não custosa.

O *Holdout* não permite avaliar o quanto varia o desempenho de uma técnica quando diferentes combinações de objetos são apresentadas em seu treinamento, dessa forma, é possível que, em uma única divisão realizada, objetos considerados “mais prováveis” tenham sido selecionados para o conjunto teste (FACELI *et al.*, 2011).

- **Amostragem Aleatória (*random subsampling*)**: são realizadas várias partições aleatórias do conjunto de dados em conjuntos de treinamento e teste, a fim de tornar o resultado menos dependente de uma única divisão. Em cada iteração, é selecionada aleatoriamente uma porcentagem dos dados para treinamento, enquanto o restante é destinado ao teste. O desempenho do modelo é avaliado pela média dos resultados de várias partições, tendo assim um método de validação de média. Este método é utilizado para mitigar o efeito da variabilidade na escolha dos conjuntos.

3.5.3 Problema de desbalanceamento

Conjuntos de dados desequilibrados se referem a categorias de classificação que não são representadas igualmente, na mesma quantidade. Algoritmos de classificação geralmente têm seu desempenho prejudicado na presença de dados desbalanceados, pois tendem a favorecer a classificação de novos dados na classe majoritária.

O viés é ainda maior para dados de alta dimensão, em que o número de variáveis excede em muito o número de amostras. É comum a maioria dos conjuntos de dados de classificação não possuir um número exatamente igual de observações nas classes.

Quando não há a possibilidade de disposição de mais exemplos reais para serem adicionados, existem duas técnicas principais (Figura 9) pelas quais os dados desbalanceados podem ser normalmente tratados utilizados durante o processo de treinamento dos dados:

- **Subamostragem:** mantêm-se todos os dados na classe minoritária e diminuindo o tamanho da classe majoritária. A maneira mais simples de fazer isso é confiar na escolha puramente aleatória. Porém, é possível que dados de grande importância para a indução do modelo correto sejam perdidos. Pode ocorrer:
 - **Underfitting:** modelos com baixa taxa de acerto, mesmo no conjunto de treino, configurando subajustamento. Pode ocorrer devido à baixa representatividade dos dados de treinamento ou o modelo usado é muito simples e não identifica padrões;
- **Superamostragem:** adicionam-se exemplos da classe minoritária. Isso pode ser feito adicionando cópias das observações da classe minoritária ou criando versões ligeiramente modificadas delas. Contudo, podem representar situações que nunca ocorrerão, induzindo um modelo inadequado para os dados (KUBAT; KUBAT, 2017). Pode ocorrer:
 - **Overfitting:** baixa capacidade de generalização devido a superajustamento. Neste caso o modelo se especializou demais aos dados de treino;

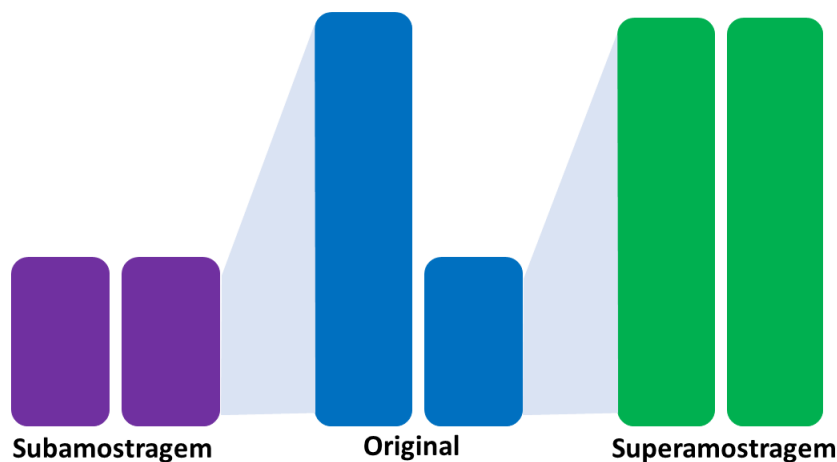


Figura 9 – Subamostragem e sobreamostragem para conjuntos de dados desequilibrados.

Há muitos algoritmos que realizam subamostragem e superamostragem, o trabalho de Fotouhi, Asadi e Kattan (2019) faz uma análise abrangente dos algoritmos e ao final concluem que a medição da classificação média de cada técnica mostrou que o uso de técnicas de sobreamostragem produzia melhores resultados do que subamostragem.

[Kothandan \(2015\)](#) relata que o método de sobreamostragem forneça uma boa medida de desempenho na etapa de treinamento, contudo quando desafiado com conjuntos de dados de teste independentes, o desempenho do classificador deteriorou-se completamente.

Um método de superamostragem popular proposto para melhorar a superamostragem aleatória e bastante utilizado em dados de genética é o SMOTE (*Synthetic Minority Oversampling TEchnique*) ([CHAWLA et al., 2002](#)), que superamostra seu evento raro usando bootstrapping e k -vizinho mais próximo para criar sinteticamente observações adicionais, contudo seu comportamento em dados de alta dimensão ainda não tinha sido investigado minuciosamente. [Blagus e Lusa \(2013\)](#) mostra que, geralmente, o SMOTE pareça benéfico com dados de baixa dimensão, ele não atenua o viés da classificação na classe majoritária para a maioria dos classificadores quando os dados são de alta dimensão e é menos eficaz do que a subamostragem aleatória.

3.5.4 Árvore de decisão

A árvore de decisão é uma técnica de aprendizado de máquina supervisionado amplamente utilizado para prever uma variável de saída com base em diversas variáveis de entrada, também conhecidas como variáveis preditoras. O termo "árvore de decisão" é derivado da sua representação gráfica que se assemelha a uma árvore. Essa técnica é aplicável tanto em problemas de classificação, quando a variável resposta é qualitativa, quanto em problemas de regressão, quando a variável resposta é quantitativa.

Em síntese, a construção da árvore de decisão binária, representada na Figura 10, começa com um nó inicial (nó pai) que contém todo o conjunto de dados de treinamento. O primeiro teste lógico é realizado neste nó com base na variável considerada mais importante para criar dois nós intermediários (nós filhos) o mais homogêneos possível, um esquerdo com os dados satisfazem o teste lógico e um direito com o restante que não satisfaz. O procedimento de divisão é repetido recursivamente usando o mesmo critério até que um ou mais critérios de parada sejam alcançados, como um número mínimo de amostras por nó ou a profundidade máxima da árvore e dessa forma se obtém o nó final (folha da árvore). Denomina-se de ramificação a linha que liga um nó pai com um nó filho. Dessa forma, cada nó representa uma decisão ou um teste sobre uma das variáveis de entrada, e cada ramo representa o resultado desse teste.

A árvore de decisão é uma técnica muito popular devido a seus aspectos positivos, como: flexibilidade ao não assumir distribuição para os dados (métodos não-paramétricos); robustez por serem invariantes a transformações nas variáveis de entrada; seleciona os atributos (variáveis) mais importantes para usar no modelo e tende a ser bastante rigoroso com atributos irrelevantes e redundantes; interpretabilidade fácil, mesmo as decisões complexas podem ser realizadas com decisões mais simples e todas baseadas nos valores dos atributos usados para descrever o modelo; eficiência por ser considerado um algoritmo guloso construído de cima para baixo (*Top-down*), usando a estratégia de dividir para conquistar sem *backtracking*².

² É um tipo de algoritmo que representa um refinamento da busca por força bruta, em que múltiplas soluções

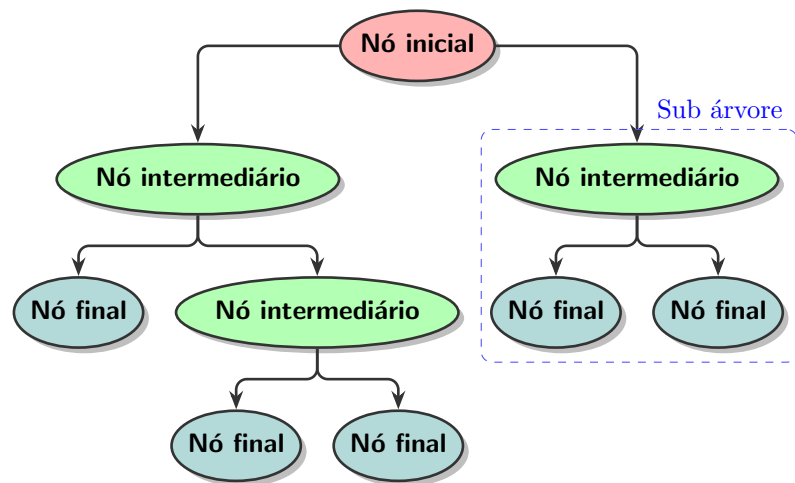


Figura 10 – Ilustração de uma árvore de decisão.

Contudo, há algumas críticas, como: replicações de testes em diferentes ramos de uma árvore de decisão, resultando em uma representação não concisa e baixa acurácia; valores ausentes acabam causando problemas em decidir que ramo seguir durante a decisão; e instabilidade, pois uma pequena variação no conjunto de treinamento pode causar grandes variações na árvore final.

Então, por exemplo, no caso de dois ou mais atributos serem detectados, durante a divisão de um nó específico, como similares, pequenas variações no conjunto de treinamento podem alterar toda a árvore e, assim, todas as subárvores abaixo desse nó podem mudar.

Em relação à estratégia da partição recursiva, os dados são divididos sucessivamente com base em determinado atributo de teste. Depois de algumas divisões, há poucos dados nos quais a decisão se baseia. Com isso, há uma forte tendência a inferências feitas próximo dos nós finais serem menos confiáveis que aquelas efetuadas próximas à raiz (FACELI *et al.*, 2011).

É crucial destacar o grau de compreensibilidade proporcionado ao ser humano pelos processos dos algoritmos. De acordo com Kubat, Bratko e Michalski (1998), os sistemas de aprendizado podem ser classificados em duas categorias principais: "Black-box" (caixa-preta), cuja representação interna pode não ser facilmente interpretada por humanos e não oferece esclarecimento ou explicação do processo de reconhecimento; e sistemas orientados ao conhecimento, que buscam criar estruturas simbólicas compreensíveis por humanos.

Após a construção da árvore, a mesma pode ser utilizada para fazer previsões em novos dados, seguindo o caminho da árvore com base nos valores das variáveis de entrada.

3.5.5 CART

O algoritmo CART (*Classification and Regression Trees*), proposto por Breiman *et al.* (1984), é um método de árvore de decisão para problemas de classificação e regressão

podem ser eliminadas sem serem explicitamente examinadas.

extremamente poderoso, versátil e popular devido ao seu bom desempenho em conjuntos de dados com grande número de observações ou de variáveis e pela resistência a *outliers*.

O algoritmo se caracteriza por gerar árvores binárias compreensíveis, simples e por possuir capacidade de encontrar relação entre os dados, mesmo em situação em que parece não haver (FONSECA, 1994). Além disso, Friedman, Kohavi e Yun (1996) relatam que, cerca de metade do código e 80% dos esforços de programação foram desenvolvidos para falta de valores.

No entanto, as árvores geradas podem ser propensas a *overfitting* (quando o modelo se ajusta demais aos dados de treinamento e não generaliza bem para novos dados) se não forem devidamente controladas. Este problema pode ser contido com técnicas como a poda da árvore (*pruning*) e o ajuste dos hiperparâmetros do modelo.

Todas as árvores e ajustes foram realizados no R utilizando os pacotes rpart (THERNEAU; ATKINSON, 2022), partykit (HOTHORN; ZEILEIS, 2015), Rattle (WILLIAMS, 2011) com a função fancyRpartPlot() e o pacote ggplot2 (WICKHAM, 2016) para plotar gráficos.

Nas próximas seções, serão apresentadas as metodologias de árvore de classificação, árvore de regressão para dados de sobrevivência, os procedimentos de crescimento e poda, e o resumo do modelo resultante gerado pelo algoritmo.

Árvore de Classificação

Árvores de classificação encontram aplicações frequentes em problemas de categorização, podendo ser binários (duas classes) ou multiclases (mais de duas classes). Essa técnica de aprendizado supervisionado auxilia na tomada de decisões com base nas características dos dados. É relevante notar que, além da capacidade de classificação, há um interesse significativo em identificar os fatores que exercem maior influência na determinação dessas categorizações.

A árvore utiliza da lógica condicional para tomar decisões a respeito da classe de um objeto. Dessa forma, cada nó na árvore de classificação representa uma decisão que deve ser tomada com base em um atributo específico (particionamento passo a passo), os ramos que partem do nó representam as possíveis saídas dessa escolha e os nós finais (folhas) representam as diferentes classes/categorias em que um objeto pode ser classificado.

Esse processo pode ser observado na Figura 11. Considera-se t como sendo o nó pai, que contém todo o banco de dados, $t_{Esquerdo}$ e $t_{Direito}$, os nós filhos esquerdo e direito ao nó pai t , respectivamente. Considere a amostra de aprendizagem uma matriz variável $\mathbf{x}$. A regra de divisão normalmente é uma desigualdade, representada por $x_j \leq x_j^R$. A divisão foi feita por meio da variável x_j , com o valor de referência x_j^R que representa o melhor valor de divisão da variável x_j . Portanto, a variável j representa o melhor valor de divisão de qualquer variável (TIMOFEEV, 2004). O $t_{esquerdo}$ representa as observações que seguem a desigualdade e $t_{direito}$ o nó representa o restante das observações do nó pai.

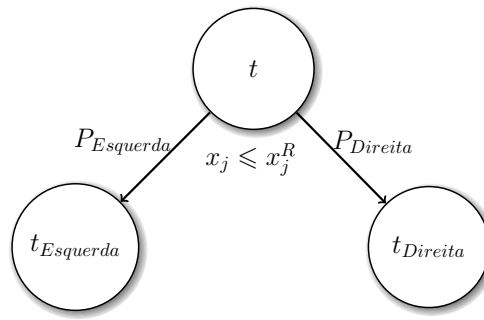


Figura 11 – Algoritmo de divisão do CART.

O nó pai tem como principal objetivo criar nós filhos com homogeneidade máxima e, assim, maximizar a diferença de impureza. A homogeneidade máxima dos nós filhos é definida pela função de impureza ($i(t)$). Como a impureza do nó pai (t) é constante para qualquer uma das divisões possíveis ($x_j \leq x_j^R$), a homogeneidade máxima dos nós filhos esquerdo e direito será equivalente à maximização da função de mudança de impureza ($\Delta i(t)$) (TIMOFEEV, 2004):

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

em que, t_c são os nós filhos juntos. Supõe-se que P_e e P_d são as probabilidades dos nós esquerdos e direitos, respectivamente. Esta fórmula pode ser obtida utilizando as probabilidades dos nós filhos. Dessa forma,

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

Logo,

$$\arg \max_{x_j \leq x_j^R, j=1, \dots, p} [i(t)] - P_e i(t_e) - P_d i(t_d). \quad (3.1)$$

A equação 3.1 revela que o CART explorará todos os valores possíveis de cada variável em busca da melhor "pergunta" para a divisão $x_j < x_j^r$ que maximizará a medida de mudança de impureza $\Delta i(t)$. Portanto, torna-se essencial definir $i(t)$. A regra de divisão de Gini (ou índice de Gini) é a regra mais amplamente usada.

Índice Gini

O algoritmo CART emprega o Índice Gini para conduzir as divisões na construção da árvore de classificação, ou seja, utiliza o Índice Gini para avaliar a impureza do nó e determinar o melhor atributo para dividir o conjunto de dados, conforme expresso pela Equação 3.2:

$$i(t) = 1 - \sum_{i=1}^{nClasses} p_i^2, \quad (3.2)$$

em que, p_i é a probabilidade para cada classe. Quando um atributo é analisado, a impureza média ponderada dos nós descendentes implícitos é subtraída de $i(t)$ e o atributo resultante na maior diminuição de impureza é selecionado (FACELI *et al.*, 2011).

Se há K classes distintas e k e l indicam os índices das classes, $P(k|t)$ indica a probabilidade condicional da classe k se estiver no nó t . Substituindo esta fórmula na Equação 3.1 obtém-se como resultado:

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

O índice Gini mede a probabilidade de classificar errado um elemento da amostra, considerando a distribuição das classes existentes. O resultado pode variar de 0 a 1, em que, 0 indica que todos os elementos pertencem à mesma classe ou apenas existe uma classe, 1 que todos os elementos são distribuídos aleatoriamente as classes e 0,5 denota que os elementos são distribuídos uniformemente em algumas classes. Quanto maior o índice de Gini, maior a impureza do conjunto de dados. É escolhido o atributo com o menor índice de Gini como o melhor para dividir o conjunto de dados, pois isso resultará em uma maior redução da impureza.

Árvore de sobrevivência

A construção de uma árvore de sobrevivência emprega não apenas os pacotes `rpart` (THERNEAU; ATKINSON, 2022) e `partykit` (HOTHORN; ZEILEIS, 2015), mas também faz uso do ambiente `survival` (THERNEAU, 2023). Esse método adota uma abordagem combinada que integra elementos da árvore de regressão, da análise de Kaplan-Meier e do modelo de Cox. Apresenta-se uma breve abordagem sobre esses métodos, a seguir:

Árvore de regressão

Inicialmente, é construída uma árvore de regressão, que ao contrário das árvores de classificação, não possuem classes predefinidas; em vez disso, lidam com vetores de resposta Y que representam os valores para cada observação da variável X .

Na árvore de regressão, a regra de divisão é determinada pelo algoritmo de minimização da soma dos quadrados dos resíduos, $SQ(s,t)$. Esse procedimento implica na minimização das variações esperadas da soma para os dois nós filhos (TIMOFEEV, 2004).

$$\arg \max_{x_j \leq x_j^R, j=1, \dots, p} [P_e Var(Y_e) + P_d Var(Y_d)] \quad (3.3)$$

em que, $Var(Y_e)$ e $Var(Y_d)$ referem-se aos vetores de resposta para os nós filhos esquerdo e direito, respectivamente, enquanto $x_j \leq x_j^R$ representa a condição ótima de divisão conforme definido na Equação 3.3.

É crucial notar que, na árvore de regressão, os nós terminais são delineados por um valor médio e seu erro quadrático médio, podendo ser percebido ou interpretado como uma pseudo-classe.

Pode-se estabelecer uma analogia entre o algoritmo de minimização dos resíduos quadrados e a regra de divisão de Gini, interpretando a função de impureza de Gini por meio da notação de variância. Ao atribuir o valor 1 aos objetos da classe atual e 0 aos demais objetos, a variável Y pode ser interpretada como uma variável aleatória que segue a lei de Bernoulli. A variância da amostra desses valores seria igual a $P(k|t)[1 - P(k|t)]$. Resumindo para K classes, obtém-se a seguinte medida de impureza, representada por $i(t)$:

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

Ao adotar esse critério de divisão, torna-se viável a expansão da árvore de regressão. Contudo, a fim de prevenir questões relacionadas ao sobreajuste, durante a implementação, é indicado analisar a necessidade de realizar poda na árvore.

Análise de Sobrevivência

A análise de sobrevivência é área que engloba diversas técnicas e modelos estatísticos. Neste contexto, a variável de interesse representa o tempo decorrido desde um ponto inicial definido até a ocorrência de um evento específico. A variável aleatória em questão é não negativa e, por sua natureza, corresponde a uma medida de tempo de acompanhamento. O tempo até a ocorrência do evento de interesse é denominado tempo de falha, e as observações resultantes são conhecidas como tempos de vida ou tempos de sobrevivência.

A análise de sobrevivência destaca-se por sua capacidade de lidar com bancos de dados contendo tempos de vida, a principal característica é a presença de censura. Censura ocorre quando não é possível observar o evento de interesse para alguns indivíduos durante todo o período de observação. Diante de censuras, a análise de sobrevivência é essencial, ao permitir incorporar essas informações nas análises estatísticas (COLOSIMO; GIOLO, 2006).

Seja t_j , para $j = 1, \dots, n$, o tempo de falha do j -ésimo indivíduo, sendo o menor valor entre t_j^* (tempo de sobrevivência até a falha) e u_j (tempo de falha), expresso como:

$$t_j = \min(t_j^*, u_j), \quad \text{para } j = 1, \dots, n.$$

A variável que indica a censura (falha) é definida por:

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

Conjuntos de dados de sobrevivência são caracterizados pelos tempos de falha, que podem ser o tempo até a morte de um paciente ou até a cura/recidiva de uma doença, e pelas censuras, constituindo, em conjunto, a variável resposta.

Estimador de Kaplan-Meier

O estimador de Kaplan-Meier (EKM), proposto por [Kaplan e Meier \(1958\)](#), desempenha um papel crucial na estimativa da função de sobrevivência ao longo do tempo, sem exigir suposições paramétricas sobre a distribuição da variável de sobrevivência. Esta abordagem é indispensável para lidar dados de sobrevivência e destaca-se como a técnica mais empregada para estimar a curva de sobrevivência na presença de censura. A expressão geral é apresentada 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,

- $t_1 < t_2 \dots < t_k$, são os k tempos distintos e ordenados de falhas;
- d_j o número de falhas em t_j , $j = 1, \dots, k$;
- n_j o número de indivíduos sob risco em t_j .

De acordo com [Colosimo e Giolo \(2006\)](#), as vantagens do uso do EKM incluem ser um estimador de máxima verossimilhança de $S(t)$, ser não viciado para amostras grandes, ser fracamente consistente e convergir assintoticamente para um processo Gaussiano.

Teste de log-rank

O teste de log-rank ([MANTEL et al., 1966](#)) é amplamente empregado na análise de sobrevivência para dados censurados, sendo um método não paramétrico. Sua aplicação é apropriada quando a razão das funções de taxa de falha entre os grupos comparados permanece aproximadamente constante, indicando característica de taxas de falhas proporcionais.

Inicialmente, realiza-se o teste de igualdade entre duas funções de sobrevivência, considerando $S_1(t)$ e $S_2(t)$. Os tempos de falha distintos, $t_1 < t_2 < \dots < t_k$, referem-se à amostra formada pela combinação de duas amostras individuais. Suponha que, no tempo t_j , ocorram d_j falhas, com n_j indivíduos sob risco antes de t_j na amostra combinada e, respectivamente, d_{ij} e n_{ij} na amostra i ; $i = 1, 2$ e $j = 1, \dots, k$. Os tempos de falha t_j podem ser representados em uma tabela de contingência 2×2 , com d_{ij} falhas e $n_{ij} - d_{ij}$ sobreviventes na coluna i , conforme apresentado na Tabela 7 ([COLOSIMO; GIOLO, 2006](#)).

Tabela 7 – Tabela de contingência obtida no 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

Relativo à falha e censura até o tempo t_j e ao número de falhas no tempo t_j , a distribuição de d_{2j} é uma hipergeométrica:

$$\frac{\binom{n_{1j}}{d_{1j}} \binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}}.$$

A média de D_{2j} é $w_{2j} = \frac{n_{2j}d_j}{n_j}$. A variância de d_{2j} obtida da distribuição hipergeométrica:

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

Assim, a estatística $d_{2j} - w_{2j}$ tem média zero e variância $(V_j)_2$. Caso as k tabelas de contingência forem independentes, um teste adequado para a equivalência das duas funções de sobrevivência pode ser baseado em:

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

A hipótese nula $H_0 : S_1(t) = S_2(t)$ para todo t no espaço de acompanhamento tem distribuição aproximada qui-quadrado com um grau de liberdade. Rejeita-se H_0 , se valor $p \leq \alpha$.

O Modelo de Riscos Proporcionais de Cox

O modelo de Cox, também conhecido como modelo de taxas de falha proporcionais, é uma técnica de regressão utilizada para modelar a taxa de risco em relação às variáveis independentes. É frequentemente empregado na análise de sobrevivência para avaliar o impacto das covariáveis no tempo até o evento de interesse (COX, 1972).

O modelo geral proposto considera p covariáveis, de modo que X seja um vetor com os componentes $\mathbf{X} = (x_1, \dots, x_p)'$. A expressão na forma genérica proposta considera (COX, 1972):

$$\lambda(t|\mathbf{x}) = \lambda_0(t)g(\mathbf{x}'\beta) \quad (3.4)$$

no qual, $g(\mathbf{X}'\beta)$ é uma função não-negativa que deve ser determinada, tal que $g(0) = 1$. A Expressão 3.4 é constituído pelo produto de dois componentes, um não paramétrico e outro

paramétrico. O componente não-paramétrico, $\lambda_0(t)$, não é estabelecido e é uma função não-negativa do tempo. É usualmente denotado função de taxa de falha de base, pois $\lambda(t|\mathbf{x}) = \lambda_0(t)$ quando $\mathbf{x} = 0$. O componente paramétrico é frequentemente utilizado na seguinte forma multivariada (COX, 1972):

$$g(\mathbf{x}'\beta) = \exp \{\mathbf{x}'\beta\} = \exp \{\beta_1 x_1 + \dots + \beta_p x_p\}, \quad (3.5)$$

em que $\beta^T = (\beta_1, \dots, \beta_p)$ é o vetor de parâmetros ligados às covariáveis. Assim, garante-se que $\lambda(t|\mathbf{x})$ seja não negativa. Observe que a constante β_0 , presente nos modelos paramétricos, não aparece no componente não-paramétrico mostrado em 3.5, porque a presença do componente não-paramétrico no modelo pode absorver este termo constante.

O modelo é também denotado o modelo de taxas de falha proporcionais porque a razão das taxas de falha de dois indivíduos diferentes é constante ao longo do tempo, assim, a razão das funções da taxa de falha para os indivíduos i e j é (GIOLO, 2017):

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp \{\mathbf{x}'_i \beta\}}{\lambda_0(t) \exp \{\mathbf{x}'_j \beta\}} = \exp \{\mathbf{x}'_i \beta - \mathbf{x}'_j \beta\}, \quad (3.6)$$

não depende do tempo. Assim, o pressuposto para usar o modelo de regressão de Cox, é que as taxas de falha sejam proporcionais.

É necessário ajustar o modelo de Cox, pois o método de máxima verossimilhança é usado para fazer inferências acerca dos parâmetros do modelo. Porém, a presença do componente não-paramétrico $\lambda_0(t)$ na função de verossimilhança torna o método de máxima verossimilhança inapropriado. Por isso, Cox (1975) propôs o método de máxima verossimilhança parcial, detalhado no livro de Giolo (2017).

A verossimilhança parcial é utilizada para maximizar a informação amostral na função de verossimilhança, levando em consideração os dados censurados. Isso significa que, ao calcular a verossimilhança parcial, o modelo leva em conta apenas as contribuições dos eventos que são observados, ao mesmo tempo em que trata os eventos censurados de maneira apropriada. Essa abordagem é crucial para garantir uma estimativa adequada dos parâmetros do modelo quando há censura nos dados de sobrevivência.

Assim, a abordagem combinada permite que a árvore de sobrevivência capture padrões complexos nos dados de sobrevivência, considerando tanto a estrutura da árvore de decisão quanto os aspectos específicos da análise de sobrevivência. Isso resulta em um modelo mais robusto para lidar com dados de sobrevivência em contextos diversos.

Crescimento e poda

Árvores induzidas sem poda apresentam algumas desvantagens, tais como a tendência ao crescimento excessivo, resultando em árvores grandes e de difícil compreensão. Além disso, estas

árvores podem classificar novos objetos de maneira pouco confiável. As estatísticas calculadas nos nós mais profundos tendem a refletir mais o conjunto de treinamento, resultando em sobreajuste e ao aumento do erro devido à variância do classificador.

A poda de uma árvore de decisão implica em substituir os nós mais profundos por folhas, o que pode resultar na classificação incorreta de alguns exemplos do conjunto de treinamento. No entanto, os benefícios desse processo se tornam evidentes ao classificar novos exemplos não utilizados durante a construção da árvore. Portanto, a poda é considerada uma etapa crucial na construção de árvores de decisão, pois geralmente leva a erros de generalização menores. Existem dois principais métodos para realizar a poda:

- Pré-poda: a árvore pára de crescer assim que algum critério definido for satisfeito;
- Pós-poda: uma árvore completa, superajustada aos dados de treinamento, é gerada e podada.

Pós-Poda

A pós-poda é o método mais comum para podar as árvores de decisão e é o método utilizado pelo CART. O trabalho de [Quinlan \(1988\)](#) relata que “construir e podar uma árvore é mais lento, mas mais confiável”.

A pós-poda é realizada por meio do custo-complexidade (CP) ([BREIMAN et al., 1984](#)). Para isso, considere $T_1, T_2, \dots, T_k$ como sub-árvores de uma T_{max} . Para qualquer $T \leq T_{max}$, a complexidade é dada como $|T|$. O parâmetro α representa a importância relativa do tamanho da árvore em relação à taxa de erro, sempre $\alpha \geq 0$ um número real. Assim, o custo-complexidade da árvore é medido por:

$$R_\alpha(T) = R(T) + \alpha|T|,$$

em que, $|T|$ é o número de nós terminais (folhas) de uma árvore T e o custo de $T = R(T) = \sum_{i=1}^k \frac{T_i}{T} R(T_i)$. O custo de complexidade de uma árvore T é, então, uma combinação linear do custo e de sua complexidade. Assim, o objetivo é encontrar $T_\alpha \leq T_{max}$ que minimize $R_\alpha(T)$, isso para cada α , ou seja,

$$R_\alpha(T_\alpha) = \operatorname{argmin}_{T \leq T_{max}} R_\alpha(T).$$

Quando, α for pequeno, a penalidade pelo número alto de nós filhos será pequena e T_α (sub-árvore podada por α) será grande. Em contrapartida, para α grande, T_α conterá apenas o nó raiz e a árvore T_{max} terá sido completamente podada.

O número de sub-árvores de T é finito, dessa forma, a poda gera uma sequência finita de subárvores aninhadas $T_1, T_2, T_3, \dots$ que contém terminais progressivamente menores. A árvore podada selecionada é aquela que minimiza $R_\alpha(T)$ na sequência de sub-árvores. Com isso, tem-se

3 propriedades importantes, provadas em (BREIMAN *et al.*, 1984) e discutidas em Therneau, Atkinson *et al.* (1997):

1. Se T_1 e T_2 são sub-árvores de T com $R_\alpha(T_1) = R_\alpha(T_2)$, então T_1 é uma subárvore de T_2 ou T_2 é uma subárvore de T_1 ; e, portanto, $|T_1| < |T_2|$ ou $|T_2| < |T_1|$;
2. Se $\alpha \leq \beta$, então $T_\alpha = T_\beta$ ou T_α é uma sub-árvore estrita de T ;
3. Dado algum conjunto de números $\alpha_1, \alpha_2, \dots, \alpha_m$; tanto $T_{\alpha_1}, \dots, T_{\alpha_m}$ quanto $R(T_{\alpha_1}), \dots, R(T_{\alpha_m})$ podem ser calculados de forma eficiente.

Assim, pode-se dizer que todos os valores possíveis de α (parâmetro de complexidade que minimiza o custo-complexidade) podem ser agrupados em m intervalos, os quais $m \leq |T|$,

$$I_1 = (0, \alpha_1], \quad I_2 = (\alpha_1, \alpha_2], \dots, I_m = (\alpha_{m-1}, \infty),$$

em que, todos $\alpha \in I_i$ provém da mesma sub-árvore podada. Assim, é empregada validação cruzada para escolher o valor de α que minimiza $R_\alpha(T)$, por meio dos seguintes passos:

1. Ajuste a árvore no conjunto de treinamento e calcule $I_1, \dots, I_m$. Seja

$$\beta_1 = 0, \quad \beta_2 = \sqrt{\alpha_1 \alpha_2}, \dots, \quad \beta_{m-1} = \sqrt{\alpha_{m-2} \alpha_{m-1}}, \quad \beta_m = \infty,$$

cada β_i é um valor típico de I_i ;

2. Divide-se os dados em s grupos $G_1, \dots, G_s$, cada um de tamanho s/n . Para cada grupo G_i , faça:
 - a) Ajuste a árvore para todos os G grupos, exceto para o G_i . Depois, determine $T_{\beta_1}, \dots, T_{\beta_m}$;
 - b) Obtenha a classe predita para as observações em G_i , para cada T_{β_j} em $1 \leq j \leq m$. Depois, calcule o custo-complexidade.
3. Some os G_i para obter a estimativa de custo-complexidade para cada β_j . Para β com menor custo-complexidade estimado, modele a árvore T_β , sendo esta a escolhida como a árvore podada.

Resumo do Modelo

O modelo de árvore de classificação gerado pela função `rpart()` do pacote **rpart** (THERNEAU; ATKINSON, 2022) oferece uma ampla gama de informações, incluindo uma tabela resumo que abrange:

- **Erro associado ao nó raiz (root node error):** representa a porcentagem de registros classificados corretamente no primeiro nó da divisão;

- **CP:** indica os valores dos parâmetros de complexidade. O modelo com o menor CP permite mais divisões (*nsplit*), enquanto o maior corresponde a uma árvore com apenas um nó raiz;
- **nsplit:** o número de divisões da árvore;
- **Rel error:** o erro relativo na previsão dos dados usados para estimar o modelo, calculado como $1 - R^2$ raiz do erro quadrático médio;
- **Xerror:** o erro de validação cruzada gerado pela validação cruzada integrada do *rpart*;
- **Xstd:** o erro padrão.

Cada linha na tabela representa uma altura ou profundidade distinta da árvore. À medida que a árvore se aprofunda, o erro de classificação durante o treinamento geralmente diminui. Entretanto, esse benefício é acompanhado por um aumento no risco de *overfitting*.

O erro de validação cruzada normalmente aumenta à medida que a árvore cresce além do nível ideal. Para a pós-poda, recomenda-se selecionar o valor de CP associado a uma árvore mais curta que minimize a taxa de erro de validação cruzada (*xerror*).

Variáveis Importantes

A avaliação da importância das variáveis desempenha um papel crucial no entendimento dos resultados gerados pelo pacote *rpart*. Essa medida é fundamentada em dois elementos essenciais, que, quando combinados, oferecem uma perspectiva abrangente sobre o papel de cada variável na construção da árvore de decisão.

- **Frequência de Divisão:** Este critério refere-se à frequência com que uma variável é selecionada para realizar divisões na árvore. Variáveis que contribuem regularmente para as divisões são consideradas mais relevantes.
- **Melhora da Pureza (ou Redução do Erro):** Este aspecto mede a eficácia das divisões realizadas com base em uma variável específica. Se uma variável conduz a divisões que resultam significativamente na redução da impureza, é classificada como mais importante.

Assim, as importâncias oferecem uma visão relativa da contribuição de cada variável na construção da árvore. Variáveis mais importantes serão associadas a valores mais elevados na medida de importância. A análise dessas medidas proporciona *insights* valiosos para identificar quais variáveis desempenham um papel crucial na previsão do modelo de árvore construído.

É importante ressaltar o conceito de variáveis substitutas, sendo as variáveis que não são escolhidas para estar na árvore, mas que estiveram próximas de serem incluídas. Por exemplo, se o algoritmo decide dividir com a variável x_1 em uma determinada divisão, e existe outra variável x_2 que seria quase tão eficaz quanto x_1 para a divisão, então x_2 é designada como substituta

e recebe a mesma importância atribuída a x_1 . Assim, as variáveis importantes podem incluir aquelas que não foram efetivamente usadas na divisão da árvore, mas que são relevantes.

A medida de importância das variáveis é derivada de uma fórmula que considera a frequência de divisões de uma variável e a melhora resultante na impureza ou erro da árvore. A fórmula geral para o cálculo da importância é expressa como:

$$\text{Importância} = \sum_j \text{Impureza}(\text{nó } j) \times \text{Frequência de divisão (nó } j)$$

em que, $\text{Impureza}(\text{nó } j)$ representa uma medida da impureza ou erro no nó j , que pode ser o índice de Gini ou a soma dos quadrados. Essa fórmula é aplicada a cada nó da árvore, resultando em valores brutos (que podem ser normalizados, se desejado). As importâncias brutas oferecem uma avaliação direta da contribuição de cada variável para a árvore de decisão.

3.5.6 Florestas Aleatórias

Florestas aleatórias são um método de *ensemble* que constrói muitas árvores de decisão utilizadas para classificar novos exemplos por meio de votação majoritária. Elas são altamente eficazes em muitos casos e se destacam por sua habilidade em lidar com sobreajuste, atenuado pela diversidade introduzida pelas múltiplas árvores e pela seleção aleatória de características.

Florestas Aleatórias são versáteis, lidando tanto com problemas de classificação quanto com regressão supervisionada. Neste trabalho, utiliza-se a classificação binária. No entanto, é crucial destacar que RF podem abordar problemas multiclasse, como o trabalho de [Díaz-Uriarte e Andrés \(2006\)](#).

Contudo, a complexidade resultante torna desafiador para os usuários interpretarem o motivo de uma decisão específica. Florestas Aleatórias são frequentemente descritas como caixas pretas, indicando que os modelos são complexos o suficiente para que os usuários não possam compreender totalmente seu funcionamento interno apenas examinando suas entradas e saídas.

Formalmente Florestas Aleatórias trata-se de um classificador composto por uma coleção de árvores $h_k(x)$, onde $k = 1, 2, \dots, L$ ([BREIMAN, 2001a](#)). Cada árvore é construída com base em amostras aleatórias independentes e identicamente distribuídas, representadas por T_k . No momento da classificação, cada árvore emite um voto, e a classe mais frequente dentre esses votos é atribuída à entrada x .

Florestas Aleatórias emprega o uso de amostras *bootstraps* ([EFRON, 1992](#)) para gerar conjuntos de treinamento aleatórios. A cada iteração, um novo conjunto de treinamento é construído com reposição, selecionado a partir do conjunto original. A construção de cada árvore na floresta ocorre utilizando este subconjunto específico e uma escolha aleatória de atributos. Em cada nó da árvore, um subconjunto de m atributos é aleatoriamente selecionado e avaliado. O atributo mais relevante é, então, escolhido para realizar a divisão no nó. Importante destacar que

o valor de m permanece constante para todos os nós. As árvores crescem sem a aplicação de poda. A Figura 12 ilustra esse funcionamento.

O erro de classificação em Florestas Aleatórias é influenciado pela força das árvores individuais que a compõem, podendo ser entendido como uma medida do desempenho de cada árvore. Aumentar a força das árvores individuais tende a reduzir a taxa de erro da floresta. Além disso, a correlação entre quaisquer duas árvores na floresta também desempenha um papel, visto que uma baixa correlação tende a diminuir a taxa de erro de classificação (BREIMAN, 2001a; BREIMAN; CUTLER, 2004).

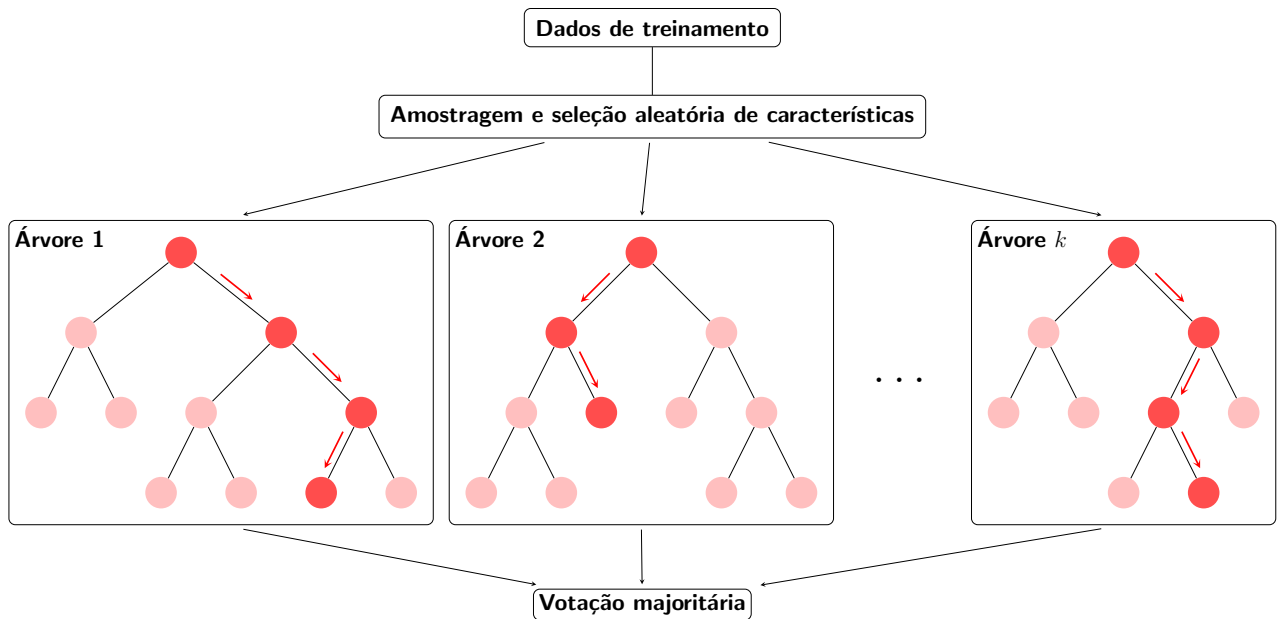


Figura 12 – Ilustração Florestas Aleatórias.

A seguir, o estudo conduzido por Biau e Scornet (2016) revisita os avanços teóricos e metodológicos mais recentes em relação a Florestas Aleatórias. Os autores apresentam e explicam a classificação binária e como funciona o algoritmo, detalhando-a resumidamente abaixo:

Classificação binária

Para o problema de classificação binária, considere que a resposta aleatória Y assume valores em $\{0, 1\}$ e, dado $\mathbf{X}$, é preciso adivinhar o valor de Y (DEVROYE; GYÖRFI; LUGOSI, 2013). Um classificador, ou regra de classificação, m_n é uma função mensurável de Borel de $\mathbf{X}$ e D_n que tenta estimar o rótulo Y de $\mathbf{X}$ e D_n . Diz-se que o classificador é consistente se sua probabilidade de erro m_n , segundo Biau e Scornet (2016):

$$L(m_n) = \mathbb{P}[m_n(\mathbf{X}) \neq Y] \xrightarrow{n \rightarrow \infty} L^*,$$

em que, L^* é o erro ótimo, mas desconhecido, do classificador de Bayes:

$$m^*(\mathbf{x}) = \begin{cases} 1 & \text{se } \mathbb{P}[Y = 1 \mid \mathbf{X} = \mathbf{x}] > \mathbb{P}[Y = 0 \mid \mathbf{X} = \mathbf{x}], \\ 0 & \text{caso contrário.} \end{cases}$$

No contexto de classificação, o classificador Florestas Aleatórias é obtido por maioria de votos entre as árvores de classificação, ou seja,

$$m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n) = \begin{cases} 1 & \text{se } \frac{1}{M} \sum_{j=1}^M m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n) > 1/2, \\ 0 & \text{caso contrário.} \end{cases}$$

Se uma folha representa a região A , então um classificador de árvore aleatório assume a forma simples:

$$m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n) = \begin{cases} 1 & \text{se } \sum_{i \in \mathcal{D}_n^*(\Theta_j)} \mathbf{1}_{\mathbf{x}_i \in A, Y_i=1} > \sum_{i \in \mathcal{D}_n^*(\Theta_j)} \mathbf{1}_{\mathbf{x}_i \in A, Y_i=0}, \mathbf{x} \in A, \\ 0 & \text{caso contrário} \end{cases},$$

em que, $\mathcal{D}_n^*(\Theta_j)$ contém os pontos de dados selecionados na etapa de reamostragem, ou seja, em cada folha, uma votação majoritária é tomada sobre todos $(\mathbf{X}_i, \mathbf{Y}_i)$ para qual $\mathbf{X}_i$ está na mesma região. Os empates são desfeitos, por convenção, em favor da classe 0.

Algoritmo

Para duas classes, pegue $\mathbf{Y} \in \{0, 1\}$ e considere uma única árvore sem etapa de subamostragem. Para qualquer célula genérica A , seja $p_{0,n}(A)$ (resp., $p_{1,n}(A)$) seja a probabilidade empírica, dado um ponto de dados em uma célula A , de que ela tenha rótulo 0 (resp., rótulo 1). Ao perceber que $\bar{Y}_A = p_{1,n}(A) = 1 - p_{0,n}(A)$, o critério de classificação CART-split lê, para qualquer $(j, z) \in C_A$

$$L_{\text{class},n}(j, z) = p_{0,n}(A)p_{1,n}(A) - \frac{N_n(A_L)}{N_n(A)} \times p_{0,n}(A_L)p_{1,n}(A_L) - \frac{N_n(A_R)}{N_n(A)} \times p_{0,n}(A_R)p_{1,n}(A_R).$$

Este critério é baseado na medida de impureza de Gini $2p_{0,n}(A)p_{1,n}(A)$ (BREIMAN *et al.*, 1984). Para classificar um ponto de dado que cai na célula A , usa-se a regra que atribui um ponto, uniformemente selecionado de $\{\mathbf{X}_i \in A : (\mathbf{X}_i, Y_i) \in \mathcal{D}_n\}$, para rotular ℓ com probabilidade $p_{\ell,n}(A)$, para $j \in \{0, 1\}$. A probabilidade estimada de que o item realmente tenha rótulo é ℓ , $p_{\ell,n}(A)$. Portanto, o erro estimado sob esta regra é o índice de Gini $2p_{0,n}(A)p_{1,n}(A)$. Note, no entanto, que a estratégia de previsão difere na classificação e na regressão: no regime de classificação, cada árvore usa um voto majoritário local, enquanto na regressão a previsão é feita por uma média local.

Variáveis Importantes

A métrica "*Mean Decrease Gini*" é uma medida crucial no pacote *randomForest* (LIAW; WIENER, 2002) no R, utilizada para avaliar a importância das variáveis em modelos de Florestas Aleatórias.

Para compreender o *Mean Decrease Gini*, é essencial lembrar o índice Gini, quanto menor o índice Gini, mais "puro" é o nó, indicando que as observações pertencem principalmente a uma única classe. O *Mean Decrease Gini* representa a diminuição média no coeficiente de Gini e é fundamental para avaliar como cada variável contribui para a homogeneidade dos nós e folhas na Florestas Aleatórias resultante.

Esta métrica é uma ferramenta valiosa para entender a importância de cada variável na classificação dos dados. As variáveis são apresentadas em ordem decrescente de importância, indicando sua influência na construção do modelo.

No contexto de um problema de duas classes, a importância é calculada com base no índice Gini (G) e é expressa como a média ponderada das reduções de Gini em todas as divisões que envolvem o preditor em questão. Essa abordagem fornece uma medida consistente e confiável da importância das variáveis em Florestas Aleatórias. A importância é então calculada como:

$$I = G_{pai} - G_{divisão1} - G_{divisão2},$$

Além disso, ao considerar a propriedade da torre: $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$, considerando X e Y como sendo variáveis aleatórias, a importância média condicional de cada variável é calculada, refletindo sua contribuição específica quando usada em comparação com outras variáveis do mesmo grupo na floresta. Essa abordagem robusta de avaliação de importância destaca o papel essencial do "*Mean Decrease Gini*" na análise de modelos de Florestas Aleatórias.

3.5.7 Hiperparâmetros

Os algoritmos de aprendizado de máquina ajustam automaticamente seus parâmetros internos, chamados de "parâmetros do modelo". No entanto, antes do início do processo de aprendizagem, existem outros parâmetros ajustados, conhecidos como "hiperparâmetros" (EL-GELDAWI *et al.*, 2021).

A grande maioria dos algoritmos de aprendizado de máquina é composta por um conjunto de hiperparâmetros, e a escolha desses valores deve ser feita cuidadosamente, pois geralmente afetam consideravelmente o desempenho preditivo do modelo utilizado (BISCHL *et al.*, 2021).

A determinação dos valores dos hiperparâmetros pode ser feita por meio das recomendações fornecidas pelos pacotes de *software*, manualmente utilizando informações de referências, experiência prévia ou método de tentativa e erro.

Os algoritmos utilizados nesse trabalho terão seus hiperparâmetros avaliados conforme os valores apresentados na Tabela 8, baseando-se no trabalho de Probst, Boulesteix e Bischl

(2019), que estabelecem padrões e sugerem medidas gerais para hiperparâmetros de algoritmos, [Biau e Scornet \(2016\)](#), e no *software* R. As breves definições apresentadas na Tabela 8 estão detalhadas em [Therneau e Atkinson \(2022\)](#) e [Liaw e Wiener \(2002\)](#).

Tabela 8 – Hiperparâmetros dos algoritmos *rpart* e *randomForest*.

Algoritmo	Hiperparâmetro	Descrição	Tipo	Variação	Padrão
rpart	cp	Parâmetro de complexidade	numérico	0 – 1	0.01
	maxdepht	Define a profundidade máxima de qualquer nó da árvore final	inteira	1 – 30	30
	minsplit	Número mínimo de observações que devem existir em um nó para seja realizada divisão	inteira	1 – 60	20
	minbucket	Número mínimo de observações em qualquer nó terminal	inteira	1 – 60	$\frac{\text{minsplit}}{3}$
randomForest	ntree	Número de árvores	numérica		500
	mtry	Quantidade de variáveis aleatórias que o algoritmo sorteia	numérica		$\sqrt{p^*}$
	replace	Reamostragem com reposição	logico		TRUE
	nodesize	Quantidade mínima de observações no nó final	numérica		1
	maxnodes	Quantidade máxima no nó terminal	numérico		máximo possível

p^* = número de variáveis.

Por fim, [Biau e Scornet \(2016\)](#) destaca que, embora não haja garantia teórica para suportar os valores padrão dos parâmetros, eles são fáceis de ajustar sem exigir um conjunto de validação independente.

3.5.8 Avaliação de modelos preditivos

Em AM é desafiador prever qual técnica superará outra na resolução de um determinado problema. Diversos algoritmos podem ser considerados para abordar uma questão específica, e mesmo ao escolher um único algoritmo, pode ser necessário ajustar seus parâmetros livres, resultando na obtenção de múltiplos modelos para os mesmos dados ([FACELI et al., 2011](#)).

A experimentação é uma etapa crucial em AM. São necessários procedimentos cuidadosos para assegurar a correção, validade e reprodutibilidade dos experimentos realizados, e, por último, para fundamentar as conclusões obtidas a partir desses resultados.

A avaliação experimental dos modelos de AM pode ser conduzida de várias maneiras, considerando critérios como acurácia, compreensibilidade do conhecimento extraído, tempo de aprendizado, entre outros. É essencial escolher métricas de avaliação que se alinhem aos objetivos específicos do problema em questão, garantindo uma análise abrangente do desempenho.

Matriz de Confusão

A matriz de confusão é uma tabela que permite a visualização do desempenho de um algoritmo de classificação. Para o caso de duas classes, geralmente uma é denotada como positiva (+) e a outra como negativa (-). A estrutura típica da matriz de confusão é ilustrada na Tabela 9:

A partir da matriz de confusão é possível obter outras medidas de desempenho, entre elas ([MONARD; BARANAUSKAS, 2003](#)):

Tabela 9 – Exemplo de uma matriz de confusão para um problema com duas classes.

		Classe Real	
		Positiva (+)	Negativa (-)
Classe Predita	Positiva (+)	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativa (-)	Falso Negativo (FN)	Verdadeiro Negativo (VN)

- **Taxa de falsos negativos (TFN):** proporção de exemplos da classe positiva incorretamente classificados pelo preditor $\hat{f}$.

$$\text{err}_+(\hat{f}) = TFN(\hat{f}) = \frac{FN}{VP + FN}.$$

- **Taxa de falsos positivos (TFP):** proporção de exemplos da classe negativa incorretamente classificados por $\hat{f}$

$$\text{err}_-(\hat{f}) = TFP(\hat{f}) = \frac{FP}{FP + VN}.$$

- **Taxa de erro total:** soma dos valores da diagonal secundária da matriz, dividida pela soma dos valores de todos os elementos da matriz.

$$\text{err}(\hat{f}) = \frac{FP + FN}{n}.$$

considerar: $n = VP + FN + FP + VN$.

- **Taxa de acerto ou acurácia total:** soma dos valores da diagonal principal da matriz, dividida pela soma dos valores de todos os elementos da matriz, ou seja, corresponde a proporção total de acertos de valores positivos e negativos. Quanto maior a acurácia de um teste, mais preciso em determinar quem tem ou não a doença.

$$\text{ac}(\hat{f}) = \frac{VP + VN}{n}.$$

- **Precisão:** proporção de exemplos positivos classificados corretamente entre todos aqueles classificados como positivos por $\hat{f}$.

$$\text{prec}(\hat{f}) = \frac{VP}{VP + FP}.$$

- **Sensibilidade ou taxa de verdadeiros positivos (TVP):** taxa de acerto na classe predita, ou seja, corresponde ao percentual de resultados positivos dentre as pessoas que tem uma determinada doença ou condição clínica.

$$\text{sens}(\hat{f}) = \text{rev}(\hat{f}) = TVP(\hat{f}) = \frac{VP}{VP + FN}.$$

- **Especificidade:** taxa de acerto na classe negativa, ou seja, é a capacidade do mesmo teste ser negativo nos indivíduos que não apresentam a doença que esta sendo investigada.

$$\text{esp}(\hat{f}) = \frac{VN}{VN + FP} = 1 - TFP.$$

A precisão é frequentemente considerada uma métrica que avalia a exatidão de um modelo, enquanto a sensibilidade é vista como uma medida de sua completude. No entanto, é comum não discutir precisão e sensibilidade separadamente; em vez disso, elas são frequentemente combinadas em uma única medida chamada F_1 :

$$F_1(\hat{f}) = \frac{2 \times \text{prec}(\hat{f}) \times \text{sens}(\hat{f})}{\text{prec}(\hat{f}) + \text{sens}(\hat{f})}.$$

No R é possível utilizar a função `confusionMatrix()` do pacote `caret` (KUHN, 2022) e `accuracy.meas()` do pacote `ROSE` (LUNARDON; MENARDI; TORELLI, 2014). O pacote `dplyr` (WICKHAM *et al.*, 2022) também foi utilizado para manipulação de dados.

Curva ROC

A Curva ROC (*Receiver Operating Characteristics*) é utilizada para avaliar classificadores em problemas binários (FAWCETT, 2006). Além de ser frequentemente empregada para propor e desenvolver novas técnicas em AM (SPACKMAN, 1989; PRATI; FLACH, 2005).

O gráfico ROC é bidimensional, representado no espaço ROC, com os eixos X e Y indicando TFP e TVP, respectivamente (Figura 13). Destacando alguns aspectos essenciais, a linha tracejada na diagonal representa classificadores que realizam previsões aleatórias. Assim, qualquer classificador abaixo dessa linha pode ser considerado pior que o aleatório. O ponto (0,1) representa a classificação perfeita (céu ROC), enquanto o ponto (1,0) representa a classificação oposta (inferno ROC). O ponto (1,1) sugere classificações sempre positivas, enquanto o ponto (0,0) sugere classificações sempre negativas.

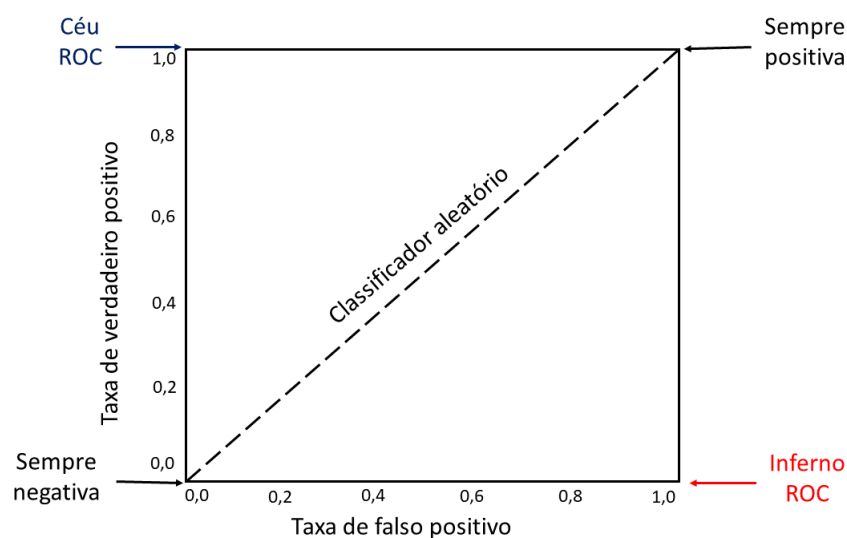


Figura 13 – Espaço ROC.

A avaliação de classificadores por meio de pontos no espaço ROC é possível, mas a abordagem mais comum envolve a criação de uma Curva ROC. Isso requer utilizar algum

esquema de ranqueamento nas classificações. Muitos classificadores geram valores contínuos, sendo discretizados para a saída da classificação. A introdução de limiares permite determinar se a previsão corresponde à classe positiva ou negativa. Ao variar esses limiares, diferentes valores de TFP e TVP são obtidos. Ao plotar esses valores para diferentes limiares, obtemos a Curva ROC (Figura 14), buscando que todos os nós se agrupem no canto superior esquerdo.

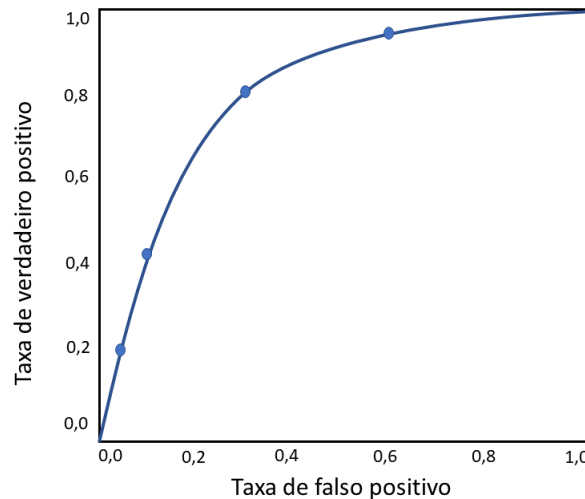


Figura 14 – Curva ROC.

A avaliação do desempenho da Curva ROC frequentemente se baseia em uma medida única: a Área Sob a Curva ROC (AUC, do inglês *Area Under ROC Curve*). Valores próximos a 1 são considerados indicativos de melhor desempenho.

A análise ROC oferece vantagens significativas, permitindo medidas de desempenho independentes do limiar de classificação e de custos associados a classificações incorretas e à distribuição das classes (PRATI, 2006). Apesar de sua limitação original a problemas de classificação binária, a análise ROC tem sido amplamente aplicada. Utilizaram-se os pacotes *Metrics* (HAMNER; FRASCO, 2018) e *yardstick* (KUHN; VAUGHAN; HVITFELDT, 2022).

4 Resultados e Discussões

Neste capítulo são apresentados os resultados e discussões do estudo. Na primeira análise, realizaram-se comparações descritivas, biológicas e estatísticas dos resultados de três métodos de identificação de expressão diferencial: Teste de Wilcoxon-Mann-Whitney, Deseq2 e EdgeR. Essas análises foram conduzidas utilizando o banco de dados LUAD miRNA-seq precursores.

As análises II e III abordaram estudos e ferramentas de bioinformática para investigar quatro bancos de dados de LUAD (miRNA-seq e RNA-seq) e LUSC (miRNA-seq e RNA-seq). O objetivo foi identificar miRNAs e genes com potencial para distinguir T vs. N em LUAD e LUSC, e diferenciar as doenças LUSC vs. LUAD.

4.1 Análise I: comparação de métodos de análise de expressão diferencial

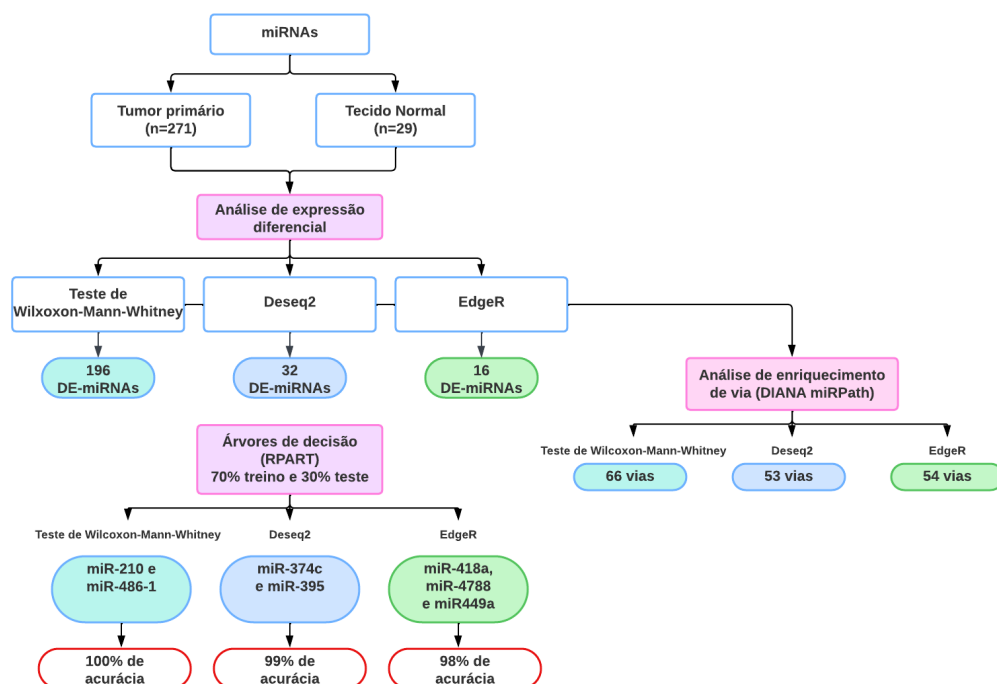


Figura 15 – Fluxograma esquemático da metodologia utilizada e resultados da análise I.

Utilizou-se o banco de dados LUAD miRNA-seq precursores ($n = 300$ e 864 miRNAs) para identificar DE-miRNAs em tecidos T vs. N usando os três diferentes e consolidadas metodologias de análise de expressão diferencial: DESeq2, EdgeR e o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni.

Primeiramente, o teste de Wilcoxon-Mann-Whitney foi realizado e cada miRNA foi avaliado independentemente para determinar a significância estatística da diferença entre as medianas dos dois grupos, considerando $p < 0,05$. Posteriormente, calculou-se $\alpha_{\text{corrigido}}$ da correção de Bonferroni:

$$\alpha_{\text{corrigido}} = \frac{\alpha_{\text{original}}}{n} = \frac{0,05}{864} \approx 0,000058.$$

Assim, identificaram-se 196 DE-miRNAs, sendo 48 *upregulated* e 148 *downregulated*, todos comparados com N e apresentados na Tabela 10. Entre os 10 principais DE-miRNAs (classificados com base nos valores de p mais baixos) (Figura 16A), 8 DE-miRNAs estavam subexpressos e 2 superexpressos.

Para as outras duas metodologias, EdgeR e Deseq2, foram selecionados DE-miRNAs usando como critérios de corte: $\log \text{Fold Change } |\log FC| > 2$ e $\text{adj.}p < 0,05$. O Deseq2 identificou 32 DE-miRNAs (11 superexpressos e 21 subexpressos, Tabela 11). Entre os 10 principais DE-miRNAs (Figura 16B), todos estão subexpressos, O EdgeR identificou 16 DE-miRNAs (11 superexpressos e 5 subexpressos, Tabela 12). Entre os 10 principais DE-miRNAs (Figura 16C), miR-1298, miR-1911, miR-516a-1, miR-516a-2 e miR-449a estavam superexpressos e miR-7641-1, miR-215, miR-7641-2, miR-4788 e miR-347c subexpressos.

O diagrama de Venn, ilustrado na Figura 16D, mostrou que o teste de Wilcoxon-Mann-Whitney obteve 22,68% DE-miRNAs, em relação a todos os miRNAs, enquanto DESeq2 e EdgeR encontraram apenas 3,70% e 1,85% DE-miRNAs, respectivamente. Entre os três métodos investigados há apenas 3 DE-miRNAs em comum miR-516a-1, miR-516a-2 e miR-522 e com a mesma direção de expressão (superexpresso). Além disso, DE-miRNAs comuns foram identificados entre cada par de métodos.

Tabela 10 – DE-miRNAs identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD miRNA-seq precursos.

miRNAs ($n = 254$)	Mediana			miRNAs	Mediana			miRNAs	Mediana		
	N	T	valor p		N	T	valor p		N	T	valor p
mir-210	3345	65	$p < 0,000058$	mir-642a	16	3	$p < 0,000058$	mir-3174	1	0	$p < 0,000058$
mir-153-2	37	3	$p < 0,000058$	mir-199b	9107	4543	$p < 0,000058$	mir-551b	8	3	$p < 0,000058$
mir-708	464	65	$p < 0,000058$	mir-3653	19	5	$p < 0,000058$	mir-1271	8	4	$p < 0,000058$
mir-96	97	15	$p < 0,000058$	mir-592	9	1	$p < 0,000058$	mir-580	2	0	$p < 0,000058$
mir-9-1	797	56	$p < 0,000058$	mir-545	1	0	$p < 0,000058$	mir-219b	1	0	$p < 0,000058$
mir-9-3	805	52	$p < 0,000058$	mir-4638	3	0	$p < 0,000058$	mir-323a	4	1	$p < 0,000058$
mir-9-2	809	53	$p < 0,000058$	mir-550a-1	9	3	$p < 0,000058$	mir-4791	2	0	$p < 0,000058$
mir-183	39603	7138	$p < 0,000058$	mir-5571	1	0	$p < 0,000058$	mir-382	35	19	$p < 0,000058$
mir-21	1503146	378809	$p < 0,000058$	mir-224	136	50	$p < 0,000058$	mir-1291	2	1	$p < 0,000058$
mir-503	31	5	$p < 0,000058$	mir-188	11	5	$p < 0,000058$	mir-4766	1	0	$p < 0,000058$
mir-301a	59	9	$p < 0,000058$	mir-3189	2	0	$p < 0,000058$	mir-651	11	6	$p < 0,000058$
mir-7-1	79	27	$p < 0,000058$	mir-937	13	3	$p < 0,000058$	mir-493	28	16	$p < 0,000058$
mir-141	6358	2045	$p < 0,000058$	mir-627	5	3	$p < 0,000058$	mir-212	26	13	$p < 0,000058$
mir-577	11	0	$p < 0,000058$	mir-3617	1	0	$p < 0,000058$	mir-615	3	0	$p < 0,000058$
mir-4668	7	0	$p < 0,000058$	mir-671	30	11	$p < 0,000058$	mir-4728	8	3	$p < 0,000058$
mir-3607	112	7	$p < 0,000058$	mir-1277	4	1	$p < 0,000058$	mir-106a	43	23	$p < 0,000058$
mir-200a	3690	834	$p < 0,000058$	mir-4661	11	4	$p < 0,000058$	mir-486-1	197	1962	$p < 0,000058$
mir-182	93420	18734	$p < 0,000058$	mir-34a	965	533	$p < 0,000058$	mir-486-2	196	1983	$p < 0,000058$
mir-616	12	1	$p < 0,000058$	mir-2355	216	117	$p < 0,000058$	mir-518e	0	2	$p < 0,000058$
mir-135b	230	36	$p < 0,000058$	mir-1229	2	0	$p < 0,000058$	mir-4732	1	10	$p < 0,000058$
mir-130b	124	26	$p < 0,000058$	mir-940	9	2	$p < 0,000058$	mir-490	0	4	$p < 0,000058$
mir-345	68	14	$p < 0,000058$	mir-4491	2	0	$p < 0,000058$	mir-184	7	85	$p < 0,000058$
mir-301b	6	0	$p < 0,000058$	mir-4652	1	0	$p < 0,000058$	mir-378a	1317	3572	$p < 0,000058$
mir-142	9857	1872	$p < 0,000058$	mir-4724	3	1	$p < 0,000058$	mir-3154	0	1	$p < 0,000058$
mir-147b	5	0	$p < 0,000058$	mir-628	60	35	$p < 0,000058$	mir-1247	43	195	$p < 0,000058$
mir-33b	14	2	$p < 0,000058$	mir-4677	33	20	$p < 0,000058$	mir-516a-1	1	7	$p < 0,000058$
mir-148a	186785	77727	$p < 0,000058$	mir-137	1	0	$p < 0,000058$	mir-139	198	601	$p < 0,000058$
mir-548v	7	1	$p < 0,000058$	mir-639	2	0	$p < 0,000058$	mir-517a	0	1	$p < 0,000058$
mir-33a	75	22	$p < 0,000058$	mir-548s	1	0	$p < 0,000058$	mir-517b	0	1	$p < 0,000058$
mir-3170	3	0	$p < 0,000058$	mir-659	6	2	$p < 0,000058$	mir-133b	3	19	$p < 0,000058$
mir-29b-1	1821	777	$p < 0,000058$	mir-4726	1	0	$p < 0,000058$	mir-4529	0	4	$p < 0,000058$
mir-29b-2	1944	846	$p < 0,000058$	mir-766	41	12	$p < 0,000058$	mir-6892	8	20	$p < 0,000058$
mir-455	620	174	$p < 0,000058$	mir-589	380	215	$p < 0,000058$	mir-133a-1	12	37	$p < 0,000058$
mir-323b	5	0	$p < 0,000058$	mir-196b	70	20	$p < 0,000058$	mir-143	282030	1819712	$p < 0,000058$
mir-331	165	65	$p < 0,000058$	mir-4473	2	0	$p < 0,000058$	mir-516a-2	1	6	$p < 0,000058$
mir-1307	5268	1488	$p < 0,000058$	mir-186	1256	775	$p < 0,000058$	let-7c	4356	11138	$p < 0,000058$
mir-579	3	0	$p < 0,000058$	mir-5698	11	2	$p < 0,000058$	mir-4800	0	1	$p < 0,000058$
mir-625	998	392	$p < 0,000058$	mir-576	38	20	$p < 0,000058$	mir-133a-2	10	27	$p < 0,000058$
mir-539	9	2	$p < 0,000058$	mir-200b	3057	1589	$p < 0,000058$	mir-378c	17	44	$p < 0,000058$
mir-590	68	31	$p < 0,000058$	mir-20a	982	463	$p < 0,000058$	let-7a-3	37892	108488	$p < 0,000058$
mir-31	14	1	$p < 0,000058$	mir-632	1	0	$p < 0,000058$	let-7a-1	37844	108907	$p < 0,000058$
mir-196a-2	21	2	$p < 0,000058$	mir-891a	7	2	$p < 0,000058$	let-7a-2	37806	108301	$p < 0,000058$
mir-556	3	0	$p < 0,000058$	mir-148b	827	489	$p < 0,000058$	mir-6720	3	9	$p < 0,000058$
mir-193b	221	56	$p < 0,000058$	mir-4326	19	4	$p < 0,000058$	mir-30a	72179	320675	$p < 0,000058$
mir-196a-1	20	1	$p < 0,000058$	mir-1287	203	115	$p < 0,000058$	mir-144	338	2831	$p < 0,000058$
mir-450a-1	17	4	$p < 0,000058$	mir-187	53	17	$p < 0,000058$	mir-607	0	3	$p < 0,000058$
mir-425	748	302	$p < 0,000058$	mir-92b	765	326	$p < 0,000058$	mir-522	0	3	$p < 0,000058$
mir-1269a	10	0	$p < 0,000058$	mir-136	77	35	$p < 0,000058$	let-7f-1	17654	76508	$p < 0,000058$
mir-424	297	101	$p < 0,000058$	mir-376b	3	1	$p < 0,000058$	let-7f-2	17931	77092	$p < 0,000058$
mir-1301	57	17	$p < 0,000058$	mir-550a-2	10	4	$p < 0,000058$	mir-584	86	213	$p < 0,000058$
mir-2277	4	0	$p < 0,000058$	mir-154	13	6	$p < 0,000058$	mir-195	164	564	$p < 0,000058$
mir-429	476	198	$p < 0,000058$	mir-3194	1	0	$p < 0,000058$	mir-6788	0	1	$p < 0,000058$
mir-93	16524	6220	$p < 0,000058$	mir-339	196	104	$p < 0,000058$	mir-4524a	0	1	$p < 0,000058$
mir-199a-1	4443	2004	$p < 0,000058$	mir-450b	55	29	$p < 0,000058$	mir-451a	1594	9416	$p < 0,000058$
mir-450a-2	17	5	$p < 0,000058$	mir-3651	2	0	$p < 0,000058$	mir-1-1	17	90	$p < 0,000058$
mir-409	47	14	$p < 0,000058$	mir-3664	1	0	$p < 0,000058$	mir-1293	0	0	$p < 0,000058$
mir-4664	2	0	$p < 0,000058$	mir-548au	1	0	$p < 0,000058$	mir-1-2	18	87	$p < 0,000058$
mir-199a-2	7146	3474	$p < 0,000058$	mir-4645	1	0	$p < 0,000058$	mir-190a	11	34	$p < 0,000058$
mir-3136	2	0	$p < 0,000058$	mir-375	70743	25292	$p < 0,000058$	mir-218-1	107	350	$p < 0,000058$
mir-505	176	71	$p < 0,000058$	mir-7705	4	2	$p < 0,000058$	mir-520a	0	1	$p < 0,000058$
mir-19a	91	42	$p < 0,000058$	mir-214	107	52	$p < 0,000058$	mir-4770	0	1	$p < 0,000058$
mir-454	36	15	$p < 0,000058$	mir-570	1	0	$p < 0,000058$	mir-559	0	0	$p < 0,000058$
mir-324	189	81	$p < 0,000058$	mir-3200	13	5	$p < 0,000058$	mir-218-2	103	328	$p < 0,000058$
mir-3127	22	8	$p < 0,000058$	mir-219a-1	12	5	$p < 0,000058$	mir-4433b	0	0	$p < 0,000058$
mir-191	1976	820	$p < 0,000058$	mir-6737	1	0	$p < 0,000058$				
mir-629	477	217	$p < 0,000058$	mir-320b-2	9	4	$p < 0,000058$				

↑ e ↓ representam a superexpressão ou subexpressão do miRNA, respectivamente.

Fonte: Camargo *et al.* (2023).

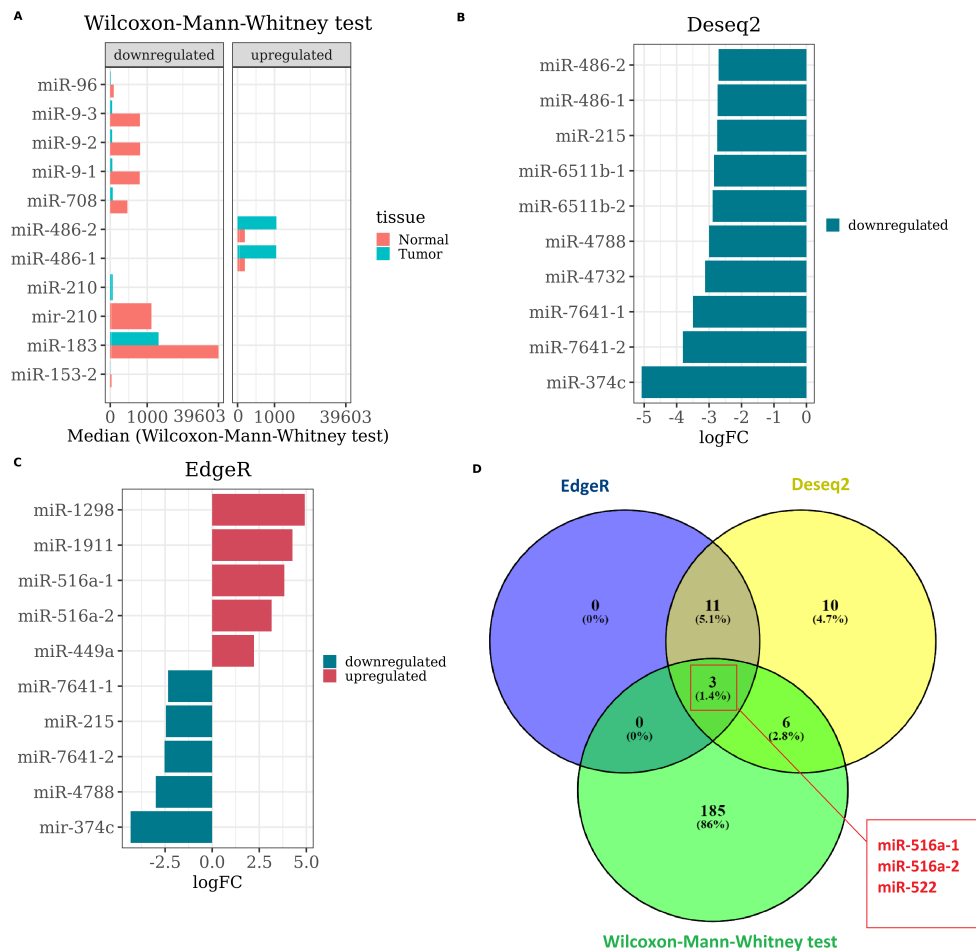


Figura 16 – (A) Gráfico de barras com as medianas da expressão dos 10 principais DE-miRNAs em tecidos T vs. N. (B, C) Gráficos de barra indicando os valores de logFC dos 10 principais DE-miRNAs. (D) Diagrama de Venn mostrando a sobreposição entre os resultados obtidos. Retirado de [Camargo et al. \(2023\)](#).

Tabela 11 – DE-miRNAs identificados com DESeq2 - LUAD miRNA-seq precursores.

miRNAs	logFC	valor <i>p</i>	Valor <i>p</i> ajustado
mir-374c ↓	-5,0769	p<0,0001	p<0,0001
mir-7641-2 ↓	-3,8094	p<0,0001	p<0,0001
mir-7641-1 ↓	-3,4943	p<0,0001	p<0,0001
mir-4732 ↓	-3,1230	p<0,0001	p<0,0001
mir-4788 ↓	-2,9999	p<0,0001	0,00031
mir-6511b-2 ↓	-2,8889	p<0,0001	p<0,0001
mir-6511b-1 ↓	-2,8471	p<0,0001	p<0,0001
mir-215 ↓	-2,7484	p<0,0001	p<0,0001
mir-486-1 ↓	-2,7405	p<0,0001	p<0,0001
mir-486-2 ↓	-2,7080	p<0,0001	p<0,0001
mir-328 ↓	-2,5748	p<0,0001	p<0,0001
mir-3609 ↓	-2,5349	p<0,0001	p<0,0001
mir-3154 ↓	-2,4961	p<0,0001	p<0,0001
mir-935 ↓	-2,4556	p<0,0001	p<0,0001
mir-197 ↓	-2,3717	p<0,0001	p<0,0001
mir-6892 ↓	-2,2234	p<0,0001	p<0,0001
mir-6788 ↓	-2,2207	p<0,0001	p<0,0001
mir-3622a ↓	-2,2013	p<0,0001	p<0,0001
mir-6894 ↓	-2,0915	p<0,0001	p<0,0001
mir-605 ↓	-2,0856	p<0,0001	p<0,0001
mir-504 ↓	-2,0556	p<0,0001	p<0,0001
mir-449a ↑	2,0922	p<0,0001	0,0010
mir-449b ↑	2,5142	0,0010	0,0073
mir-516a-2 ↑	2,6437	0,0006	0,0052
mir-519a-2 ↑	2,8288	0,0054	0,0307
mir-372 ↑	3,3660	0,0002	0,0022
mir-516a-1 ↑	3,5435	p<0,0001	0,0005
mir-522 ↑	3,5455	0,0009	0,0071
mir-1911 ↑	4,0720	p<0,0001	0,0001
mir-1298 ↑	5,0724	p<0,0001	0,0002
mir-1912 ↑	5,3226	p<0,0001	0,0010
mir-122 ↑	7,2576	p<0,0001	0,0001

↑ e ↓ representam a superexpressão ou subexpressão do miRNA, respectivamente.

Fonte: Camargo *et al.* (2023).

Tabela 12 – DE-miRNAs identificados com EdgeR - LUAD miRNA-seq precursores.

miRNAs	logFC	valor <i>p</i>	Valor <i>p</i> ajustado
mir-374c ↓	-4,3344	p<0,0001	p<0,0001
mir-215 ↓	-2,4613	p<0,0001	p<0,0001
mir-4788 ↓	-2,9893	p<0,0001	p<0,0001
mir-7641-2 ↓	-2,5342	p<0,0001	p<0,0001
mir-7641-1 ↓	-2,3452	p<0,0001	p<0,0001
mir-449a ↑	2,2246	0,0004	0,0059
mir-1911 ↑	4,2613	0,0004	0,0059
mir-516a-1 ↑	3,8279	0,0005	0,0067
mir-516a-2 ↑	3,1572	0,0009	0,0096
mir-1298 ↑	4,9219	0,0014	0,0142
mir-449b ↑	2,6260	0,0025	0,0205
mir-372 ↑	3,6213	0,0025	0,0206
mir-522 ↑	4,1671	0,0034	0,0254
mir-122 ↑	6,8609	0,0037	0,0267
mir-1912 ↑	4,7074	0,0060	0,0376
mir-519a-2 ↑	3,4594	0,0077	0,0467

↑ e ↓ representam a superexpressão ou subexpressão do miRNA, respectivamente.

Fonte: Camargo *et al.* (2023).

Para identificar e compreender o papel biológico, dos DE-miRNAs identificados, em cada método em relação aos genes-alvo e vias no câncer, as sequências precursoras de miRNA foram convertidas em sequências maduras, Tabela 13, por meio da ferramenta miEAA2.0.

Tabela 13 – MiRNAs precursores e maduros por metodologia.

Precursor	Maduro	Precursor	Maduro
Teste de Wilcoxon-Mann-Whitney			
mir-210	miR-210-5p;miR-210-3p	mir-627	miR-627-5p;miR-627-3p
mir-153-2	miR-153-5p;miR-153-3p	mir-3617	miR-3617-5p;miR-3617-3p
mir-708	miR-708-5p;miR-708-3p	mir-671	miR-671-5p;miR-671-3p
mir-96	miR-96-5p;miR-96-3p	mir-133a-2	miR-133a-5p;miR-133a-3p
mir-9-1	miR-9-5p;miR-9-3p	mir-1277	miR-1277-5p;miR-1277-3p
mir-9-3	miR-9-5p;miR-9-3p	mir-4661	miR-4661-5p;miR-4661-3p
mir-9-2	miR-9-5p;miR-9-3p	mir-378c	miR-378c
mir-183	miR-183-5p;miR-183-3p	mir-34a	miR-34a-5p;miR-34a-3p
mir-486-1	miR-486-5p;miR-486-3p	let-7a-3	let-7a-5p;let-7a-3p
mir-486-2	miR-486-5p;miR-486-3p	let-7a-1	let-7a-5p;let-7a-3p
mir-21	miR-21-5p;miR-21-3p	let-7a-2	let-7a-5p;let-7a-2-3p
mir-503	miR-503-5p;miR-503-3p	mir-6720	miR-6720-5p;miR-6720-3p
mir-301a	miR-301a-5p;miR-301a-3p	mir-2355	miR-2355-5p;miR-2355-3p
mir-7-1	miR-7-5p;miR-7-1-3p	mir-1229	miR-1229-5p;miR-1229-3p
mir-141	miR-141-5p;miR-141-3p	mir-940	miR-940
mir-577	miR-577	mir-4491	miR-4491
mir-4668	miR-4668-5p;miR-4668-3p	mir-4652	miR-4652-5p;miR-4652-3p
mir-200a	miR-200a-5p;miR-200a-3p	mir-4724	miR-4724-5p;miR-4724-3p
mir-182	miR-182-5p;miR-182-3p	mir-628	miR-628-5p;miR-628-3p
mir-616	miR-616-5p;miR-616-3p	mir-4677	miR-4677-5p;miR-4677-3p
mir-135b	miR-135b-5p;miR-135b-3p	mir-137	miR-137-5p;miR-137-3p
mir-130b	miR-130b-5p;miR-130b-3p	mir-30a	miR-30a-5p;miR-30a-3p
mir-345	miR-345-5p;miR-345-3p	mir-639	miR-639
mir-518e	miR-518e-5p;miR-518e-3p	mir-548s	miR-548s
mir-301b	miR-301b-5p;miR-301b-3p	mir-144	miR-144-5p;miR-144-3p
mir-142	miR-142-5p;miR-142-3p	mir-659	miR-659-5p;miR-659-3p
mir-4732	miR-4732-5p;miR-4732-3p	mir-607	miR-607
mir-147b	miR-147b-5p;miR-147b-3p	mir-4726	miR-4726-5p;miR-4726-3p
mir-33b	miR-33b-5p;miR-33b-3p	mir-766	miR-766-5p;miR-766-3p
mir-148a	miR-148a-5p;miR-148a-3p	mir-589	miR-589-5p;miR-589-3p
mir-548v	miR-548v	mir-196b	miR-196b-5p;miR-196b-3p
mir-33a	miR-33a-5p;miR-33a-3p	mir-4473	miR-4473
mir-3170	miR-3170	mir-186	miR-186-5p;miR-186-3p
mir-29b-1	miR-29b-3p;miR-29b-1-5p	mir-5698	miR-5698
mir-29b-2	miR-29b-3p;miR-29b-2-5p	mir-522	miR-522-5p;miR-522-3p
mir-455	miR-455-5p;miR-455-3p	mir-576	miR-576-5p;miR-576-3p
mir-323b	miR-323b-5p;miR-323b-3p	mir-200b	miR-200b-5p;miR-200b-3p
mir-331	miR-331-5p;miR-331-3p	mir-20a	miR-20a-5p;miR-20a-3p
mir-490	miR-490-5p;miR-490-3p	mir-632	miR-632
mir-1307	miR-1307-5p;miR-1307-3p	mir-891a	miR-891a-5p;miR-891a-3p
mir-184	miR-184	let-7f-1	let-7f-5p;let-7f-1-3p
mir-579	miR-579-5p;miR-579-3p	mir-148b	miR-148b-5p;miR-148b-3p
mir-378a	miR-378a-5p;miR-378a-3p	let-7f-2	let-7f-5p;let-7f-2-3p
mir-625	miR-625-5p;miR-625-3p	mir-4326	miR-4326
mir-539	miR-539-5p;miR-539-3p	mir-584	miR-584-5p;miR-584-3p
mir-3154	miR-3154	mir-195	miR-195-5p;miR-195-3p
mir-590	miR-590-5p;miR-590-3p	mir-1287	miR-1287-5p;miR-1287-3p
mir-31	miR-31-5p;miR-31-3p	mir-187	miR-187-5p;miR-187-3p
mir-1247	miR-1247-5p;miR-1247-3p	mir-6788	miR-6788-5p;miR-6788-3p
mir-196a-2	miR-196a-5p;miR-196a-3p	mir-92b	miR-92b-5p;miR-92b-3p
mir-516a-1	miR-516a-5p;miR-516a-3p	mir-136	miR-136-5p;miR-136-3p
mir-556	miR-556-5p;miR-556-3p	mir-376b	miR-376b-5p;miR-376b-3p
mir-193b	miR-193b-5p;miR-193b-3p	mir-550a-2	miR-550a-5p;miR-550a-3p
mir-139	miR-139-5p;miR-139-3p	mir-154	miR-154-5p;miR-154-3p
mir-196a-1	miR-196a-5p;miR-196a-1-3p	mir-3194	miR-3194-5p;miR-3194-3p
mir-450a-1	miR-450a-5p;miR-450a-1-3p	mir-339	miR-339-5p;miR-339-3p
mir-425	miR-425-5p;miR-425-3p	mir-450b	miR-450b-5p;miR-450b-3p
mir-517a	miR-517a-3p;miR-517-5p	mir-3651	miR-3651
mir-1269a	miR-1269a	mir-3664	miR-3664-5p;miR-3664-3p
mir-517b	miR-517b-3p;miR-517-5p	mir-548au	miR-548au-5p;miR-548au-3p
mir-424	miR-424-5p;miR-424-3p	mir-4645	miR-4645-5p;miR-4645-3p
mir-1301	miR-1301-5p;miR-1301-3p	mir-4524a	miR-4524a-5p;miR-4524a-3p
mir-133b	miR-133b	mir-375	miR-375-5p;miR-375-3p
mir-2277	miR-2277-5p;miR-2277-3p	mir-7705	miR-7705
mir-429	miR-429	mir-451a	miR-451a
mir-4529	miR-4529-5p;miR-4529-3p	mir-214	miR-214-5p;miR-214-3p
mir-93	miR-93-5p;miR-93-3p	mir-1-1	miR-1-5p;miR-1-3p

continuação			
Precursor	Maduro	Precursor	Maduro
mir-199a-1	miR-199a-5p;miR-199a-3p	mir-570	miR-570-5p;miR-570-3p
mir-450a-2	miR-450a-5p;miR-450a-2-3p	mir-1293	miR-1293
mir-409	miR-409-5p;miR-409-3p	mir-3200	miR-3200-5p;miR-3200-3p
mir-4664	miR-4664-5p;miR-4664-3p	mir-1-2	miR-1-3p
mir-199a-2	miR-199a-5p;miR-199a-3p	mir-190a	miR-190a-5p;miR-190a-3p
mir-3136	miR-3136-5p;miR-3136-3p	mir-219a-1	miR-219a-5p;miR-219a-1-3p
mir-505	miR-505-5p;miR-505-3p	mir-6737	miR-6737-5p;miR-6737-3p
mir-6892	miR-6892-5p;miR-6892-3p	mir-320b-2	miR-320b
mir-19a	miR-19a-5p;miR-19a-3p	mir-3174	miR-3174
mir-454	miR-454-5p;miR-454-3p	mir-551b	miR-551b-5p;miR-551b-3p
mir-133a-1	miR-133a-5p;miR-133a-3p	mir-218-1	miR-218-5p;miR-218-1-3p
mir-324	miR-324-5p;miR-324-3p	mir-1271	miR-1271-5p;miR-1271-3p
mir-3127	miR-3127-5p;miR-3127-3p	mir-580	miR-580-5p;miR-580-3p
mir-143	miR-143-5p;miR-143-3p	mir-219b	miR-219b-5p;miR-219b-3p
mir-191	miR-191-5p;miR-191-3p	mir-323a	miR-323a-5p;miR-323a-3p
mir-629	miR-629-5p;miR-629-3p	mir-4791	miR-4791
mir-516a-2	miR-516a-5p;miR-516a-3p	mir-382	miR-382-5p;miR-382-3p
mir-642a	miR-642a-5p;miR-642a-3p	mir-1291	miR-1291
mir-199b	miR-199b-5p;miR-199b-3p	mir-520a	miR-520a-5p;miR-520a-3p
mir-592	miR-592	mir-4766	miR-4766-5p;miR-4766-3p
mir-545	miR-545-5p;miR-545-3p	mir-4770	miR-4770
mir-4638	miR-4638-5p;miR-4638-3p	mir-651	miR-651-5p;miR-651-3p
mir-550a-1	miR-550a-5p;miR-550a-3p	mir-559	miR-559
mir-5571	miR-5571-5p;miR-5571-3p	mir-493	miR-493-5p;miR-493-3p
mir-224	miR-224-5p;miR-224-3p	mir-218-2	miR-218-5p;miR-218-2-3p
let-7c	let-7c-5p;let-7c-3p	mir-4433b	miR-4433b-5p;miR-4433b-3p
mir-188	miR-188-5p;miR-188-3p	mir-212	miR-212-5p;miR-212-3p
mir-3189	miR-3189-5p;miR-3189-3p	mir-615	miR-615-5p;miR-615-3p
mir-937	miR-937-5p;miR-937-3p	mir-4728	miR-4728-5p;miR-4728-3p
mir-4800	miR-4800-5p;miR-4800-3p	mir-106a	miR-106a-5p;miR-106a-3p
DESeq2			
mir-122	miR-122-5p;miR-122-3p	mir-6892	miR-6892-5p;miR-6892-3p
mir-1911	miR-1911-5p;miR-1911-3p	mir-486-1	miR-486-5p;miR-486-3p
mir-1298	miR-1298-5p;miR-1298-3p	mir-486-2	miR-486-5p;miR-486-3p
mir-516a-1	miR-516a-5p;miR-516a-3p	mir-374c	miR-374c-5p;miR-374c-3p
mir-1912	miR-1912-5p;miR-1912-3p	mir-4732	miR-4732-5p;miR-4732-3p
mir-449a	miR-449a	mir-605	miR-605-5p;miR-605-3p
mir-372	miR-372-5p;miR-372-3p	mir-504	miR-504-5p;miR-504-3p
mir-516a-2	miR-516a-5p;miR-516a-3p	mir-3609	miR-3609
mir-522	miR-522-5p;miR-522-3p	mir-935	miR-935
mir-449b	miR-449b-5p;miR-449b-3p	mir-3154	miR-3154
mir-519a-2	miR-519a-3p;miR-519a-2-5p	mir-215	miR-215-5p;miR-215-3p
mir-328	miR-328-5p;miR-328-3p	mir-6788	miR-6788-5p;miR-6788-3p
mir-6511b-1	miR-6511b-5p;miR-6511b-3p	mir-6894	miR-6894-5p;miR-6894-3p
mir-6511b-2	miR-6511b-5p;miR-6511b-3p	mir-3622a	miR-3622a-5p;miR-3622a-3p
mir-197	miR-197-5p;miR-197-3p	mir-4788	miR-4788
EgdeR			
mir-1911	miR-1911-5p;miR-1911-3p	mir-522	miR-522-5p;miR-522-3p
mir-449a	miR-449a	mir-122	miR-122-5p;miR-122-3p
mir-516a-1	miR-516a-5p;miR-516a-3p	mir-1912	miR-1912-5p;miR-1912-3p
mir-516a-2	miR-516a-5p;miR-516a-3p	mir-519a-2	miR-519a-3p;miR-519a-2-5p
mir-1298	miR-1298-5p;miR-1298-3p	mir-374c	miR-374c-5p;miR-374c-3p
mir-449b	miR-449b-5p;miR-449b-3p	mir-215	miR-215-5p;miR-215-3p
mir-372	miR-372-5p;miR-372-3p	mir-4788	miR-4788

Fonte: Camargo *et al.* (2023).

Em seguida, a análise de enriquecimento de vias foi realizada utilizando a ferramenta DIANA-miRPath e dados da KEGG. Foram selecionadas as vias KEGG mais enriquecidas, considerando $p < 0,05$ ajustado por Benjamini-Hochberg, Tabela 14.

Tabela 14 – Vias enriquecidas para cada método.

KEGG pathway	valor <i>p</i>	genes	miRNAs
Teste de Wilcoxon-Mann-Whitney			
ECM-receptor interaction	p<0,0001	59	53
Proteoglycans in cancer	p<0,0001	150	73
Mucin type O-Glycan biosynthesis	p<0,0001	22	33
Hippo signaling pathway	p<0,0001	115	66
Axon guidance	p<0,0001	97	63
Morphine addiction	p<0,0001	71	65
GABAergic synapse	p<0,0001	68	66
Pathways in cancer	p<0,0001	284	80
ErbB signaling pathway	p<0,0001	72	64
Signaling pathways regulating pluripotency of stem cells	p<0,0001	106	74
Focal adhesion	p<0,0001	155	73
Bacterial invasion of epithelial cells	p<0,0001	63	59
Adherens junction	p<0,0001	57	62
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	p<0,0001	54	55
Pancreatic cancer	p<0,0001	53	62
Thyroid hormone signaling pathway	p<0,0001	91	70
Fatty acid biosynthesis	p<0,0001	8	16
Glioma	p<0,0001	50	61
Gap junction	p<0,0001	67	64
TGF-beta signaling pathway	0,0002	60	62
Adrenergic signaling in cardiomyocytes	0,0007	103	71
Colorectal cancer	0,0008	49	62
PI3K-Akt signaling pathway	0,0014	231	78
Ras signaling pathway	0,0014	154	72
Endocytosis	0,0015	146	70
Glutamatergic synapse	0,0015	81	69
Lysine degradation	0,0019	35	51
Estrogen signaling pathway	0,0019	71	72
cGMP-PKG signaling pathway	0,0027	117	76
Renal cell carcinoma	0,0030	52	57
Phosphatidylinositol signaling system	0,0031	59	58
Neurotrophin signaling pathway	0,0031	88	72
Rap1 signaling pathway	0,0032	144	70
Amoebiasis	0,0033	74	59
Prolactin signaling pathway	0,0033	51	63
Prostate cancer	0,0033	66	67
Regulation of actin cytoskeleton	0,0033	149	72
Choline metabolism in cancer	0,0038	76	70
Retrograde endocannabinoid signaling	0,0045	73	70
Oxytocin signaling pathway	0,0053	112	75
Sphingolipid signaling pathway	0,0057	83	66
Prion diseases	0,0085	18	32
Non-small cell lung cancer	0,0085	41	59
Glycosphingolipid biosynthesis - ganglio series	0,0089	10	22
Small cell lung cancer	0,0099	63	58
Chronic myeloid leukemia	0,0101	55	62
Dorso-ventral axis formation	0,0135	23	42
Regulation of autophagy	0,0150	31	38
Wnt signaling pathway	0,0150	96	70
MAPK signaling pathway	0,0161	171	77
Melanoma	0,0248	52	60
Platelet activation	0,0248	89	69
Hepatitis B	0,0248	95	76
Cholinergic synapse	0,0262	79	74
mTOR signaling pathway	0,0268	45	61
Endometrial cancer	0,0272	39	57
Shigellosis	0,0297	45	53
Cocaine addiction	0,0297	33	55
Long-term depression	0,0305	42	52
cAMP signaling pathway	0,0327	133	74
Melanogenesis	0,0367	68	66
Acute myeloid leukemia	0,0392	42	59
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	0,0418	18	34
p53 signaling pathway	0,0418	50	59
mRNA surveillance pathway	0,0483	62	64
AMPK signaling pathway	0,0483	87	68
Deseq2			
Morphine addiction	p<0,0001	55	35
Proteoglycans in cancer	p<0,0001	113	38
Axon guidance	p<0,0001	74	32
ErbB signaling pathway	p<0,0001	59	34
Glioma	p<0,0001	42	32

KEGG pathway	valor <i>p</i>	continuação	
		genes	miRNAs
Pathways in cancer	p<0,0001	215	43
Thyroid hormone signaling pathway	p<0,0001	70	34
GABAergic synapse	p<0,0001	52	34
Pancreatic cancer	p<0,0001	45	33
Retrograde endocannabinoid signaling	0,0002	62	34
MAPK signaling pathway	0,0002	140	42
Chronic myeloid leukemia	0,0002	46	32
Long-term depression	0,0003	38	27
Melanoma	0,0005	46	29
Renal cell carcinoma	0,0006	44	27
Lysine degradation	0,0007	28	29
Wnt signaling pathway	0,0007	76	36
Gap junction	0,0009	50	32
Glutamatergic synapse	0,0022	65	37
TGF-beta signaling pathway	0,0025	42	28
Choline metabolism in cancer	0,0025	61	35
Nicotine addiction	0,0033	24	26
Prostate cancer	0,0033	52	32
Hepatitis B	0,0033	74	36
FoxO signaling pathway	0,0033	71	32
Prion diseases	0,0039	8	14
Signaling pathways regulating pluripotency of stem cells	0,0043	76	37
Dopaminergic synapse	0,0043	73	37
Phosphatidylinositol signaling system	0,0047	42	33
Amphetamine addiction	0,0047	37	27
Hippo signaling pathway	0,0064	75	33
Adrenergic signaling in cardiomyocytes	0,0064	76	37
Ras signaling pathway	0,0064	112	41
Regulation of actin cytoskeleton	0,0075	114	39
Non-small cell lung cancer	0,0097	36	27
p53 signaling pathway	0,0097	41	28
Acute myeloid leukemia	0,0137	34	32
Estrogen signaling pathway	0,0178	50	38
cGMP-PKG signaling pathway	0,0179	87	38
Prolactin signaling pathway	0,0207	42	33
N-Glycan biosynthesis	0,0237	24	24
PI3K-Akt signaling pathway	0,0238	166	40
Colorectal cancer	0,0247	38	31
Focal adhesion	0,0278	105	39
Mucin type O-Glycan biosynthesis	0,0300	17	15
Bladder cancer	0,0338	24	22
cAMP signaling pathway	0,0427	101	38
Circadian entrainment	0,0434	52	35
Rap1 signaling pathway	0,0436	105	41
Neurotrophin signaling pathway	0,0442	65	35
Synaptic vesicle cycle	0,0466	31	23
SNARE interactions in vesicular transport	0,0473	20	23
Transcriptional misregulation in cancer	0,0494	84	38
EdgeR			
Proteoglycans in cancer	p<0,0001	87	17
Prion diseases	p<0,0001	8	9
ErbB signaling pathway	p<0,0001	46	16
Glioma	p<0,0001	33	16
Pathways in cancer	p<0,0001	157	18
MAPK signaling pathway	p<0,0001	106	19
Amphetamine addiction	0,0002	31	15
N-Glycan biosynthesis	0,0002	21	16
Hepatitis B	0,0003	60	16
cGMP-PKG signaling pathway	0,0003	72	20
Thyroid hormone signaling pathway	0,0004	51	15
Estrogen signaling pathway	0,0004	40	19
Melanoma	0,0006	37	13
Retrograde endocannabinoid signaling	0,0007	46	17
Morphine addiction	0,0007	38	17
Glutamatergic synapse	0,0008	49	16
TGF-beta signaling pathway	0,0008	35	13
Renal cell carcinoma	0,0008	33	13
Prostate cancer	0,0008	41	14
Dopaminergic synapse	0,0008	59	18
Axon guidance	0,0017	52	14
Wnt signaling pathway	0,0020	55	16
Protein processing in endoplasmic reticulum	0,0024	67	16
Pancreatic cancer	0,0025	33	15

KEGG pathway	valor <i>p</i>	genes	continuação
			miRNAs
Chronic myeloid leukemia	0,0029	34	15
Long-term depression	0,0036	28	13
Signaling pathways regulating pluripotency of stem cells	0,0036	55	15
p53 signaling pathway	0,0086	31	15
Phosphatidylinositol signaling system	0,0089	34	17
Biosynthesis of unsaturated fatty acids	0,0108	8	5
Choline metabolism in cancer	0,0108	45	14
Circadian entrainment	0,0108	43	18
GABAergic synapse	0,0109	33	16
Mucin type O-Glycan biosynthesis	0,0117	11	9
Serotonergic synapse	0,0117	44	15
Hippo signaling pathway	0,0119	59	15
Rap1 signaling pathway	0,0130	80	18
Adrenergic signaling in cardiomyocytes	0,0135	55	18
Ras signaling pathway	0,0162	83	18
Lysine degradation	0,0185	16	13
Long-term potentiation	0,0185	30	16
PI3K-Akt signaling pathway	0,0185	119	17
Bladder cancer	0,0186	20	12
FoxO signaling pathway	0,0197	53	15
HIF-1 signaling pathway	0,0216	43	15
SNARE interactions in vesicular transport	0,0245	15	13
Transcriptional misregulation in cancer	0,0272	63	16
Prolactin signaling pathway	0,0289	30	16
cAMP signaling pathway	0,0296	74	18
Gap junction	0,0334	32	15
Vascular smooth muscle contraction	0,0334	44	17
Oxytocin signaling pathway	0,0421	60	19
Acute myeloid leukemia	0,0463	24	14
Neurotrophin signaling pathway	0,0463	47	17

Fonte: Camargo *et al.* (2023).

A Figura 17A apresenta os 10 principais caminhos KEGG com os valores *p* mais baixos para cada método de análise de expressão diferencial. Os três métodos identificaram vias enriquecidas semelhantes, incluindo proteoglicanos no câncer, vias no câncer, via de sinalização ErbB, biossíntese de O-Glicano tipo mucina, via de sinalização PI3K-Akt e via de sinalização Ras. A Figura 17B apresenta o diagrama de Venn e destaca a sobreposição de 47,6% entre as vias enriquecidas identificadas derivadas dos três métodos.

Há diferentes vias identificadas por cada método. Por exemplo, as vias da dependência da nicotina e do ciclo da vesícula sináptica foram identificadas usando apenas os DE-miRNAs derivados do Deseq2. As vias de processamento de proteínas no retículo endoplasmático, biossíntese de ácidos graxos insaturados, sinapse serotoninérgica, potencialização de longo prazo e via de sinalização do HIF-1 foram encontradas usando apenas os resultados gerados pelo EdgeR. O teste de Wilcoxon-Mann-Whitney identificou um total de 66 vias enriquecidas, número superior de vias enriquecidas identificadas quando comparado ao Deseq2 (53 vias enriquecidas) e EdgeR (54 vias enriquecidas). A lista de vias únicas por cada método pode ser encontrada na Tabela 15.

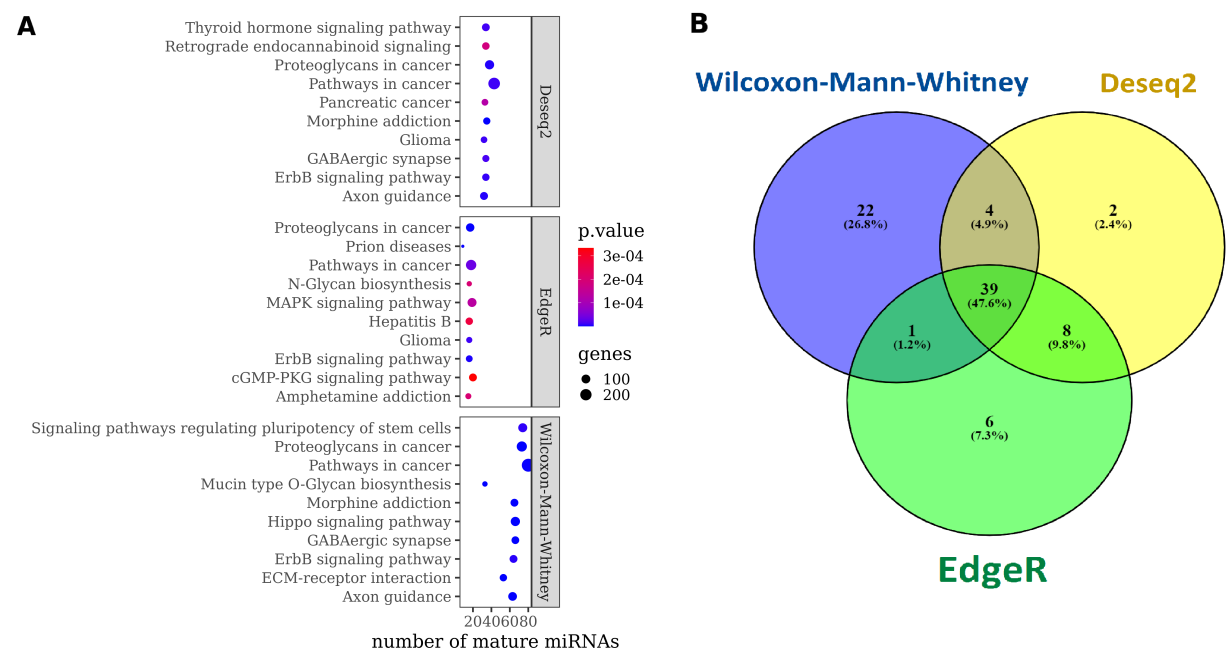


Figura 17 – (A) As 10 vias com os menores valores de p identificados pela análise KEGG dos resultados das três metodologias. (B) Diagrama de Venn destacando a existência de sobreposição. Retirado de Camargo *et al.* (2023).

Tabela 15 – Lista de caminhos únicos por cada método.

Teste de Wilcoxon-Mann-Whitney
ECM-receptor interaction
Bacterial invasion of epithelial cells
Adherens junction
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Fatty acid biosynthesis
Endocytosis
Amoebiasis
Sphingolipid signaling pathway
Glycosphingolipid biosynthesis - ganglio series
Small cell lung cancer
Dorso-ventral axis formation
Regulation of autophagy
Platelet activation
Cholinergic synapse
mTOR signaling pathway
Endometrial cancer
Shigellosis
Cocaine addiction
Melanogenesis
Glycosaminoglycan biosynthesis - heparan sulfate / heparin
mRNA surveillance pathway
AMPK signaling pathway
Deseq2
Nicotine addiction
Synaptic vesicle cycle
EdgeR
Protein processing in endoplasmic reticulum
Biosynthesis of unsaturated fatty acids
Serotonergic synapse
Long-term potentiation
HIF-1 signaling pathway
Vascular smooth muscle contraction

Fonte: Camargo *et al.* (2023).

Em seguida, os bancos de dados formados pelos DE-miRNAs precursores identificados por cada método e o tipo de tecido que a amostra pertencia (T ou N) foram divididos aleatoriamente em 70% para treinamento e 30% para teste. Para cada conjunto de DE-miRNAs, foram geradas 10 mil árvores de classificação, geradas utilizando o hiperparâmetro $cp = 0,001$, a fim de diferenciar os tecidos em T vs. N e, assim, comparar as acurácias dos resultados obtidos pelos três diferentes métodos.

Analisando as 10 mil acurácias geradas pelas árvores de decisão dos DE-miRNAs identificados para cada método, Figura 18, verificou-se que o teste de Wilcoxon-Mann-Whitney apresenta para esses dados os melhores níveis de acurácia (83-100%, média de aproximadamente 96,5%) (Figura 18A) e observa-se que a curva do histograma contém mais de 75% das acurácias acima de 95%. Enquanto os métodos DESeq2 e EdgeR, as acurácias foram de 71-99% (média de aproximadamente 89,5%) e 74-99% (média de aproximadamente 91,5%), respectivamente, e ambos não continham muitos valores acima de 95% de acurácia. Para todos os métodos, as médias estão acima de 89%, indicando uma alta taxa de acerto para a classificação do tecido.

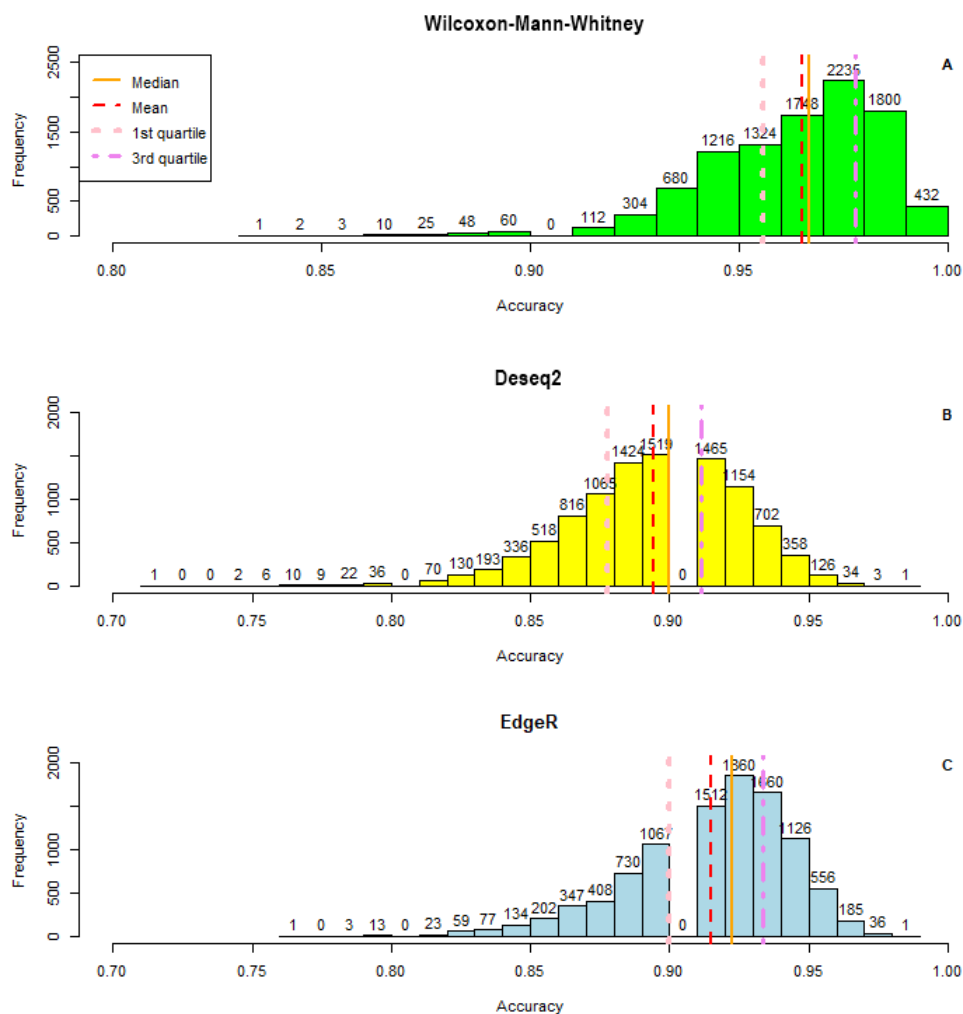


Figura 18 – Histogramas com as acurácias das simulações de árvore de classificação para o teste Wilcoxon-Mann-Whitney, Deseq2 e EdgeR. Retirado de [Camargo et al. \(2023\)](#).

As árvores de classificação com as acurácias mais elevadas, indicando maior precisão na distinção entre T vs. N, foram representadas e estão visualizadas na Figura 19. Elas revelam miRNAs distintos para a separação dos tecidos, e é importante destacar que nenhum miRNA se repetiu entre as diferentes árvores.

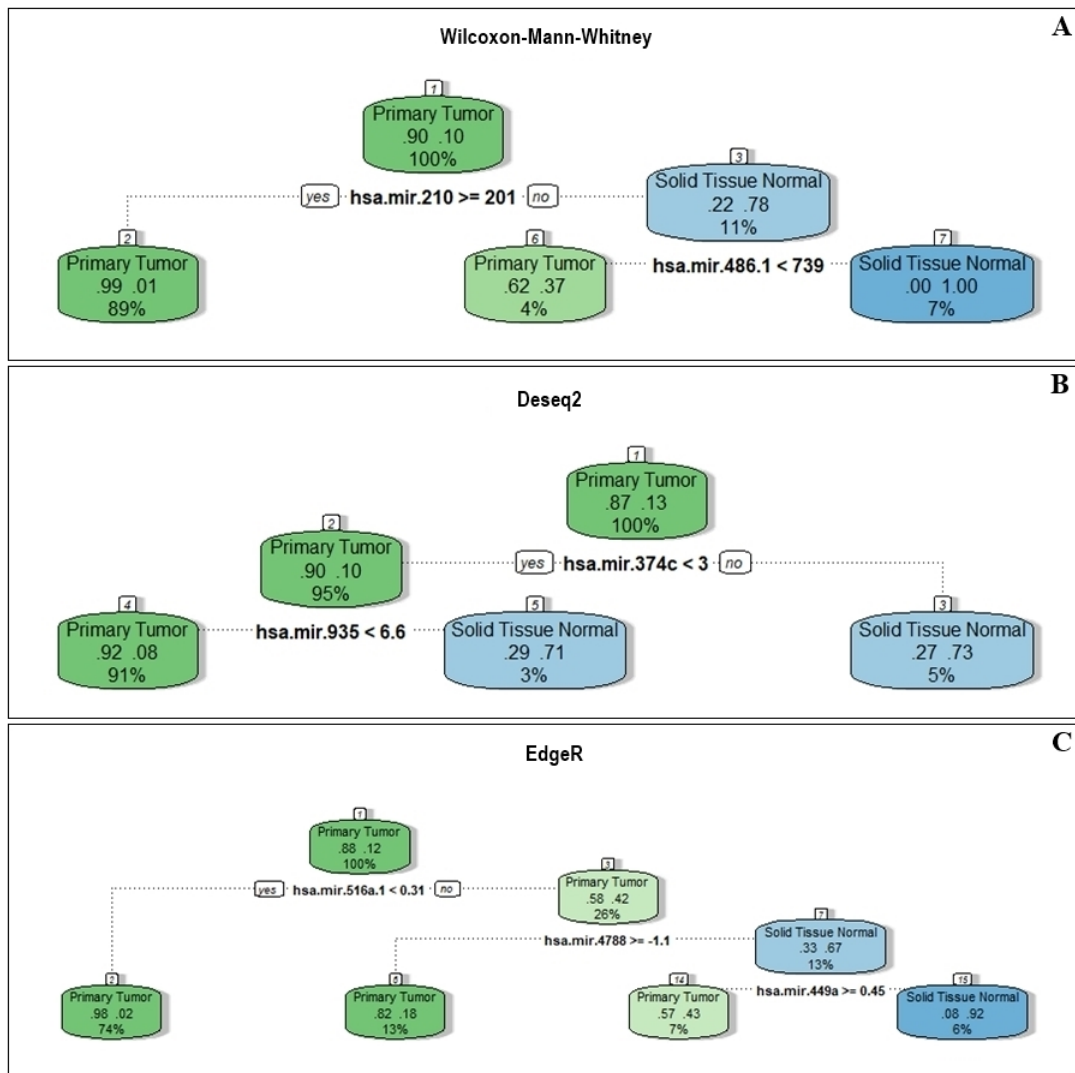


Figura 19 – (A) Árvore de decisão com 100% de acurácia para o teste de Wilcoxon-Mann-Whitney. (B) Árvore de decisão com 99% de acurácia para Deseq2. (C) Árvore de decisão com 98% de acurácia para EdgeR. Retirado de [Camargo et al. \(2023\)](#).

A árvore com os DE-miRNAs identificados no teste de Wilcoxon-Mann-Whitney (100% de acurácia, Figura 19A) é formado por apenas dois miRNAs, o mir-210 representa o nó raiz com a classe prevista de tumor primário, pacientes com $\text{mir-210} \geq 201$ seguem para a esquerda e os demais para a direita. No nó filho $\text{mir-486-1} < 739$, pacientes com expressões que satisfazem essa condição seguem para a esquerda sendo classificados como tumor primário e os demais para a direita sendo classificados como tecido normal. Mesmo o método tendo identificado 196 DE-miRNAs, a árvore tornou-se simples, e tanto ela quanto a de Deseq2 (Figura 19B) não houve necessidade de realizar podas. No entanto, a árvore EdgeR (Figura 19C) tornou-se complexa

e houve necessidade de poda em relação ao custo-complexidade da árvore ($cp = 0,0714$), provavelmente devido à quantidade de DE-miRNAs que a metodologia identificou.

A identificação de DE-miRNAs é uma etapa crucial em estudos que buscam identificar genes como biomarcadores, e é essencial analisar cuidadosamente qual método de análise de dados é mais adequado para cada questão de pesquisa.

Neste estudo, ao comparar amostras de adenocarcinoma pulmonar em estágio inicial com tecidos pulmonares normais adjacentes ao tumor, observou-se que o teste de Wilcoxon-Mann-Whitney, Deseq2 e EdgeR divergem no conjunto de DE-miRNAs identificados, com apenas 1,4% de DE-miRNAs sendo comumente identificados por todos os três métodos. Assim, a escolha do método de análise de dados pode levar a diferentes interpretações biológicas.

O teste de Wilcoxon-Mann-Whitney identificou 196 DE-miRNAs (22% dos 864 miRNAs expressos nos tecidos pulmonares), mostrando-se conservador durante o teste de hipótese em comparação com outros testes paramétricos. Além disso, este método identificou 22 vias enriquecidas que não foram identificadas pelos outros métodos (Figura 17B). Utilizando um classificador baseado em árvore, alcançou melhores acurácias (83-100%) para classificar tecidos T vs. N. Este resultado está de acordo com os dados apresentados por [Li et al. \(2022\)](#), indicando que o teste de Wilcoxon superou os outros métodos em termos de potência. Esses autores mostraram que os métodos DESeq2 e EdgeR têm taxas de FDR sugeriram que estudos de RNA-seq ao nível populacional com grandes tamanhos de amostra deveriam usar o teste de soma de classificação de Wilcoxon para análise de dados.

Deseq2 e EdgeR foram considerados mais sucintos quando comparados ao teste de Wilcoxon-Mann-Whitney, pois identificaram apenas 3,7% e 1,85% de DE-miRNAs dos 864 miRNAs expressos, respectivamente. Todos os DE-miRNAs identificados pelo EdgeR também foram encontrados pelo DESeq2, e, em ambos os conjuntos de dados, os DE-miRNAs foram regulados negativamente. Mesmo o EdgeR detectando menos DE-miRNAs, apresentou boas acurácias nas simulações de árvore de decisão (74-99%), ligeiramente superiores às de Deseq2 (71-99%).

Enquanto os três métodos identificam vias enriquecidas semelhantes associadas ao câncer de pulmão, o teste de Wilcoxon-Mann-Whitney identificou vias mais enriquecidas ($P < 0,05$) e expandiu as possibilidades de compreensão biológica dos dados.

Além disso, ao usar árvores de decisão para classificar T vs. N e ter um classificador com saídas (pontos de corte) com dados de expressão de miRNA-seq originais (contagens brutas), o teste Wilcoxon-Mann-Whitney torna o entendimento mais simples e fácil, pois se baseia nos dados originais do que ter saídas de expressões normalizadas.

Os trabalhos de [Geng et al. \(2014\)](#) e [Zhang et al. \(2017\)](#) sugeriram que miR-20a, miR-145, miR-21, miR-223 e miR-221 podem ser potenciais biomarcadores para NSCLC em estágio inicial. Embora esses estudos incluam diferentes subtipos histológicos de NSCLC, além do

adenocarcinoma, dois miRNAs em comum foram identificados, miR-21 e miR-20a, em T vs N, usando o teste de Wilcoxon-Mann-Whitney. Nenhum dos miRNAs mencionados acima foi identificado por Deseq2 e EdgeR.

Syednasrollah, Laiho e Elo (2015) relataram que ainda não há consenso sobre as melhores práticas de expressão diferencial para RNA-seq, tornando a escolha de um método adequado uma tarefa difícil. Muitos estudos na literatura usam Deseq2 e EdgeR para a identificação de DE-miRNAs (HU *et al.*, 2018; BOGEDALE *et al.*, 2019; HAMFJORD *et al.*, 2012; HE *et al.*, 2018; MA *et al.*, 2021). No entanto, alguns estudos relataram menos conservadores ao usar conjuntos de dados simulados com tamanhos de amostra pequenos (SCHURCH *et al.*, 2016; CORCHETE *et al.*, 2020; GAUTHIER *et al.*, 2020).

Os resultados mostram que os três métodos são úteis para a identificação de DE-miRNAs. No entanto, o teste de Wilcoxon-Mann-Whitney com a correção de Bonferroni demonstrou melhores resultados na identificação de DE-miRNAs nesse conjunto de dados, considerando os casos de adenocarcinoma pulmonar em estágio inicial. Destaca-se a importância de pesquisar, analisar e estudar diferentes métodos para identificação de DE-miRNAs, pois o método escolhido pode gerar resultados diferentes, conforme demonstrado aqui. Além disso, recomenda-se que os pesquisadores validem os resultados (experimentalmente ou *in silico*), usando conjuntos de dados independentes. Os métodos não são suficientes e completos em si e devem ser utilizados conjuntamente com outros métodos no processo de identificação dos potenciais biomarcadores para o adenocarcinoma de pulmão, em seu estágio inicial.

4.2 Análises II e III

Nesta seção dos resultados, foram utilizadas os bancos de dados descritos nas Tabelas 4 e 5, a primeira tem o propósito de distinguir T vs. N em cada banco de dados, enquanto a segunda foi empregada para separar LUSC vs. LUAD.

Visando garantir a consistência dos resultados, foram aplicadas metodologias idênticas em todos os casos. O fluxograma das etapas realizadas ao longo desse processo é apresentado na Figura 20.

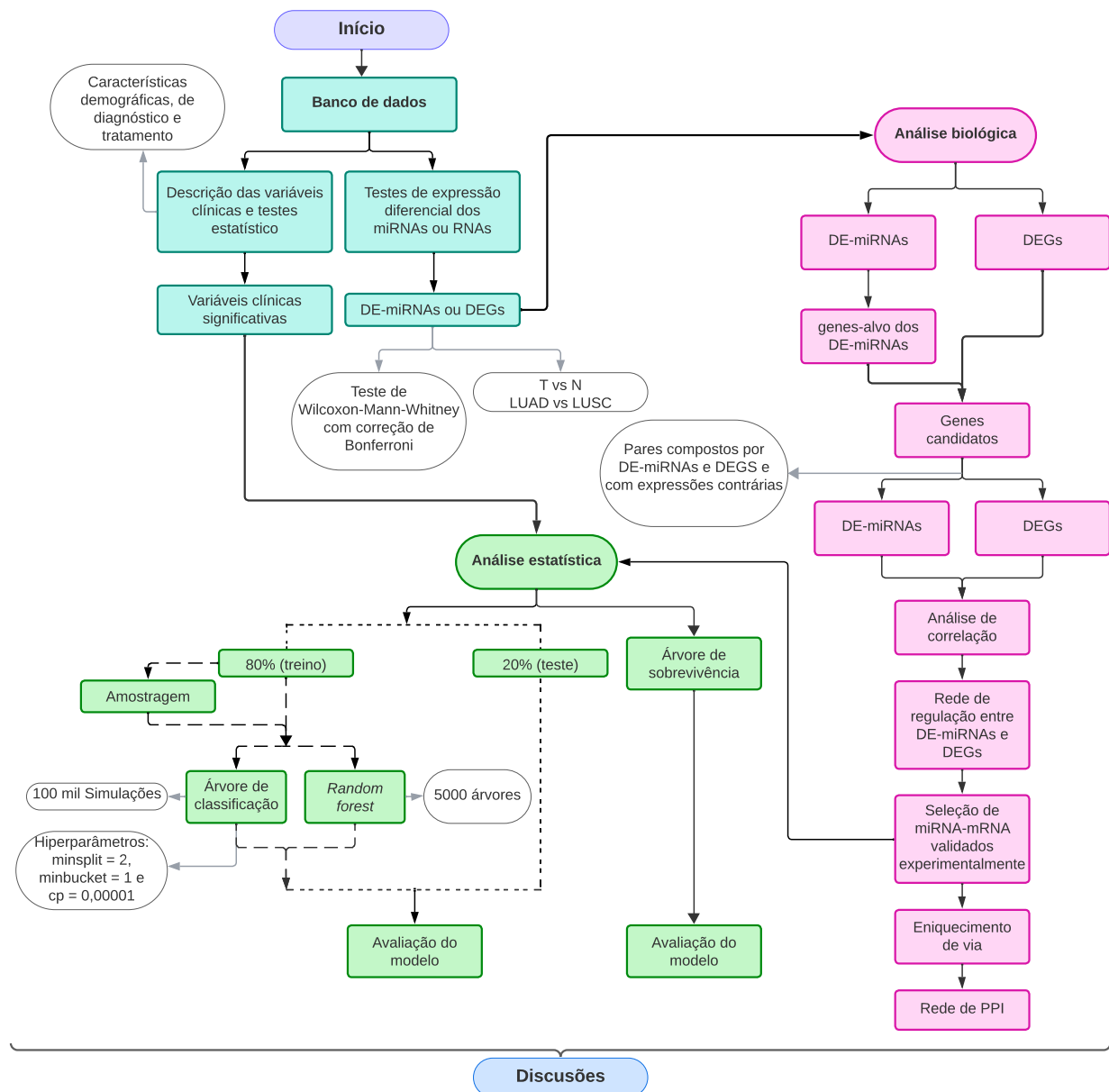


Figura 20 – Fluxograma esquemático da metodologia utilizada nas análises II e III.

4.2.1 Análise de expressão diferencial, biológica e estatística de LUAD

Esta seção apresenta os resultados e discussões obtidos para os bancos de dados LUAD miRNA-seq e RNA-seq. O banco de dados de miRNA-seq é composto por 299 amostras de pacientes, divididas em dois grupos: 270 amostras de T e 29 de N. Desses, 14 pacientes possuem amostras em ambos os grupos. Há expressão de 1057 miRNA-seq. O banco de dados de RNA-seq é formado por 300 pacientes, divididos em 270 amostras de T e 30 de N. Neste caso, 26 amostras possuem amostras em ambos os grupos, e há expressões de 40.599 RNA-seq. Ambos os bancos incluem 12 variáveis clínicas que descrevem características demográficas, de diagnóstico e tratamento dos pacientes, cujas definições podem ser encontradas na Tabela 2.

A Tabela 16 apresenta as características dos pacientes dos bancos de dados. É observada

uma predominância de amostras de pacientes do sexo feminino em ambas os bancos. A variável raça indica a maioria das pessoas como brancas e o estado vital predominante é "vivo". Em relação às variáveis de diagnóstico, o estágio IB, tamanho T2, diagnóstico primário de adenocarcinoma e localização no lobo superior ocorrem com maior frequência. A maioria dos pacientes não possui malignidade anterior e não recebeu tratamento ou terapia.

A média das idades dos pacientes é próxima, variando de 65 a 66 anos, e os dias do último acompanhamento também têm médias próximas, com variações de 1019 a 1162 dias, para ambos os bancos de dados. As Figuras 21 e 22 exibem esses comportamentos.

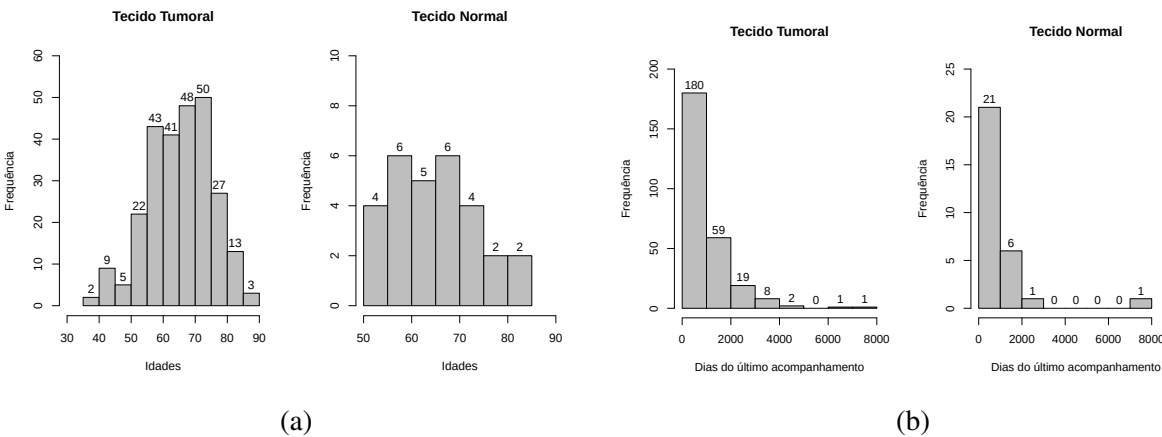


Figura 21 – Histograma da idade e dos dias do último acompanhamento - LUAD miRNA-seq.

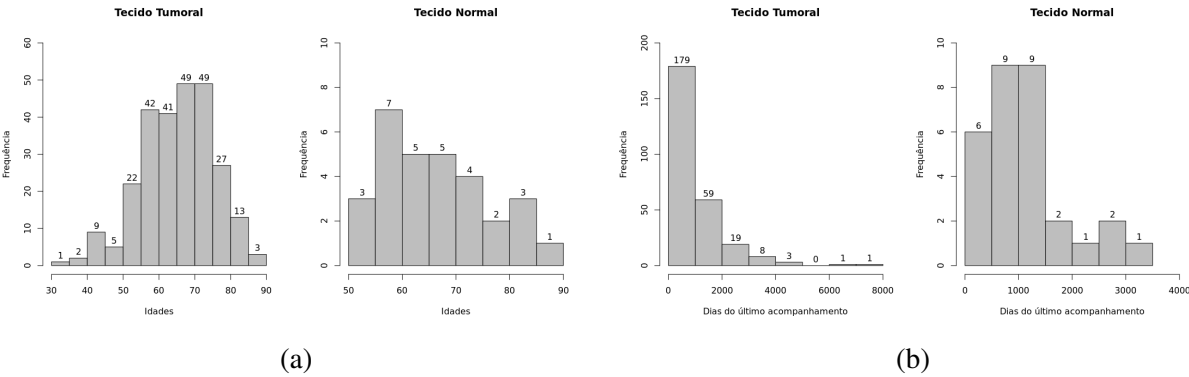


Figura 22 – Histograma da idade e do número de dias até o último acompanhamento - LUAD RNA-seq.

Tabela 16 – Características demográficas, diagnóstico e tratamento das amostras dos pacientes dos bancos de dados LUAD miRNA-seq e RNA-seq.

Variáveis	miRNA-seq					RNA-seq					
	T (n = 270)		N (n = 29)		Valor p	T (n = 270)		N (n = 30)		Valor p	
	n	%	n	%		n	%	n	%		
Sexo											
Feminino	158	90,29%	17	9,71%	0,99	159	89,83%	18	10,17%	0,91*	
Masculino	112	90,32%	12	9,68%		111	90,24%	12	9,76%		
Raça											
Branco	214	89,54%	25	10,46%	0,73	213	88,38%	28	11,62%	0,86**	
Negro ou afro-americano	25	86,21%	4	13,79%		26	92,86%	2	7,14%		
Asiático	5	100,00%	0	0,00%		5	100,00%	0	0,00%		
Estado Vital											
Vivo	204	89,87%	23	10,13%	0,65	205	89,91%	23	10,09%	0,93*	
Morto	66	91,67%	6	8,33%		65	90,28%	7	9,72%		
Estadiamento											
IB	134	91,16%	13	8,84%	0,82	135	89,40%	16	10,60%	0,91**	
IA	131	89,12%	16	10,88%		130	90,28%	14	9,72%		
I	5	100,00%	0	0,00%		5	100,00%	0	0,00%		
Tamanho do tumor (T)											
T2	86	92,47%	7	7,53%	0,95	87	85,29%	15	14,71%	0,04**	
T2a	51	89,47%	6	10,53%		51	98,08%	1	1,92%		
T1	50	89,29%	6	10,71%		50	84,75%	9	15,25%		
T1b	42	89,36%	5	10,64%		42	93,33%	3	6,67%		
T1a	41	89,13%	5	10,87%		40	95,24%	2	4,76%		
Diagnóstico primário											
Adenocarcinoma	163	85,79%	27	14,21%	0,04	163	87,17%	24	12,83%	0,08**	
Adenocarcinoma com subtipos mistos	51	98,08%	1	1,92%		51	98,08%	1	1,92%		
Carcinoma de células acinares	14	100,00%	0	0,00%		14	100,00%	0	0,00%		
Carcinoma bronquiolo-alveolar,não mucinoso	13	100,00%	0	0,00%		13	81,25%	3	18,75%		
Adenocarcinoma papilar	11	100,00%	0	0,00%		11	91,67%	1	8,33%		
Outros [◊]	18	94,74%	1	5,26%		18	94,74%	1	5,26%		
Tecido ou órgão de origem											
Lobo superior	171	90,96%	17	9,04%	0,42	172	90,05%	19	9,95%	0,79**	
Lobo inferior	86	90,53%	9	9,47%		85	90,43%	9	9,57%		
Outros [⊕]	13	81,25%	3	18,75%		13	86,67%	2	13,33%		
Malignidade anterior											
Não	227	91,90%	20	8,10%	0,04	227	90,80%	23	9,20%	0,30**	
Sim	43	82,69%	9	17,31%		43	86,00%	7	14,00%		
Tratamento ou terapia											
Não	211	89,03%	26	10,97%	0,59	211	89,03%	26	10,97%	0,59**	
Sim	39	92,86%	3	7,14%		39	92,86%	3	7,14%		
Idade											
Média	65,90		65,80		0,77	65,80		66,80		0,89•	
Mediana	67		65			67		65,5			
Dias para o último acompanhamento											
Média	1019,00		1062,00		0,79	1035,00		1162,00		0,08•	
Mediana	702		719			702		1073			

*Teste qui-quadrado. **Teste exato de Fisher. •Teste de Wilcoxon-Mann-Whitney.

Legenda: A categoria outros[◊] contém: adenocarcinoma de células claras, carcinoma micropapilar, carcinoma de células em anel de sinete, adenocarcinoma mucinoso, adenocarcinoma bronquioloalveolar, carcinoma sólido e carcinoma bronquioalveolar - mucinoso. A categoria outros[⊕] contém: brônquio principal, lesão sobreposta do pulmão e lobo médio.

Foram observadas diferenças estatisticamente significativas em relação ao tipo de tecido (T vs. N), com um valor de $p < 0,05$, para as seguintes variáveis: diagnóstico primário ($p = 0,04$) e malignidade anterior ($p = 0,04$) no banco de dados de miRNA-seq, e tamanho do tumor ($p = 0,04$) no banco de dados RNA-seq.

Análise de Expressão Diferencial

Realizou-se a análise de expressão diferencial de forma independente para cada miRNA e RNA, utilizando o teste de Wilcoxon-Mann-Whitney, a fim de identificar DE-miRNAs e DEGs na comparação entre as condições T vs. N. A significância estatística foi definida como $p < 0,05$. Para diminuir os efeitos de comparações múltiplas nas 1057 expressões de miRNA-seq, aplicou-se a correção de Bonferroni. Adotamos um novo nível de significância $\alpha_{\text{corrigido}}$:

$$\alpha_{\text{corrigido}} = \frac{\alpha_{\text{original}}}{n} = \frac{0,05}{1057} \approx 0,000047.$$

Observando um valor de $p < 0,000047$, identificamos a expressão de 254 DE-miRNAs, 206 *upregulated* e 48 *downregulated* em comparação com a mediana de N. Os resultados detalhados estão apresentados nas Tabelas 17 e 18, organizados com base no valor p .

Tabela 17 – DE-miRNAs *downregulated* (↓) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD.

miRNAs	Mediana		valor p	miRNAs	Mediana		valor p	miRNAs	Mediana		valor p
	N	T			N	T			N	T	
miR-139-3p	↓ 273,0	42,0	$p < 0,000047$	miR-143-3p	↓ 1819693,0	280713,5	$p < 0,000047$	miR-1537-5p	↓ 0,0	0,0	$p < 0,000047$
miR-486-5p	↓ 3951,0	386,5	$p < 0,000047$	miR-133a-3p	↓ 62,0	22,0	$p < 0,000047$	miR-139-5p	↓ 295,0	136,0	$p < 0,000047$
miR-30c-2-3p	↓ 725,0	106,0	$p < 0,000047$	miR-218-1-3p	↓ 12,0	4,0	$p < 0,000047$	miR-4524a-3p	↓ 1,0	0,0	$p < 0,000047$
miR-4732-3p	↓ 9,0	1,0	$p < 0,000047$	let-7c-5p	↓ 11128,0	4336,5	$p < 0,000047$	let-7b-3p	↓ 167,0	84,0	$p < 0,000047$
miR-378a-3p	↓ 3167,0	1075,0	$p < 0,000047$	miR-221-5p	↓ 30,0	10,0	$p < 0,000047$	miR-374a-5p	↓ 348,0	130,5	$p < 0,000047$
miR-1247-3p	↓ 149,0	29,5	$p < 0,000047$	miR-6720-3p	↓ 9,0	3,0	$p < 0,000047$	miR-338-5p	↓ 108,0	29,0	$p < 0,000047$
miR-30a-3p	↓ 81198,0	14830,5	$p < 0,000047$	miR-378c	↓ 44,0	17,0	$p < 0,000047$	miR-451a	↓ 9415,0	1578,0	$p < 0,000047$
miR-3154	↓ 1,0	0,0	$p < 0,000047$	let-7a-5p	↓ 325288,0	113621,0	$p < 0,000047$	miR-1-3p	↓ 177,0	34,5	$p < 0,000047$
miR-184	↓ 85,0	7,0	$p < 0,000047$	miR-6788-3p	↓ 1,0	0,0	$p < 0,000047$	miR-30a-5p	↓ 216167,0	55658,0	$p < 0,000047$
miR-516a-5p	↓ 11,0	1,0	$p < 0,000047$	miR-144-5p	↓ 2649,0	311,0	$p < 0,000047$	miR-4770	↓ 1,0	0,0	$p < 0,000047$
miR-4732-5p	↓ 1,0	0,0	$p < 0,000047$	let-7f-5p	↓ 153576,0	35700,5	$p < 0,000047$	miR-190a-5p	↓ 33,0	10,0	$p < 0,000047$
miR-490-3p	↓ 4,0	0,0	$p < 0,000047$	miR-195-5p	↓ 548,0	149,0	$p < 0,000047$	miR-144-3p	↓ 182,0	25,5	$p < 0,000047$
miR-133b	↓ 19,0	3,0	$p < 0,000047$	miR-3194-5p	↓ 0,0	0,0	$p < 0,000047$	miR-340-5p	↓ 202,0	71,5	$p < 0,000047$
miR-6892-5p	↓ 20,0	8,0	$p < 0,000047$	miR-584-5p	↓ 213,0	85,0	$p < 0,000047$	miR-218-5p	↓ 692,0	201,0	$p < 0,000047$
miR-4529-3p	↓ 4,0	0,0	$p < 0,000047$	miR-585-5p	↓ 1,0	0,0	$p < 0,000047$	miR-4523	↓ 0,0	0,0	$p < 0,000047$
miR-423-5p	↓ 444,0	223,0	$p < 0,000047$	miR-5010-5p	↓ 1,0	0,0	$p < 0,000047$	miR-584-3p	↓ 2,0	0,0	$p < 0,000047$

Tabela 18 – DE-miRNAs *upregulated* (↑) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD.

miRNAs	Mediana		valor p	miRNAs	Mediana		valor p	miRNAs	Mediana		valor p
	N	T			N	T			N	T	
miR-210-3p	↑ 64,0	3517,0	p< 0,000047	miR-625-3p	↑ 388,0	988,0	p< 0,000047	miR-542-5p	↑ 3,0	8,0	p< 0,000047
miR-1287-3p	↑ 2,0	27,0	p< 0,000047	miR-9-3p	↑ 0,0	4,0	p< 0,000047	miR-766-3p	↑ 12,0	39,5	p< 0,000047
miR-153-5p	↑ 2,0	35,5	p< 0,000047	miR-3653-5p	↑ 1,0	7,0	p< 0,000047	miR-377-3p	↑ 0,0	2,0	p< 0,000047
miR-21-3p	↑ 2199,0	12942,5	p< 0,000047	miR-625-5p	↑ 5,0	16,5	p< 0,000047	miR-5698	↑ 2,0	11,0	p< 0,000047
miR-708-3p	↑ 47,0	313,5	p< 0,000047	miR-2277-5p	↑ 0,0	3,0	p< 0,000047	miR-196b-5p	↑ 20,0	71,5	p< 0,000047
miR-96-5p	↑ 15,0	96,2	p< 0,000047	miR-579-5p	↑ 0,0	2,0	p< 0,000047	miR-4326	↑ 4,0	18,5	p< 0,000047
miR-301a-5p	↑ 1,0	16,0	p< 0,000047	miR-19b-1-5p	↑ 11,0	28,0	p< 0,000047	miR-576-5p	↑ 20,0	38,0	p< 0,000047
let-7g-3p	↑ 21,0	131,5	p< 0,000047	miR-200b-5p	↑ 25,0	74,5	p< 0,000047	miR-186-5p	↑ 775,0	1254,0	p< 0,000047
miR-9-5p	↑ 152,0	2397,0	p< 0,000047	miR-450a-5p	↑ 8,0	33,0	p< 0,000047	miR-182-3p	↑ 0,0	1,0	p< 0,000047
miR-7-1-3p	↑ 17,0	61,0	p< 0,000047	miR-425-5p	↑ 272,0	701,5	p< 0,000047	miR-16-1-3p	↑ 9,0	15,0	p< 0,000047
miR-183-5p	↑ 7137,0	40145,0	p< 0,000047	miR-424-5p	↑ 94,0	285,0	p< 0,000047	miR-200b-3p	↑ 1576,0	2988,0	p< 0,000047
miR-708-5p	↑ 22,0	131,3	p< 0,000047	miR-1269a	↑ 0,0	10,0	p< 0,000047	miR-5571-3p	↑ 0,0	1,0	p< 0,000047
miR-21-5p	↑ 376965,0	1485338,0	p< 0,000047	miR-503-3p	↑ 0,0	3,0	p< 0,000047	miR-891a-5p	↑ 2,0	7,0	p< 0,000047
miR-503-5p	↑ 5,0	27,0	p< 0,000047	miR-1301-3p	↑ 16,0	52,5	p< 0,000047	miR-3607-5p	↑ 0,0	0,3	p< 0,000047
miR-409-5p	↑ 1,0	15,0	p< 0,000047	miR-103a-2-5p	↑ 1,0	6,0	p< 0,000047	miR-589-5p	↑ 213,0	370,5	p< 0,000047
miR-577	↑ 0,0	11,0	p< 0,000047	miR-505-5p	↑ 8,0	26,0	p< 0,000047	miR-148b-3p	↑ 481,0	809,0	p< 0,000047
miR-4668-3p	↑ 0,0	7,0	p< 0,000047	miR-199a-5p	↑ 1226,0	3120,5	p< 0,000047	miR-20a-5p	↑ 460,0	965,5	p< 0,000047
miR-3607-3p	↑ 7,0	110,0	p< 0,000047	miR-214-5p	↑ 30,0	78,5	p< 0,000047	miR-187-3p	↑ 16,0	51,0	p< 0,000047
miR-26b-3p	↑ 22,0	102,5	p< 0,000047	miR-181b-2-3p	↑ 4,0	10,0	p< 0,000047	miR-148a-5p	↑ 125,0	256,5	p< 0,000047
miR-182-5p	↑ 18734,0	93499,0	p< 0,000047	miR-93-5p	↑ 6211,0	16511,0	p< 0,000047	miR-22-5p	↑ 67,0	121,5	p< 0,000047
miR-33a-5p	↑ 9,0	62,5	p< 0,000047	miR-429	↑ 195,0	466,0	p< 0,000047	miR-4726-5p	↑ 0,0	1,0	p< 0,000047
miR-20a-3p	↑ 3,0	19,0	p< 0,000047	miR-34a-3p	↑ 1,0	5,0	p< 0,000047	miR-92b-3p	↑ 325,0	759,5	p< 0,000047
miR-130b-5p	↑ 11,0	58,0	p< 0,000047	miR-3136-5p	↑ 0,0	2,0	p< 0,000047	miR-30e-5p	↑ 18047,0	29253,0	p< 0,000047
miR-200a-5p	↑ 451,0	2119,0	p< 0,000047	miR-365a-5p	↑ 0,0	2,0	p< 0,000047	miR-590-3p	↑ 12,0	21,5	p< 0,000047
miR-29b-1-5p	↑ 4,0	24,0	p< 0,000047	miR-181b-3p	↑ 44,0	91,0	p< 0,000047	miR-210-5p	↑ 2,0	5,0	p< 0,000047
miR-671-5p	↑ 3,0	16,0	p< 0,000047	miR-642a-5p	↑ 2,0	16,0	p< 0,000047	miR-450b-5p	↑ 28,0	55,0	p< 0,000047
miR-141-5p	↑ 645,0	2252,0	p< 0,000047	miR-26a-1-3p	↑ 1,0	4,0	p< 0,000047	miR-193b-5p	↑ 6,0	15,0	p< 0,000047
miR-142-3p	↑ 1732,0	9196,5	p< 0,000047	miR-4638-3p	↑ 0,0	2,0	p< 0,000047	miR-130a-5p	↑ 1,0	3,0	p< 0,000047
miR-33b-5p	↑ 1,0	12,0	p< 0,000047	miR-199b-3p	↑ 4101,0	8164,0	p< 0,000047	miR-219a-5p	↑ 1,0	3,0	p< 0,000047
miR-301a-3p	↑ 8,0	39,5	p< 0,000047	miR-199a-3p	↑ 4123,0	8174,5	p< 0,000047	miR-545-5p	↑ 0,0	1,0	p< 0,000047
miR-135b-5p	↑ 36,0	223,5	p< 0,000047	miR-3127-5p	↑ 7,0	21,0	p< 0,000047	miR-7705	↑ 2,0	4,0	p< 0,000047
miR-1307-5p	↑ 213,0	1134,0	p< 0,000047	miR-191-5p	↑ 812,0	1959,0	p< 0,000047	miR-29c-5p	↑ 61,0	107,5	p< 0,000047
miR-183-3p	↑ 0,0	5,0	p< 0,000047	miR-19a-3p	↑ 41,0	88,0	p< 0,000047	miR-548au-3p	↑ 0,0	1,0	p< 0,000047
miR-345-5p	↑ 13,0	65,5	p< 0,000047	miR-505-3p	↑ 62,0	146,5	p< 0,000047	miR-375-3p	↑ 25292,0	69901,0	p< 0,000047
miR-616-5p	↑ 1,0	8,0	p< 0,000047	miR-17-5p	↑ 607,0	1416,0	p< 0,000047	miR-556-5p	↑ 0,0	1,0	p< 0,000047
miR-143-5p	↑ 38,0	192,0	p< 0,000047	miR-3617-3p	↑ 0,0	1,0	p< 0,000047	miR-320b	↑ 7,0	16,0	p< 0,000047
miR-200a-3p	↑ 374,0	1330,0	p< 0,000047	miR-224-5p	↑ 45,0	128,5	p< 0,000047	miR-628-5p	↑ 30,0	49,0	p< 0,000047
miR-141-3p	↑ 1340,0	3564,0	p< 0,000047	miR-491-3p	↑ 0,0	3,0	p< 0,000047	miR-376b-3p	↑ 0,0	2,0	p< 0,000047
miR-629-3p	↑ 2,0	18,0	p< 0,000047	miR-592	↑ 1,0	9,0	p< 0,000047	miR-382-5p	↑ 14,0	28,0	p< 0,000047
miR-127-5p	↑ 86,0	375,0	p< 0,000047	miR-550a-5p	↑ 5,0	13,0	p< 0,000047	miR-4473	↑ 0,0	2,0	p< 0,000047
miR-135b-3p	↑ 0,0	7,0	p< 0,000047	miR-1307-3p	↑ 1198,0	3841,5	p< 0,000047	miR-3200-3p	↑ 5,0	13,0	p< 0,000047
miR-589-3p	↑ 2,0	10,0	p< 0,000047	miR-423-3p	↑ 189,0	451,5	p< 0,000047	miR-1291	↑ 0,0	1,0	p< 0,000047
miR-93-3p	↑ 10,0	38,0	p< 0,000047	miR-556-3p	↑ 0,0	1,0	p< 0,000047	miR-1248	↑ 0,0	1,0	p< 0,000047
miR-151a-5p	↑ 143,0	382,5	p< 0,000047	miR-4664-3p	↑ 0,0	1,0	p< 0,000047	miR-3651	↑ 0,0	2,0	p< 0,000047
miR-744-3p	↑ 10,0	40,5	p< 0,000047	miR-629-5p	↑ 213,0	468,5	p< 0,000047	miR-632	↑ 0,0	1,0	p< 0,000047
miR-301b-3p	↑ 0,0	4,0	p< 0,000047	miR-454-3p	↑ 11,0	29,5	p< 0,000047	miR-570-3p	↑ 0,0	1,0	p< 0,000047
miR-148a-3p	↑ 77566,0	186054,5	p< 0,000047	miR-937-3p	↑ 3,0	13,0	p< 0,000047	miR-6815-3p	↑ 0,0	1,0	p< 0,000047
let-7a-2-3p	↑ 7,0	30,0	p< 0,000047	miR-3189-3p	↑ 0,0	2,0	p< 0,000047	miR-424-3p	↑ 5,0	10,0	p< 0,000047
miR-539-5p	↑ 1,0	8,0	p< 0,000047	miR-1277-3p	↑ 1,0	4,0	p< 0,000047	miR-20b-3p	↑ 0,0	2,0	p< 0,000047
miR-548v	↑ 1,0	7,0	p< 0,000047	miR-331-5p	↑ 9,0	17,0	p< 0,000047	miR-409-3p	↑ 13,0	32,0	p< 0,000047
miR-130b-3p	↑ 20,0	66,0	p< 0,000047	miR-31-3p	↑ 1,0	5,0	p< 0,000047	miR-4645-3p	↑ 0,0	1,0	p< 0,000047
miR-590-5p	↑ 15,0	45,0	p< 0,000047	miR-4661-5p	↑ 4,0	11,0	p< 0,000047	miR-551b-3p	↑ 3,0	8,0	p< 0,000047
miR-147b-3p	↑ 0,0	4,0	p< 0,000047	miR-2355-5p	↑ 107,0	193,5	p< 0,000047	miR-199b-5p	↑ 328,0	627,5	p< 0,000047
miR-324-5p	↑ 32,0	101,5	p< 0,000047	miR-1229-3p	↑ 0,0	2,0	p< 0,000047	miR-4791	↑ 0,0	2,0	p< 0,000047
miR-29b-3p	↑ 1579,0	3595,0	p< 0,000047	miR-34a-5p	↑ 532,0	958,5	p< 0,000047	miR-105-5p	↑ 0,0	2,0	p< 0,000047
miR-3170	↑ 0,0	3,0	p< 0,000047	miR-769-3p	↑ 3,0	6,0	p< 0,000047	miR-493-3p	↑ 5,0	11,0	p< 0,000047
miR-29a-5p	↑ 13,0	38,0	p< 0,000047	miR-4491	↑ 0,0	2,0	p< 0,000047	miR-345-3p	↑ 0,0	1,0	p< 0,000047
miR-455-3p	↑ 99,0	512,0	p< 0,000047	miR-940	↑ 2,0	9,0	p< 0,000047	miR-132-5p	↑ 4,0	10,0	p< 0,000047
miR-155-3p	↑ 0,0	2,0	p< 0,000047	miR-339-5p	↑ 51,0	141,5	p< 0,000047	miR-758-3p	↑ 10,0	19,0	p< 0,000047
miR-136-5p	↑ 14,0	52,5	p< 0,000047	miR-4652-5p	↑ 0,0	1,0	p< 0,000047	miR-1271-5p	↑ 4,0	7,5	p< 0,000047
miR-301b-5p	↑ 0,0	2,0	p< 0,000047	miR-4677-3p	↑ 18,0	32,0	p< 0,000047	miR-4728-3p	↑ 3,0	8,0	p< 0,000047
miR-616-3p	↑ 0,0	3,0	p< 0,000047	miR-4724-5p	↑ 1,0	3,0	p< 0,000047	miR-1254	↑ 1,0	2,3	p< 0,000047
miR-188-3p	↑ 0,0	3,0	p< 0,000047	miR-137-3p	↑ 0,0	1,0	p< 0,000047	miR-651-5p	↑ 6,0	11,0	p< 0,000047
miR-331-3p	↑ 53,0	147,5	p< 0,000047	miR-548s	↑ 0,0	1,0	p< 0,000047	miR-323a-3p	↑ 1,0	4,0	p< 0,000047
miR-323b-3p	↑ 0,0	5,0	p< 0,000047	miR-639	↑ 0,0	2,0	p< 0,000047	miR-3934-3p	↑ 5,0	11,0	p< 0,000047
miR-31-5p	↑ 0,0	9,5	p< 0,000047	miR-627-5p	↑ 2,0	5,0	p< 0,000047	miR-3664-5p	↑ 0,0	1,0	p< 0,000047
miR-196a-5p	↑ 3,0	41,5	p< 0,000047	miR-659-5p	↑ 2,0	6,0	p< 0,000047	miR-361-3p	↑ 601,0	953,0	p< 0,000047
miR-200c-5p	↑ 54,0	118,0	p< 0,000047	miR-148b-5p	↑ 4,0	9,0	p< 0,000047	miR-501-5p	↑ 9,0	16,0	p< 0,000047
miR-193b-3p	↑ 52,0	203,5	p< 0,000047	miR-24-2-5p	↑ 46,0	94,0	p< 0,000047				

O teste de Wilcoxon-Mann-Whitney com correção de Bonferroni também foi realizado nas 40.599 expressões de RNA-seq, para isso, considerou-se como estatisticamente significativo o valor de $\alpha_{\text{corrigido}}$:

$$\alpha_{\text{corrigido}} = \frac{\alpha_{\text{original}}}{n} = \frac{0,05}{40.599} \approx 0,0000012.$$

Considerando, valor de $p < 0,0000012$ identificou-se 7.374 DEGs, sendo que 5.158 *upregulated* e 2.216 *downregulated*. As Tabelas 19 e 20 apresentam os 200 DEGs mais importantes (considerando menores valores de p), observa-se mais classificações como *downregulated*.

Tabela 19 – DEGs *upregulated* (↑) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD.

RNAs	Mediana		valor p	RNAs	Mediana		valor p	RNAs	Mediana		valor p
	N	T			N	T			N	T	
UBE2T	↑ 63,5	665,5	p<0,0000012	MEX3A	↑ 55,0	911,5	p<0,0000012	KIF20A	↑ 74,5	631,5	p<0,0000012
PYCR1	↑ 224,0	3354,0	p<0,0000012	ANLN	↑ 110,0	1229,5	p<0,0000012	CDC6	↑ 69,0	683,5	p<0,0000012
AC004816.1	↑ 26,0	196,5	p<0,0000012	FAM83A	↑ 33,0	2853,0	p<0,0000012	AL162595.1	↑ 152,0	735,0	p<0,0000012
SAPCD2	↑ 30,5	332,0	p<0,0000012	ADM2	↑ 48,5	697,5	p<0,0000012	TSACC	↑ 3,5	26,0	p<0,0000012
TEDC2	↑ 19,0	229,7	p<0,0000012	SPAG4	↑ 79,0	585,5	p<0,0000012	EZH2	↑ 128,5	788,0	p<0,0000012
ERCC6L	↑ 18,5	151,5	p<0,0000012	GOLM1	↑ 1741,0	11173,0	p<0,0000012	HJURP	↑ 30,5	463,0	p<0,0000012
EFNA3	↑ 19,5	147,5	p<0,0000012	IQGAP3	↑ 74,5	1106,3	p<0,0000012	CDC20	↑ 82,5	1036,3	p<0,0000012
ETV4	↑ 128,5	1881,0	p<0,0000012	LINC02014	↑ 2,0	34,5	p<0,0000012	KIF4A	↑ 45,5	592,5	p<0,0000012
GINS1	↑ 57,5	557,5	p<0,0000012	NUF2	↑ 29,5	366,5	p<0,0000012	LGR4	↑ 359,5	2628,0	p<0,0000012
CDC25C	↑ 10,0	131,5	p<0,0000012	PSAT1	↑ 91,0	824,0	p<0,0000012	CTHRC1	↑ 158,5	2553,5	p<0,0000012
TOP2A	↑ 289,0	3741,0	p<0,0000012	DEPDC1	↑ 17,0	263,0	p<0,0000012	SRPK1	↑ 578,5	1867,5	p<0,0000012
GPT2	↑ 257,5	2243,0	p<0,0000012	ALDH18A1	↑ 1299,0	4704,0	p<0,0000012	NEK2	↑ 38,0	439,5	p<0,0000012
DEPDC1B	↑ 20,5	218,0	p<0,0000012	AC145343.1	↑ 7,0	44,0	p<0,0000012	TROAP	↑ 31,0	340,0	p<0,0000012
ATP10B	↑ 8,5	261,0	p<0,0000012	PPAT	↑ 173,5	777,0	p<0,0000012	DLGAP5	↑ 37,0	490,5	p<0,0000012
PAICS	↑ 1310,5	5211,0	p<0,0000012	EFNA4	↑ 125,5	781,0	p<0,0000012	EXO1	↑ 27,5	290,0	p<0,0000012
OCIAD2	↑ 827,5	4529,0	p<0,0000012	XRCC2	↑ 28,0	232,5	p<0,0000012	UBE2C	↑ 65,5	1033,0	p<0,0000012
NME1	↑ 127,5	839,5	p<0,0000012	FAM83A.AS1	↑ 2,0	96,0	p<0,0000012				
LINC01977	↑ 3,0	73,8	p<0,0000012	RDM1	↑ 2,0	26,5	p<0,0000012				

Tabela 20 – DEGs *downregulated* (↓) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD.

miRNAs	Mediana		valor p	miRNAs	Mediana		valor p	miRNAs	Mediana		valor p
	N	T			N	T			N	T	
GDF2	↓ 3,0	0,0	p<0,0000012	PALM2AKAP2	↓ 20525,0	3614,0	p<0,0000012	C14orf132	↓ 3817,5	591,5	p<0,0000012
AL355499.2	↓ 5,5	0,0	p<0,0000012	FHL1	↓ 11960,0	1147,0	p<0,0000012	HTR3C	↓ 59,0	1,0	p<0,0000012
AC095050.1	↓ 21,5	0,0	p<0,0000012	PTPRQ	↓ 342,0	10,0	p<0,0000012	PRX	↓ 5218,5	485,0	p<0,0000012
RNU6.955P	↓ 3,5	0,0	p<0,0000012	PTPN21	↓ 3687,5	747,2	p<0,0000012	INMT	↓ 8844,0	730,5	p<0,0000012
AC010776.3	↓ 7,5	0,0	p<0,0000012	LANCL1.AS1	↓ 241,5	16,0	p<0,0000012	EPAS1	↓ 105828,5	14900,5	p<0,0000012
AC093787.1	↓ 11,0	0,0	p<0,0000012	ARHGEF26	↓ 3917,5	598,0	p<0,0000012	LDB2	↓ 3101,5	577,0	p<0,0000012
AC104257.1	↓ 11,5	0,0	p<0,0000012	LGI3	↓ 3564,0	41,5	p<0,0000012	OVCH1	↓ 109,5	5,0	p<0,0000012
AC097537.1	↓ 3,0	0,0	p<0,0000012	GRK5	↓ 4744,0	830,5	p<0,0000012	SEMA3G	↓ 2959,5	334,5	p<0,0000012
NCAPGP2	↓ 19,5	0,0	p<0,0000012	SLC6A4	↓ 6467,0	24,0	p<0,0000012	AC084871.3	↓ 106,0	9,0	p<0,0000012
AC128709.3	↓ 30,5	0,0	p<0,0000012	ADRB2	↓ 1746,5	226,0	p<0,0000012	MGAT3	↓ 2777,0	219,5	p<0,0000012
LINC01082	↓ 9,0	0,0	p<0,0000012	UPK3B	↓ 5578,0	124,5	p<0,0000012	AC112206.2	↓ 188,5	17,0	p<0,0000012
AC104237.2	↓ 14,0	0,0	p<0,0000012	LYVE1	↓ 2347,5	233,5	p<0,0000012	NECAB1	↓ 735,0	68,0	p<0,0000012
AL606469.1	↓ 99,5	1,0	p<0,0000012	WNT3A	↓ 610,0	16,0	p<0,0000012	OTUD1	↓ 2948,0	823,0	p<0,0000012
SSTR4	↓ 12,0	0,3	p<0,0000012	C10orf67	↓ 245,0	10,0	p<0,0000012	GIMAP8	↓ 3064,5	667,5	p<0,0000012
LINC01070	↓ 4,0	0,0	p<0,0000012	EDNRB	↓ 7106,5	532,0	p<0,0000012	PECAM1	↓ 36988,5	7847,0	p<0,0000012
AC010776.2	↓ 15,5	0,0	p<0,0000012	AC044810.2	↓ 34,5	2,0	p<0,0000012	CD300LG	↓ 497,5	8,0	p<0,0000012
AC104984.4	↓ 29,5	1,0	p<0,0000012	LIMS2	↓ 4603,0	669,5	p<0,0000012	HEG1	↓ 12600,5	3063,0	p<0,0000012
STX11	↓ 3098,0	439,5	p<0,0000012	LINC02570	↓ 17,5	0,0	p<0,0000012	ROBO4	↓ 4964,0	673,5	p<0,0000012
AGER	↓ 85101,0	1219,0	p<0,0000012	AC135012.3	↓ 47,0	2,0	p<0,0000012	ADAMTS8	↓ 2163,5	148,5	p<0,0000012
HBM	↓ 4,5	0,0	p<0,0000012	RSPO1	↓ 252,0	21,0	p<0,0000012	TMEM100	↓ 4241,0	198,0	p<0,0000012
LINC01996	↓ 121,5	3,0	p<0,0000012	JAM2	↓ 2011,5	448,5	p<0,0000012	AL354714.3	↓ 15,5	1,0	p<0,0000012
AC104237.3	↓ 17,5	1,0	p<0,0000012	ST8SIA6	↓ 1228,5	79,5	p<0,0000012	AQP4	↓ 22543,5	2100,5	p<0,0000012
SERTM1	↓ 672,0	5,0	p<0,0000012	AL355499.1	↓ 21,0	1,0	p<0,0000012	RGCC	↓ 11951,0	1805,5	p<0,0000012
ITLN2	↓ 1022,0	7,0	p<0,0000012	CNTN6	↓ 609,0	27,5	p<0,0000012	RAMP3	↓ 3951,0	557,0	p<0,0000012
LINC02016	↓ 135,0	2,0	p<0,0000012	GPD1	↓ 2286,0	106,0	p<0,0000012	BTNL9	↓ 2553,0	184,5	p<0,0000012
AC093110.1	↓ 550,5	52,0	p<0,0000012	FENDRR	↓ 2414,0	115,5	p<0,0000012	CALCRL	↓ 8116,0	1254,0	p<0,0000012
TNNC1	↓ 2550,5	214,0	p<0,0000012	WWC2	↓ 6251,0	1261,0	p<0,0000012	LINC00656	↓ 30,5	2,0	p<0,0000012
AC084030.1	↓ 8,5	0,0	p<0,0000012	MYZAP	↓ 441,5	49,0	p<0,0000012	SEMA6A	↓ 1881,5	272,0	p<0,0000012
SFTPC	↓ 892145,5	10123,0	p<0,0000012	FMO2	↓ 9367,0	1003,0	p<0,0000012	LIN7A	↓ 998,5	122,5	p<0,0000012
CLEC3B	↓ 1877,5	131,0	p<0,0000012	SGCG	↓ 162,5	11,0	p<0,0000012	AL590226.1	↓ 129,0	11,5	p<0,0000012
CAV1	↓ 52848,0	3618,5	p<0,0000012	AC093890.1	↓ 55,5	8,0	p<0,0000012	CABP5	↓ 1,5	0,0	p<0,0000012
GPM6A	↓ 2258,5	58,0	p<0,0000012	STXBP6	↓ 1863,0	104,5	p<0,0000012	KCNA4	↓ 146,0	4,0	p<0,0000012
AL354714.1	↓ 57,5	1,0	p<0,0000012	AC067752.1	↓ 16,5	1,0	p<0,0000012	ANKRD1	↓ 2618,5	42,5	p<0,0000012
GYPE	↓ 146,5	15,0	p<0,0000012	ACADL	↓ 1160,0	108,5	p<0,0000012	S1PR1	↓ 6624,5	1002,0	p<0,0000012
TEK	↓ 4212,5	415,0	p<0,0000012	MYOC	↓ 154,5	3,0	p<0,0000012	CDH5	↓ 10879,5	1822,0	p<0,0000012
FAM107A	↓ 7568,0	280,5	p<0,0000012	FREM3	↓ 101,5	3,0	p<0,0000012	OR6N1	↓ 6,5	0,0	p<0,0000012
FABP4	↓ 4550,0	83,5	p<0,0000012	CA4	↓ 1687,5	45,5	p<0,0000012	CCDC141	↓ 318,5	38,0	p<0,0000012
RS1	↓ 158,0	4,0	p<0,0000012	CD5L	↓ 92,0	4,0	p<0,0000012	VEGFD	↓ 1489,0	116,0	p<0,0000012
SH3GL3	↓ 242,5	5,0	p<0,0000012	MCEMP1	↓ 3320,5	126,5	p<0,0000012	FGD5	↓ 3655,0	791,5	p<0,0000012
NCKAP5	↓ 2510,5	130,0	p<0,0000012	SLC39A8	↓ 25477,5	4690,0	p<0,0000012	ABI3BP	↓ 5398,0	1122,2	p<0,0000012
RTKN2	↓ 10609,5	526,5	p<0,0000012	RAMP2	↓ 2961,5	513,5	p<0,0000012	KANK3	↓ 955,5	147,5	p<0,0000012
AL445307.1	↓ 17,0	2,0	p<0,0000012	CLDN18	↓ 38860,0	665,0	p<0,0000012	ANOS1	↓ 7671,5	1365,5	p<0,0000012
AC022325.2	↓ 5,5	0,0	p<0,0000012	CHRNA2	↓ 30,0	1,0	p<0,0000012	ACVRL1	↓ 7492,5	1310,0	p<0,0000012
LINC01837	↓ 4,5	0,0	p<0,0000012	FAM189A2	↓ 3587,5	547,5	p<0,0000012	AC009093.3	↓ 564,5	21,2	p<0,0000012
ANGPT4	↓ 258,5	11,0	p<0,0000012	AL136369.1	↓ 20,5	0,0	p<0,0000012	CD36	↓ 7667,0	716,0	p<0,0000012
CAVIN2	↓ 13919,5	1274,0	p<0,0000012	CHRM1	↓ 145,5	3,0	p<0,0000012	CYP1A2	↓ 57,0	1,0	p<0,0000012
CLIC5	↓ 14214,0	991,5	p<0,0000012	HSPA12B	↓ 1408,5	220,5	p<0,0000012	ADH1B	↓ 13623,0	1185,0	p<0,0000012
SCUBE1	↓ 1914,5	92,5	p<0,0000012	GPIHBP1	↓ 1954,5	91,0	p<0,0000012	ADRB1	↓ 128,5	92,0	p<0,0000012
EMP2	↓ 53714,0	7876,5	p<0,0000012	LINC00968	↓ 223,0	24,0	p<0,0000012				
SPOCK2	↓ 25663,0	2848,5	p<0,0000012	TAL1	↓ 641,5	98,5	p<0,0000012				

Análises Biológicas

As análises biológicas foram conduzidas com o objetivo de integrar e analisar os DE-miRNAs (n=254, sendo 206 *upregulated* e 48 *downregulated*) e DEGs (n=7.374, sendo 5.158 *upregulated* e 2.216 *downregulated*) previamente identificados.

Realizou-se a predição dos genes-alvos dos DE-miRNAs por meio da ferramenta mirDIP. Obteve-se uma lista com 44.873 genes-alvos dos DE-miRNAs (pares de combinações miRNA - gene-alvo), dos quais, sem duplicatas, contêm 9.841 genes-alvos. Foi construído um Diagrama de Venn (Figura 23) entre os genes-alvos e os DEGs para identificar sobreposições (interseções)

entre eles. Observa-se sobreposição de 18% (2.627 DEGs) entre os genes-alvos dos DE-miRNAs e os DEGs. Os genes que fazem parte dessa sobreposição serão denominados DEGs candidatos e serão submetidos às próximas análises.

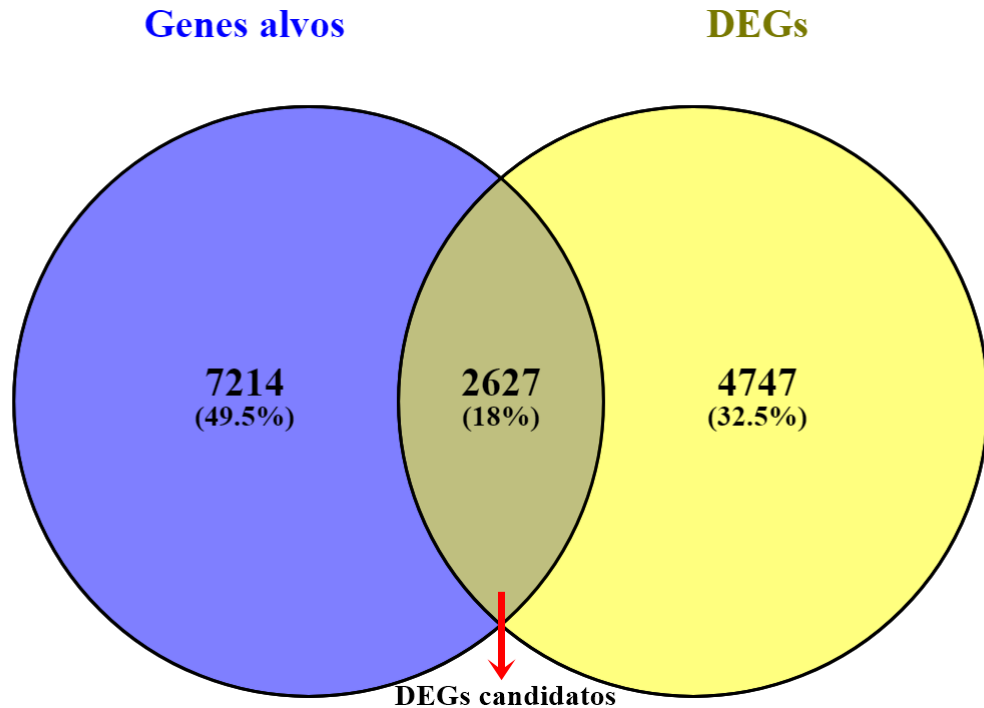


Figura 23 – Diagrama de Venn entre os genes-alvos dos DE-miRNAs e os DEGs - LUAD.

A lista gerada pelo mirDIP com os genes-alvos foi filtrada com base nos DEGs candidatos, resultando na identificação de 10.963 pares formados pelos DEGs candidatos e pelos DE-miRNAs que os regulam. Posteriormente a essa filtragem, foram selecionados DEGs candidatos cuja expressão é contrária à dos DE-miRNAs. Por exemplo, se a expressão do miRNA está diminuída, foram selecionados DEGs candidatos com expressão aumentada. Essa abordagem resultou na identificação de 4.935 pares com expressões contrárias.

Para realizar a análise de correlação linear entre 4.935 pares com expressões contrárias, foi necessário unir os bancos de dados de miRNA-seq ($n = 299$) e RNA-seq ($n = 300$) com base no ID da amostra do paciente. O banco de dados unificado consistiu em $n = 274$ amostras de pacientes que eram comuns a ambos os conjuntos.

A correlação linear de Spearman foi calculada considerando os pares que apresentaram correlações negativas com $\rho < -0,3$ e $p < 0,05$, como estatisticamente significativos. Essa análise identificou 231 pares (Tabela S1, no material suplementar) e os 50 pares com as maiores correlações negativas estão apresentados na Tabela 21, em que a menor correlação foi $\rho = -0,46$ entre miR-130b-3p e *TGFBR2* (*transforming Growth Factor Beta Receptor 2*).

Tabela 21 – As 50 maiores correlações negativas entre os pares - LUAD.

miRNA	Gene	ρ	p
miR-130b-3p	TGFBR2	-0,46	p<0,05
miR-139-3p	LAGE3	-0,46	p<0,05
miR-147b-3p	NFIX	-0,43	p<0,05
miR-130b-3p	ROBO2	-0,43	p<0,05
miR-30a-5p	R3HDM1	-0,43	p<0,05
miR-30c-2-3p	MCM6	-0,42	p<0,05
miR-30c-2-3p	E2F2	-0,42	p<0,05
miR-30a-3p	INTS8	-0,42	p<0,05
miR-30c-2-3p	MTCH2	-0,41	p<0,05
miR-9-5p	FLRT3	-0,41	p<0,05
miR-378a-3p	UBE2W	-0,41	p<0,05
miR-184	SLC7A5	-0,41	p<0,05
miR-195-5p	CDCA4	-0,40	p<0,05
miR-30a-3p	NAA25	-0,40	p<0,05
miR-1247-3p	B4GALT3	-0,40	p<0,05
miR-200b-3p	CFL2	-0,40	p<0,05
miR-1247-3p	PTGES2	-0,39	p<0,05
miR-139-3p	ALYREF	-0,39	p<0,05
miR-195-5p	BZW1	-0,39	p<0,05
miR-195-5p	TLK1	-0,39	p<0,05
let-7c-5p	ERO1A	-0,39	p<0,05
miR-30a-3p	ARMC1	-0,39	p<0,05
miR-195-5p	NAA25	-0,39	p<0,05
miR-195-5p	USP14	-0,39	p<0,05
let-7c-5p	MARS2	-0,38	p<0,05
let-7a-5p	TARBP2	-0,38	p<0,05
miR-30a-3p	LMNB1	-0,38	p<0,05
let-7f-5p	TARBP2	-0,38	p<0,05
miR-221-5p	CREBZF	-0,38	p<0,05
let-7c-5p	CDC25A	-0,37	p<0,05
miR-130b-3p	SECISBP2L	-0,37	p<0,05
miR-17-5p	TGFBR2	-0,37	p<0,05
miR-139-3p	PYM1	-0,37	p<0,05
miR-130b-5p	GRIA1	-0,37	p<0,05
miR-671-5p	DLC1	-0,37	p<0,05
miR-30c-2-3p	MRPL19	-0,37	p<0,05
miR-30a-5p	YWHAZ	-0,37	p<0,05
miR-378a-3p	ZDHHC9	-0,37	p<0,05
miR-30a-5p	CALU	-0,37	p<0,05
miR-1-3p	HSPD1	-0,37	p<0,05
let-7c-5p	CDC34	-0,37	p<0,05
miR-141-3p	QKI	-0,37	p<0,05
miR-96-5p	TACC1	-0,37	p<0,05
miR-183-5p	ZEB2	-0,37	p<0,05
miR-1247-3p	AP5Z1	-0,37	p<0,05
miR-141-3p	ZEB2	-0,37	p<0,05
miR-96-5p	FHL1	-0,37	p<0,05
miR-195-5p	KIF23	-0,37	p<0,05
miR-96-5p	MITF	-0,37	p<0,05
miR-143-3p	MOGS	-0,37	p<0,05

Os 231 pares foram submetidos à análise no miRTarBase, com o objetivo de selecionar interações entre DE-miRNAs e DEGs candidatos que sejam validadas experimentalmente. Dessa análise, identificaram-se 81 pares, dos quais 28 consistem em DE-miRNAs *upregulated* e DEGs candidatos *downregulated* (Tabela 22), enquanto 53 envolvem DE-miRNAs *downregulated* e DEGs candidatos *upregulated* (Tabela 25).

Tabela 22 – DE-miRNAs *up* e DEGs *down* validados experimentalmente - LUAD.

ID	miRNA		Gene	
MIRT020143	miR-130b-3p	↑	TGFBR2	↓
MIRT020145	miR-130b-3p	↑	SECISBP2L	↓
MIRT004591	miR-17-5p	↑	TGFBR2	↓
MIRT099339	miR-141-3p	↑	QKI	↓
MIRT000316	miR-141-3p	↑	ZEB2	↓
MIRT735992	miR-96-5p	↑	FHL1	↓
MIRT001989	miR-96-5p	↑	MITF	↓
MIRT027974	miR-93-5p	↑	TGFBR2	↓
MIRT020226	miR-130b-3p	↑	ZBTB4	↓
MIRT001990	miR-182-5p	↑	MITF	↓
MIRT224755	miR-130b-3p	↑	DPYSL2	↓
MIRT004619	miR-590-5p	↑	TGFBR2	↓
MIRT733880	miR-9-5p	↑	STARD13	↓
MIRT099340	miR-200a-3p	↑	QKI	↓
MIRT091701	miR-301b-3p	↑	TGFBR2	↓
MIRT438055	miR-17-5p	↑	ETV1	↓
MIRT001785	miR-20a-5p	↑	TGFBR2	↓
MIRT021484	miR-9-5p	↑	ID4	↓
MIRT218291	miR-200b-3p	↑	QKI	↓
MIRT054654	miR-96-5p	↑	ZEB1	↓
MIRT091690	miR-301a-3p	↑	TGFBR2	↓
MIRT000048	miR-429	↑	ZEB2	↓
MIRT299337	miR-20a-5p	↑	CYBRD1	↓
MIRT091700	miR-454-3p	↑	TGFBR2	↓
MIRT224763	miR-301b-3p	↑	DPYSL2	↓
MIRT003783	miR-200b-3p	↑	ETS1	↓
MIRT438054	miR-20a-5p	↑	ETV1	↓
MIRT001035	miR-200a-3p	↑	ZEB2	↓

↑ *upregulated* e ↓ *downregulated*.

A rede de regulatória entre DE-miRNAs e DEGs candidatos, apresentados na Tabela 22, foi construída e está apresentada na Figura 24. A rede destaca o gene *TGFBR2*, *dowregulated* ligado a 8 miRNAs *upregulated*: miR-93-5p, miR-590-5p, miR-301a-3p, miR-454-3p, miR-301b-3p e miR-17-5p, miR-20a-5p e miR-130b-3p. Além disso, também destacam o miR-130b-3p ligado a *TGFRB2*, *ZBTB4*, *SECISBP2L* e *DPYSL2* e o *QKI* está ligado aos miR-141-3p, miR-200a-3p e miR-200b-3p e o miR-96-5p está ligado aos genes *ZEB1*, *FHL1* e *MITF*.

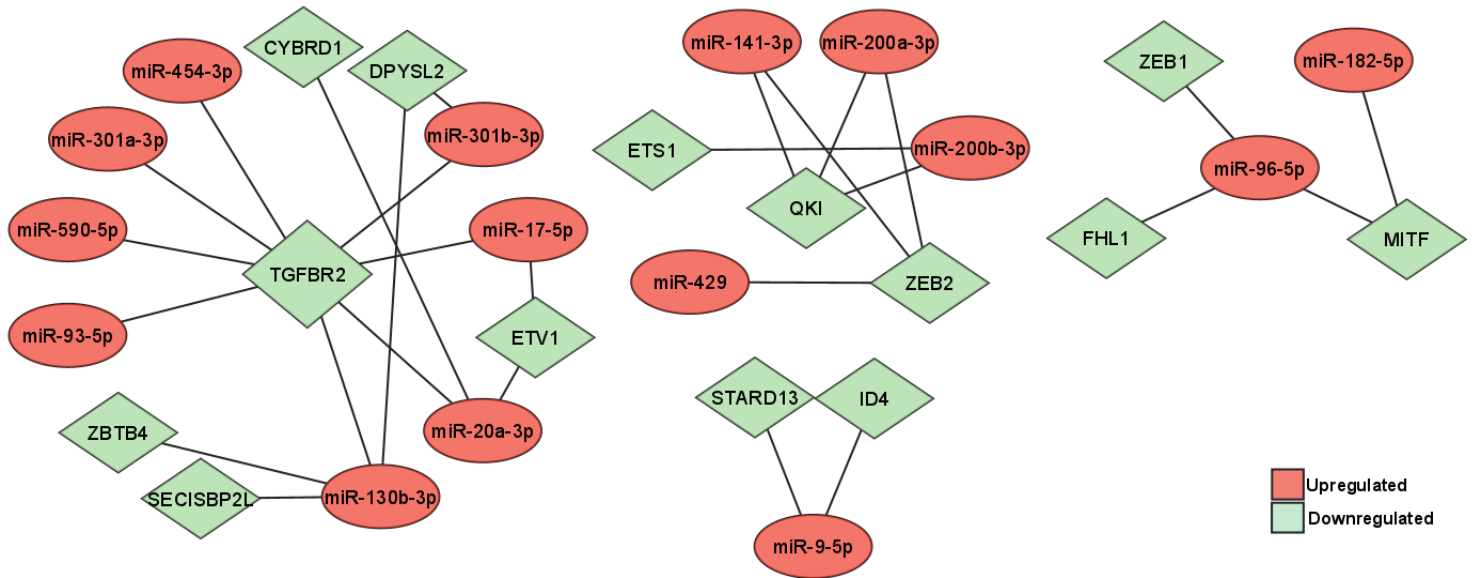


Figura 24 – Rede regulatória entre DE-miRNA *up* e DEGs *down* - LUAD.

Posteriormente, o enriquecimento das vias foi realizado para os DE-miRNAs e DEGs presentes na Tabela 22. Para os DE-miRNAs foi utilizado a ferramenta miRPath e para os DEGs candidatos a ferramenta Enrichr. A Tabela 23 apresenta as principais vias identificadas com os DE-miRNAs que estão relacionadas a patogênese do LUAD, a Tabela S2 (do material suplementar) apresenta a lista completa, e 24 apresentam todas as vias identificadas com os RNAs, considerando valor $p < 0,05$.

Tabela 23 – Principais vias enriquecidas para os DE-miRNAs *up* - LUAD.

<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes</i>	<i>miRNAs (n)</i>	FDR
Cell cycle	129	75	15	$p < 0,01$
Pathways in cancer	555	221	15	$p < 0,01$
Proteoglycans in cancer	220	103	15	$p < 0,01$
Focal adhesion	213	100	15	$p < 0,01$
HIF-1 signaling pathway	112	58	15	$p < 0,01$
p53 signaling pathway	75	43	15	$p < 0,01$
EGFR tyrosine kinase inhibitor resistance	82	45	15	$p < 0,01$
Small cell lung cancer	100	52	15	$p < 0,01$
MAPK signaling pathway	329	132	15	$p < 0,01$
AMPK signaling pathway	130	61	15	$p < 0,01$
mTOR signaling pathway	177	77	15	$p < 0,01$
ErbB signaling pathway	86	44	15	$p < 0,01$
PI3K-Akt signaling pathway	372	140	15	$p < 0,01$
Non-small cell lung cancer	81	41	15	$p < 0,01$
Transcriptional misregulation in cancer	206	78	15	0,001
Wnt signaling pathway	173	67	15	0,002
Viral carcinogenesis	265	96	15	0,002

Tabela 24 – Vias enriquecidas para os DEGs *down* - LUAD.

KEGG pathway	Overlap	Adjusted p value	Genes
Transcriptional misregulation in cancer	4/192	0,000	ZEB1;MITF;ETV1;TGFB2
TGF-beta signaling pathway	2/94	0,033	ID4;TGFB2
Osteoclast differentiation	2/127	0,037	MITF;TGFB2
Cellular senescence	2/156	0,005	ETS1;TGFB2
Pathways in cancer	3/531	0,037	MITF;ETS1;TGFB2

Os 53 pares de DE-miRNAs *downregulated* e DEGs candidatos *upregulated* estão apresentados na Tabela 25. A rede de regulatória entre DE-miRNAs e DEGs candidatos foi construída e está apresentada na Figura 25. Observa-se vários DE-miRNAs ligados á vários DEGs, miR-195-5p está relacionada a 14 DEGs, que são: CDCA4, BZW1, TLK1, NAA25, KIF23, CHEK1, E2F3, UBE2Q1, CCNE1, E2F7, SLC25A22, TBL1XR1, PEX13 e CCNT2. Além desse, o let-7c-5p esta ligado a ATP2A2, SLC20A1, E2F5, HMGA1, CDC25A, ERO1A, BZW1 e MARS2. O miR-30a-5p está com YWHAZ, PHTF2, CELSR3, TWF1, LCLAT1, EED, CCNE2 e TFDP1.

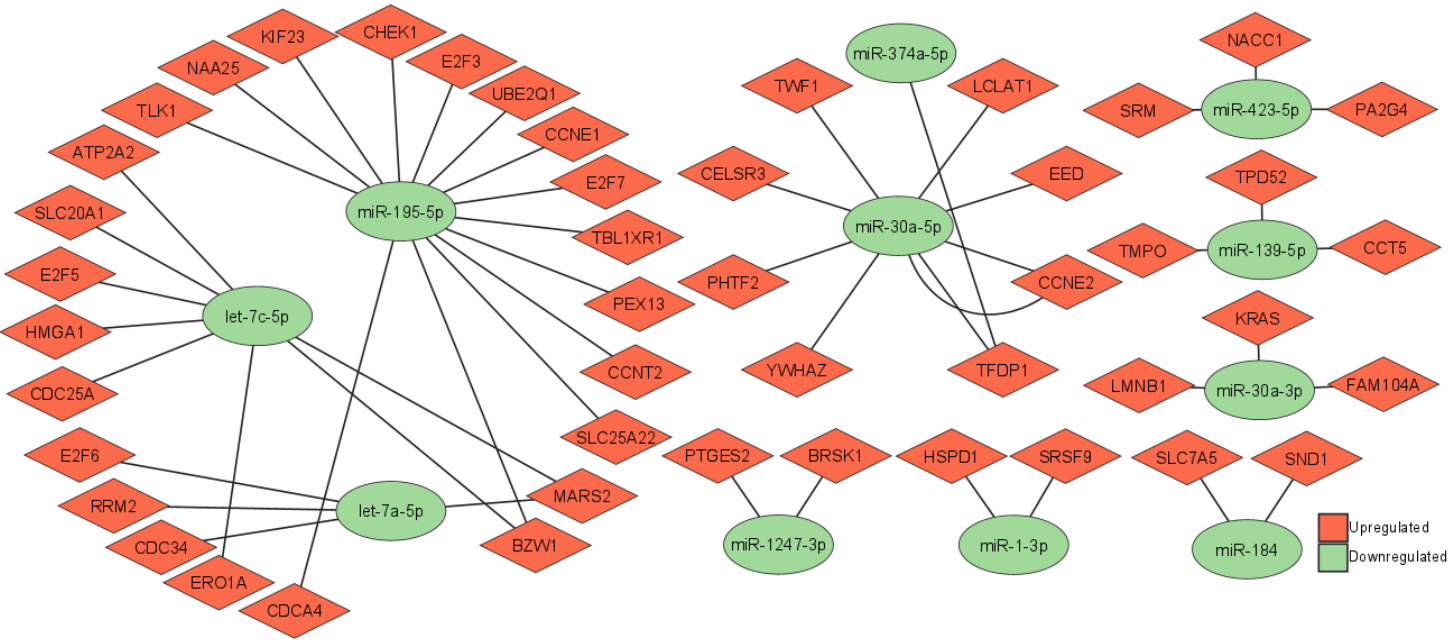


Figura 25 – Rede regulatória entre DE-miRNA *downregulated* e DEGs *upregulated* - LUAD.

Tabela 25 – DE-miRNAs *downregulated* e DEGs *upregulated* validados experimentalmente - LUAD.

ID	miRNA		Gene	
MIRT467590	miR-184	↓	SLC7A5	↑
MIRT502917	miR-195-5p	↓	CDCA4	↑
MIRT107246	miR-1247-3p	↓	PTGES2	↑
MIRT204743	miR-195-5p	↓	BZW1	↑
MIRT570465	miR-195-5p	↓	TLK1	↑
MIRT735620	let-7c-5p	↓	ERO1A	↑
MIRT472666	miR-195-5p	↓	NAA25	↑
MIRT735626	let-7c-5p	↓	MARS2	↑
MIRT214147	miR-30a-3p	↓	LMNB1	↑
MIRT512113	miR-221-5p	↓	CREBZF	↑
MIRT000408	let-7c-5p	↓	CDC25A	↑
MIRT028537	miR-30a-5p	↓	YWHAZ	↑
MIRT003976	miR-1-3p	↓	HSPD1	↑
MIRT506855	miR-195-5p	↓	KIF23	↑
MIRT265078	miR-195-5p	↓	CHEK1	↑
MIRT000223	miR-195-5p	↓	E2F3	↑
MIRT168581	let-7c-5p	↓	HMGA1	↑
MIRT736658	miR-139-5p	↓	TMPO	↑
MIRT733026	miR-30a-5p	↓	CCNE2	↑
MIRT736272	miR-30a-5p	↓	CCNE2	↑
MIRT171348	miR-30a-5p	↓	PHTF2	↑
MIRT006056	let-7a-5p	↓	CDC34	↑
MIRT574778	miR-30a-3p	↓	FAM104A	↑
MIRT005072	miR-1-3p	↓	SRSF9	↑
MIRT300035	let-7c-5p	↓	BZW1	↑
MIRT225969	let-7c-5p	↓	E2F5	↑
MIRT464747	miR-195-5p	↓	UBE2Q1	↑
MIRT297442	let-7c-5p	↓	SLC20A1	↑
MIRT735573	let-7a-5p	↓	MARS2	↑
MIRT562489	miR-30a-5p	↓	CELSR3	↑
MIRT007237	miR-195-5p	↓	CCNE1	↑
MIRT566992	miR-30a-3p	↓	KRAS	↑
MIRT502569	miR-195-5p	↓	E2F7	↑
MIRT726546	miR-195-5p	↓	SLC25A22	↑
MIRT731070	miR-184	↓	SND1	↑
MIRT727798	miR-30a-5p	↓	EED	↑
MIRT067980	miR-374a-5p	↓	TFDP1	↑
MIRT038148	miR-423-5p	↓	SRM	↑
MIRT538618	miR-139-5p	↓	CCT5	↑
MIRT038693	miR-30c-2-3p	↓	MARCKSL1	↑
MIRT052385	let-7a-5p	↓	RRM2	↑
MIRT051808	let-7c-5p	↓	ATP2A2	↑
MIRT731901	miR-139-5p	↓	TPD52	↑
MIRT005164	miR-30a-5p	↓	TWF1	↑
MIRT726354	miR-195-5p	↓	TBL1XR1	↑
MIRT067975	miR-30a-5p	↓	TFDP1	↑
MIRT159592	miR-195-5p	↓	PEX13	↑
MIRT085217	miR-195-5p	↓	CCNT2	↑
MIRT567716	let-7a-5p	↓	E2F6	↑
MIRT485950	miR-1247-3p	↓	BRSK1	↑
MIRT743262	miR-423-5p	↓	NACC1	↑
MIRT028546	miR-30a-5p	↓	LCLAT1	↑
MIRT038112	miR-423-5p	↓	PA2G4	↑

↑ *upregulated* e ↓ *downregulated*.

O enriquecimento das vias foi realizado para os DE-miRNAs e DEGs presentes na Tabela 25. A Tabela 26 apresenta as principais vias relacionadas a LUAD identificados para os DE-miRNAs, a Tabela S3 apresenta a lista completa, e 27 apresenta todas as vias identificadas com os DEGs, considerando valor $p < 0,05$.

Tabela 26 – Principais vias enriquecidas para DE-miRNAs *down* - LUAD.

<i>Term Name</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Ubiquitin mediated proteolysis	142	85	11	p<0,01
Cell cycle	129	78	12	p<0,01
Pathways in cancer	555	243	12	p<0,01
Proteoglycans in cancer	220	116	12	p<0,01
Autophagy - animal	146	84	11	p<0,01
Neurotrophin signaling pathway	124	74	12	p<0,01
Hippo signaling pathway	164	91	12	p<0,01
Protein processing in endoplasmic reticulum	194	103	12	p<0,01
Focal adhesion	213	109	12	p<0,01
FoxO signaling pathway	139	78	11	p<0,01
p53 signaling pathway	75	49	11	p<0,01
AMPK signaling pathway	130	71	11	p<0,01
TGF-beta signaling pathway	103	58	12	p<0,01
Prostate cancer	101	57	11	p<0,01
Colorectal cancer	88	51	11	p<0,01
Insulin signaling pathway	153	77	11	p<0,01
mTOR signaling pathway	177	86	11	p<0,01
MAPK signaling pathway	329	142	11	p<0,01
EGFR tyrosine kinase inhibitor resistance	82	47	11	p<0,01
Small cell lung cancer	100	54	12	p<0,01
HIF-1 signaling pathway	112	58	11	p<0,01
Chronic myeloid leukemia	79	43	11	p<0,01
PI3K-Akt signaling pathway	372	151	12	p<0,01
Apoptosis	151	69	12	p<0,01
Transcriptional misregulation in cancer	206	89	12	p<0,01
Wnt signaling pathway	173	76	11	p<0,01
Circadian rhythm	31	20	11	p<0,01
Central carbon metabolism in cancer	74	38	11	p<0,01
Bladder cancer	43	25	11	0,001
Non-small cell lung cancer	81	40	11	0,001
Endometrial cancer	61	31	11	0,002
Breast cancer	163	68	12	0,003
Thyroid cancer	43	23	11	0,004
VEGF signaling pathway	61	29	11	0,009

Tabela 27 – Vias enriquecidas para os DEGs *up* - LUAD.

<i>Term</i>	<i>Overlap</i>	<i>Adjusted p value</i>	<i>Genes</i>
Cell cycle	8/124	0,000	TGFBP1;CCNE2;CCNE1;CHEK1;E2F3;E2F5;YWHAZ;CDC25A
Cellular senescence	7/156	0,000	CCNE2;CCNE1;CHEK1;E2F3;KRAS;E2F5;CDC25A
Viral carcinogenesis	6/203	0,000	CCNE2;CCNE1;CHEK1;KRAS;YWHAZ;SND1
p53 signaling pathway	4/73	0,001	RRM2;CCNE2;CCNE1;CHEK1
Hepatitis B	5/162	0,001	CCNE2;CCNE1;E2F3;KRAS;YWHAZ
Prostate cancer	4/97	0,002	CCNE2;CCNE1;E2F3;KRAS
MicroRNAs in cancer	6/310	0,002	CCNE2;CCNE1;KIF23;E2F3;KRAS;CDC25A
Human T-cell leukemia virus 1 infection	5/219	0,003	CCNE2;CCNE1;CHEK1;E2F3;KRAS
Gastric cancer	4/149	0,006	CCNE2;CCNE1;E2F3;KRAS
Small cell lung cancer	3/92	0,018	CCNE2;CCNE1;E2F3
Oocyte meiosis	3/129	0,044	CCNE2;CCNE1;YWHAZ
Bladder cancer	2/41	0,047	E2F3;KRAS

Nas tabelas de vias identificadas para os DE-miRNAs e DEGs, independente da expressão,

foram identificados importantes vias relacionada ao LUAD, como: Proteoglicanos no câncer, adesão focal, via de sinalização p53, via de sinalização TGF-beta, câncer de pulmão de células pequenas, via de sinalização AMPK, via de sinalização MAPK, entre outras.

Para entender a relação funcional dos 81 DEGs (*ups* e *down*) e os fenômenos biológicos envolvidos, investigaram-se as interações funcionais das proteínas codificadas por esses genes. A rede de genes relacionada está apresentada na Figura 26, possui 63 nós e 106 arestas.

A rede apresenta a média do coeficiente de clusterização local = 0,421 e grau médio = 3,37. O valor de enriquecimento do PPI foi $p < 0,0001$, indicando que as proteínas estão biologicamente conectadas de forma significativa. A interação com maior pontuação combinada foi entre as proteínas TFDP1 e E2F5 (pontuação combinada = 0,999).

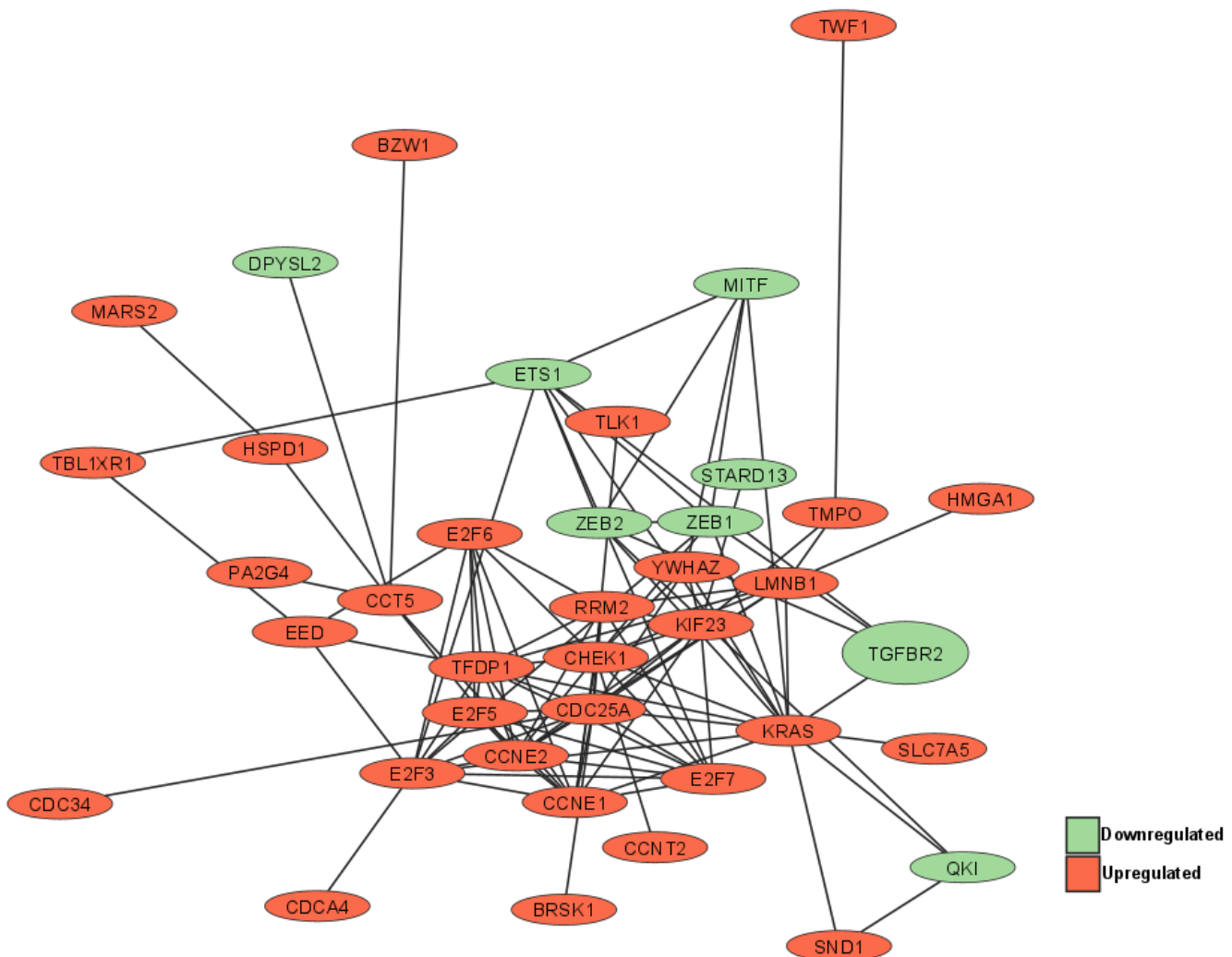


Figura 26 – Rede de interação proteína-proteína - LUAD.

Os processos biológicos (ontologia genética) da rede (Tabela 28) estão relacionados principalmente à regulação genética, ciclo celular e biossíntese de macromoléculas, todos os quais são fundamentais para o desenvolvimento e a manutenção das células.

Tabela 28 – Processos biológicos relacionados a Rede PPI - LUAD.

GO-termo	Descrição	Contagem na rede	Força	FDR
IR:0000083	Regulação da transcrição envolvida na transição G1/S do ciclo celular mitótico	3 de 21	1,65	0,03
IR:0000082	Transição G1/S do ciclo celular mitótico	9 de 86	1,51	0,00
IR:0044772	Transição de fase do ciclo celular mitótico	11 de 173	1,30	0,00
IR:0006260	Replicação de DNA	6 de 203	0,97	0,03
IR:1903047	Processo do ciclo celular mitótico	12 de 537	0,84	0,00
IR:1901990	Regulação da transição de fase do ciclo celular mitótico	7 de 332	0,82	0,04
IR:0051726	Regulação do ciclo celular	14 de 1108	0,60	0,02
IR:0000122	Regulação negativa da transcrição pela RNA polimerase II	12 de 971	0,59	0,03
IR:0045892	Regulação negativa da transcrição, modelada por DNA	15 de 1322	0,55	0,02
IR:0007049	Ciclo de célula	14 de 1246	0,55	0,02
IR:0051253	Regulação negativa do processo metabólico do RNA	16 de 1434	0,54	0,02
IR:0045934	Regulação negativa do processo metabólico composto contendo nucleobases	17 de 1562	0,53	0,01
IR:0010558	Regulação negativa do processo biossintético de macromoléculas	16 de 1532	0,51	0,02
IR:0031327	Regulação negativa do processo biossintético celular	16 de 1592	0,50	0,02
IR:0006357	Regulação da transcrição pela RNA polimerase II	21 de 2586	0,40	0,02
IR:0010556	Regulação do processo biossintético de macromoléculas	28 de 3980	0,34	0,02
IR:0051252	Regulação do processo metabólico do RNA	26 de 3759	0,34	0,02
IR:0006355	Regulação da transcrição, modelada por DNA	24 de 3460	0,34	0,04
IR:0010468	Regulação da expressão genética	33 de 4899	0,32	0,01
IR:0031326	Regulação do processo biossintético celular	28 de 4143	0,32	0,02
IR:0019219	Regulação do processo metabólico composto contendo nucleobases	27 de 4074	0,32	0,03
IR:0048523	Regulação negativa do processo celular	30 de 4736	0,30	0,02
IR:0048519	Regulação negativa do processo biológico	33 de 5313	0,29	0,02
IR:0051171	Regulação do processo metabólico do composto de nitrogênio	35 de 5734	0,28	0,02
IR:0031323	Regulação do processo metabólico celular	35 de 5681	0,28	0,02
IR:0060255	Regulação do processo metabólico de macromoléculas	37 de 6249	0,27	0,02
IR:0080090	Regulação do processo metabólico primário	35 de 5899	0,27	0,02
IR:0050794	Regulação do processo celular	50 de 11025	0,15	0,04

Análises estatísticas

Foram criadas árvores de classificação, Florestas Aleatórias e árvores de sobrevivência para identificar potenciais DE-miRNAs e DEGs importantes na separação de T vs. N e na sobrevivência dos pacientes com LUAD precoce.

Os bancos de dados de miRNA-seq e RNA-seq ($n = 274$) foram filtrados pelos 28 DE-miRNAs e 63 DEGs (sem repetições) validados experimentalmente. Foram analisados separadamente, independente da direção da expressão e considerando as variáveis clínicas significativas para cada técnica (diagnóstico primário, malignidade anterior e tamanho do tumor) identificadas na Tabela 16.

Para construir a árvore de classificação e Florestas Aleatórias (construído a fim de complementar os resultados obtidos na árvore) em relação ao T vs. N, o banco de dados foi dividida aleatoriamente em 80% para treino ($n = 219$) e 20% para teste ($n = 55$), a fim de comparar o desempenho do banco de treino em novos dados (banco de teste). Para corrigir o problema de desbalanceamento no banco de treino, realizou-se subamostragem.

Árvore de classificação

Foram simuladas 100 mil árvores de classificação, tanto para o conjunto de DE-miRNAs quanto para de DEGs, a fim de identificar a árvore de classificação com maiores acurácias na

separação de T vs. N. Para essas árvores consideraram-se os hiperparâmetros: $minsplit=2$, $minbucket=1$ e $cp=0,00001$.

A Figura 27 mostra as acurácias dessas simulações para o conjunto de DE-miRNAs. Observam-se variações de 5 – 100% de acurácia, a média próxima a 80% e destaca-se que 3201 continham acurácia de 95 – 100%.

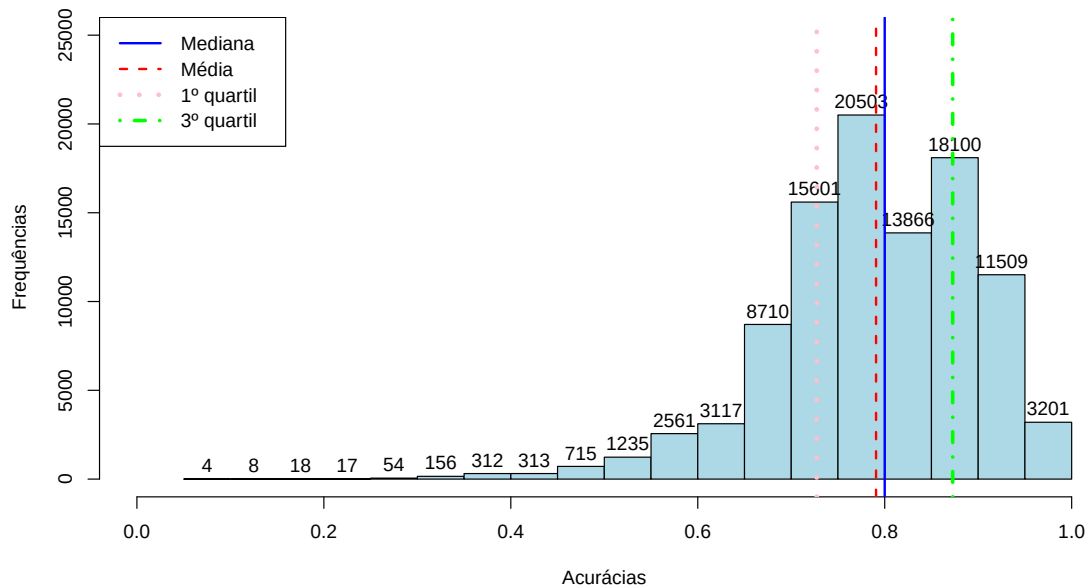


Figura 27 – Histograma das 100 mil acurácias para árvores de classificação dos DE-miRNAs - LUAD.

A Figura 28a apresenta a árvore de classificação de DE-miRNAs com 100% de acurácia ao testar no banco de teste. A árvore contém o miR-9-5p e miR-1247-3p e funciona da seguinte forma: todas as observações estão no primeiro nó (miR-9-5p), as que satisfazem a condição de $miR-9-5p \geq 602,5$ seguem para a direita e são classificadas como tumor primário (azul). As observações que não satisfazem seguem para a esquerda e se $miR-1247-3p \geq 63,5$ são classificados como tecido normal (cinza) e, caso contrário, como tumor.

Os 9 DE-miRNAs mais importantes para o modelo de árvore de classificação estão apresentados na Figura 28b, os com maior importância são o miR-9-5p seguido do miR-200a-3p.

A Figura 29 mostra as acurácias obtidas para as simulações das 100 mil árvores feitas com os DEGs, observando-se variações de 5 – 100%. Destaca-se que o 1º quartil está acima de 70% e há 5264 árvores entre 95 – 100%.

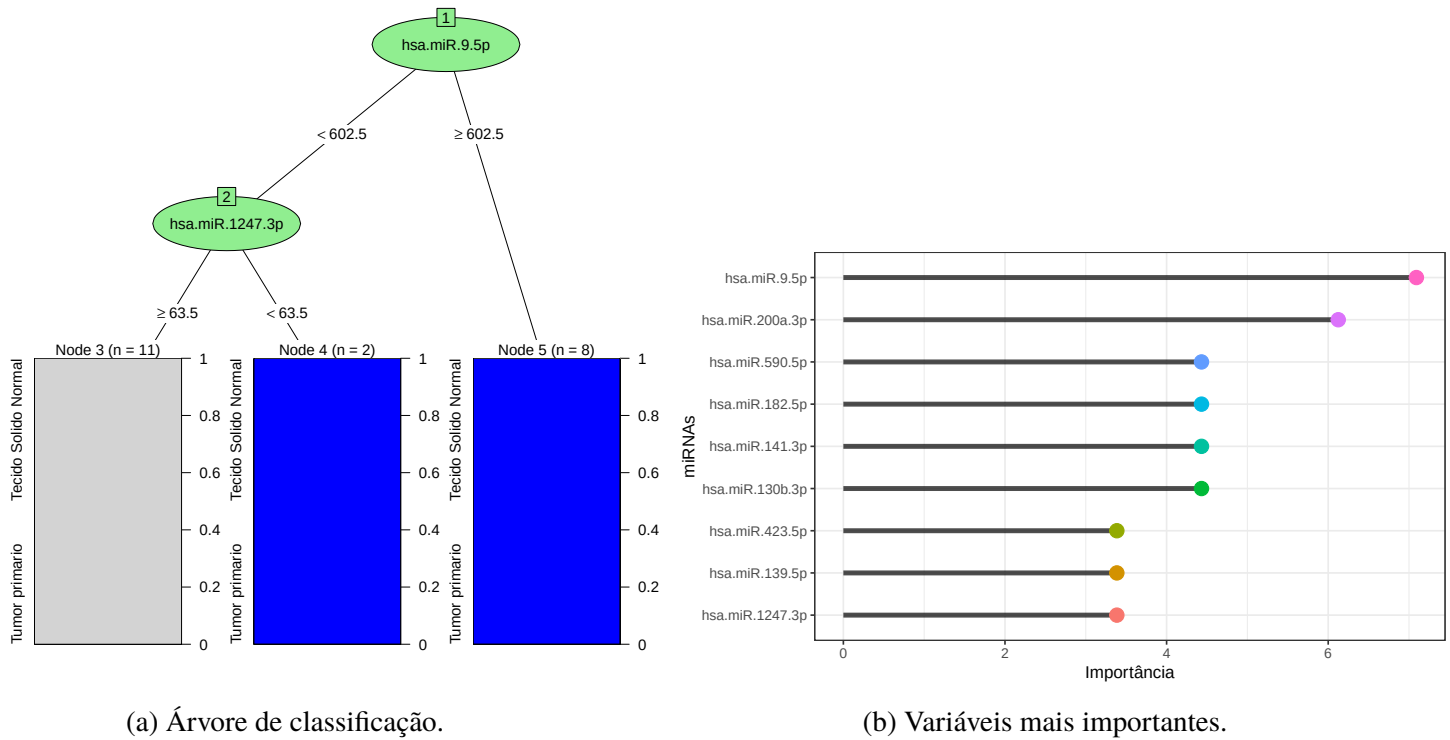
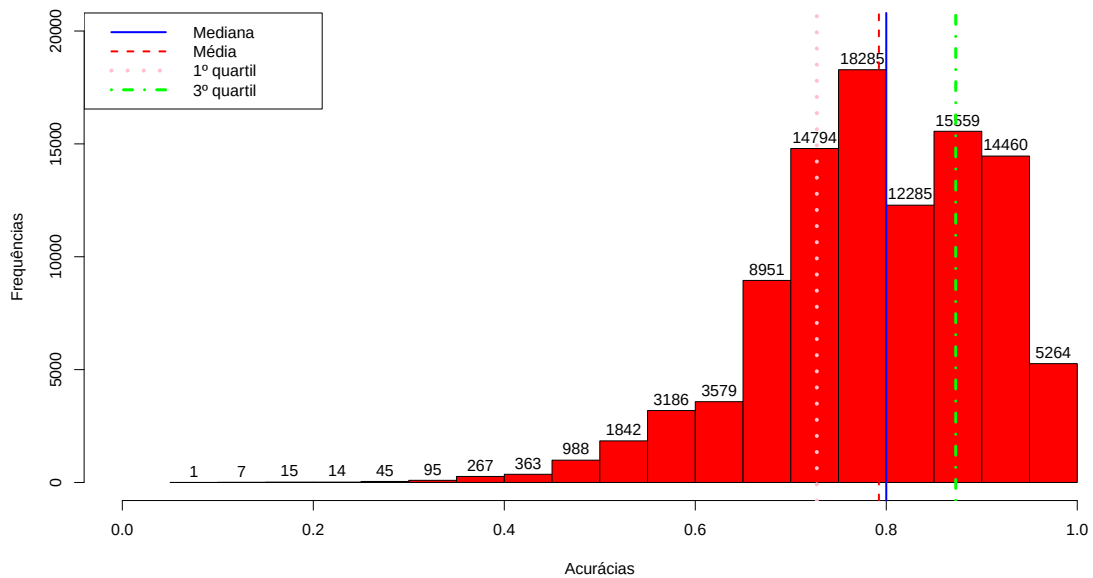


Figura 28 – Árvore de classificação e variáveis mais importantes para os DE-miRNAs- LUAD.



A árvore de classificação com 100% de acurácia formado pelos DEGs está apresentada na Figura 30a. A árvore é composta por *FHL1* (*Four and a Half LIM domains 1*) e *BZW1* (*Bent out of shape homolog 1*). Todas as amostras estão no primeiro nó, se *FHL1* < 5417,5 as observações seguem para a direita e são classificados como tumor primário (azul), caso contrário, seguem para

o proximo nó. Por fim, se se $BZW1 < 7041$ é classificado como tecido normal e caso contrário como tumor.

A Figura 30b apresenta os DEGs mais importantes para a construção do modelo de árvore de classificação na separação dos tecidos (T vs. N), destaca-se *FHL1* como os mais importantes e depois *TGFBR2*, *STARD13* e *ID4* com a mesma importância.

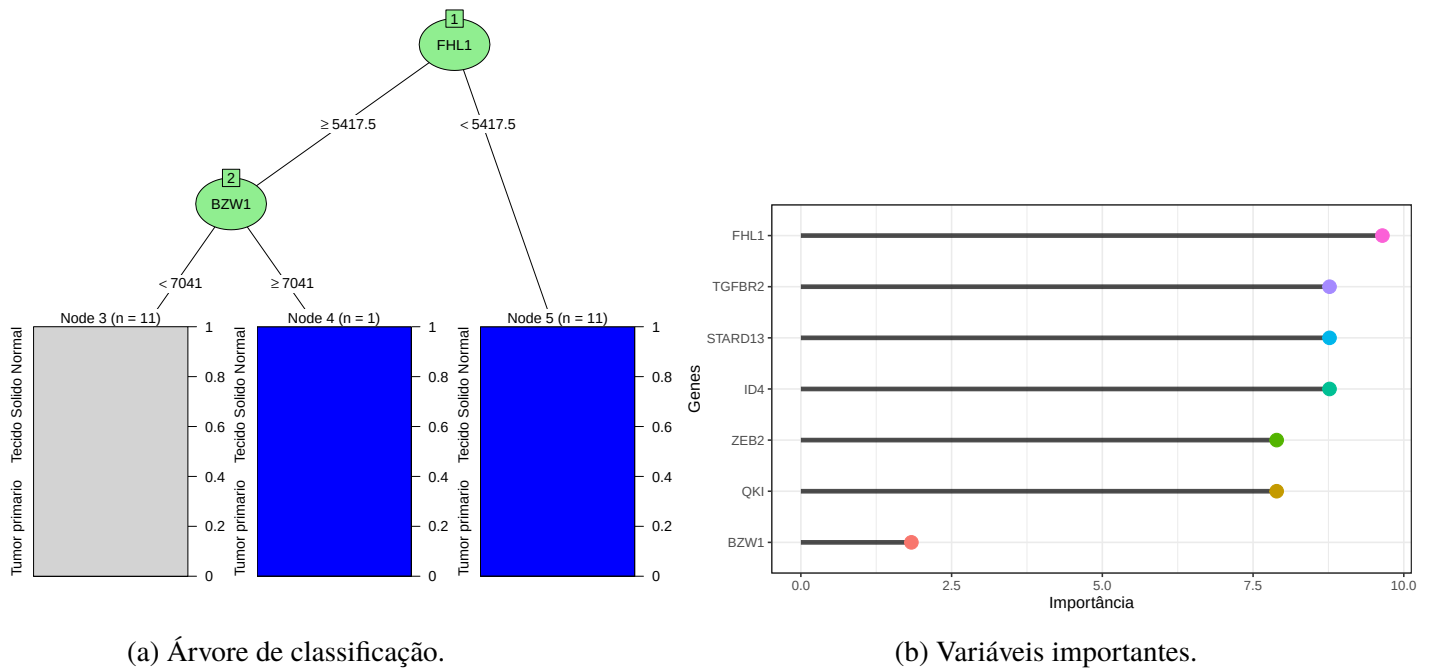


Figura 30 – Árvore de classificação e variáveis importantes para os DEGs - LUAD.

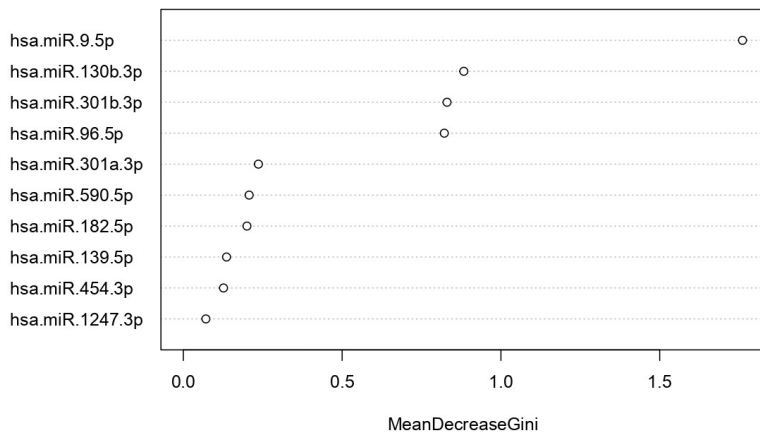
Florestas Aleatórias

O modelo de Florestas Aleatórias também foi realizado com os DE-miRNAs e DEGs para a separação de T vs. N. Foram geradas 5.000 árvores de classificação, o modelo identificou os 10 DE-miRNAs e 10 DEGs que são mais importantes para a diferenciação dos tecidos, ilustradas nas Figuras 31a e 32a.

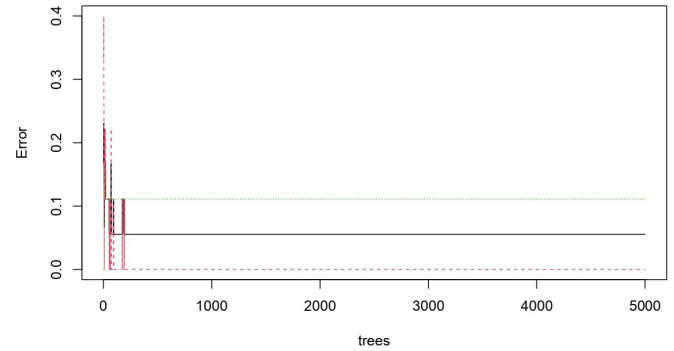
Os DE-miRNAs com maior importância, que obtiveram maior a medida *mean decrease gini* maior foram: miR-9-5p, miR-130b-3p, miR-301b-3p e miR-96-5p. Em comparação com árvore de classificação, o miR-9-5p continuou sendo o miRNA mais importante na separação de T vs.N e o miR-1247-3p também foi identificado com o modelo de Florestas Aleatórias.

Os 3 DEGs que obtiveram maiores importâncias e se destacaram com maiores *mean decrease gini* foram o *TGFBR2* e *FHL1*. Além do *FHL1*, que apareceu do modelo de árvore de classificação o *ZEB1* também foi identificado em Florestas Aleatórias.

Os erros dos modelos estão ilustrados nas Figuras 31b e 32b. Em vermelho está o tumor primário e verde tecido normal observa-se que os erros variam de 0 a 0.4 e estabilizando com menos de 500 árvores.

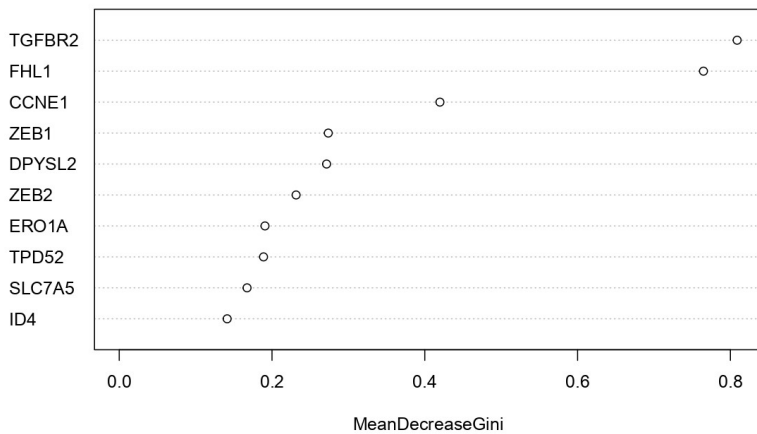


(a) DE-miRNAs importantes.

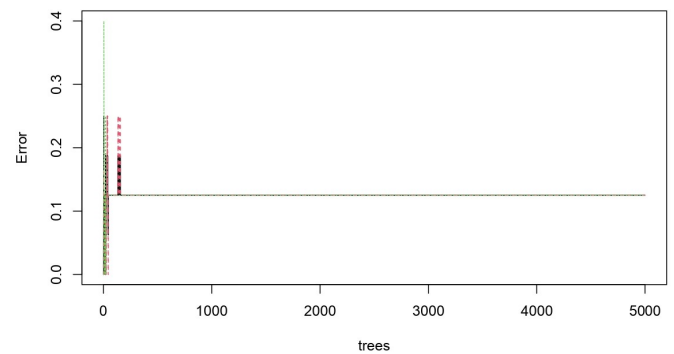


(b) Taxas de erros.

Figura 31 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DE-miRNA - LUAD.



(a) DEGs importantes.



(b) Taxas de erros.

Figura 32 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DEGs - LUAD.

Árvore de sobrevivência

Árvores de sobrevivência foram construídas com o objetivo de identificar os DE-miRNAs e DEGs relacionados com a sobrevivência de pacientes em estágio inicial de LUAD. A variável tempo utilizada para a construção da árvore 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 LUAD, status = 1 (composto por 66 casos), e caso o contrário, status = 0 (composto por 208 casos).

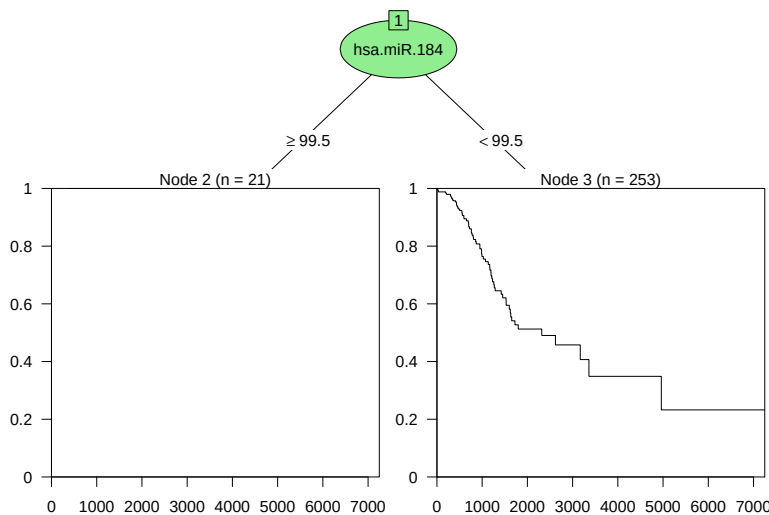
A árvore de sobrevivência gerada para o conjunto de DE-miRNAs ficou complexa com 21 nós, como pode ser observado na coluna *nsplit* na Tabela 29. Diante dessa complexidade, foi necessário realizar a pós-poda da árvore. Para tal, identificou-se o menor erro de validação cruzada, resultando em um valor para realizar a poda de $cp = 0,025$. A árvore de sobrevivência pós-poda contém apenas um nó e está ilustrada na Figura 33a.

Tabela 29 – Resumo do modelo da árvore de sobrevivência para os DE-miRNAs - LUAD.

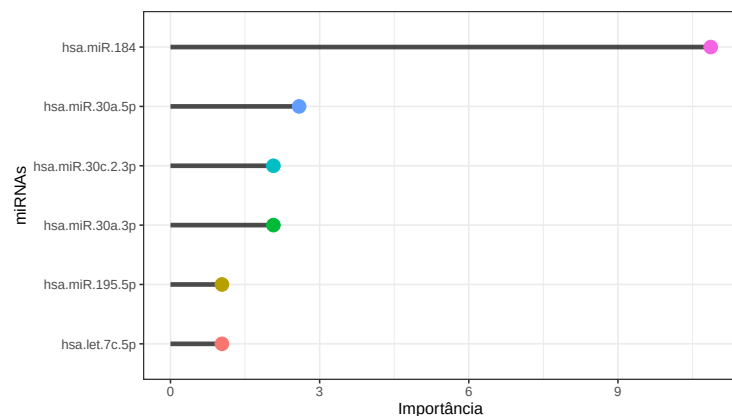
CP	Nº de divisões	Erro relativo	Erro de validação cruzada	Erro padrão
0,037	0	1,000	1,004	0,090
0,025	1	0,963	1,223	0,122
0,024	10	0,717	1,319	0,128
0,023	12	0,669	1,329	0,129
0,021	14	0,624	1,327	0,131
0,020	15	0,602	1,299	0,130
0,018	16	0,583	1,310	0,132
0,013	17	0,565	1,322	0,132
0,012	18	0,552	1,342	0,134
0,011	19	0,540	1,331	0,134
0,010	20	0,529	1,331	0,134
0,010	21	0,518	1,332	0,133

Todas as observações estão presentes no nó inicial da (Figura 33a), se $\text{miR-184} \geq 99.5$ as observações possuem maior chance de sobrevivência e, caso contrário, possuem maior risco de óbito com o decorrer dos dias, observa-se que após 2000 dias as chances estão próximas a 50%.

A Figura 33b mostra os miRNAs mais importantes para a sobrevivência dos pacientes, observa-se claramente o destaque do miR-184 em relação aos demais que também são importantes, como o miR-30a-5p, miR-30c-2-9p e miR-30a-3p.



(a) Árvore de sobrevivência pós-poda.



(b) Variáveis importantes.

Figura 33 – Árvore de sobrevivência pós-poda e variáveis mais importantes para dos DE-miRNAs - LUAD.

A árvore de sobrevivência construída para o conjunto de DEGs demonstrou uma complexidade com 15 nós, como mostra a Tabela 30. Diante dessa complexidade, foi necessário podar a árvore. Para isso, identificou-se o menor erro de validação cruzada, resultando em $cp = 0,041$

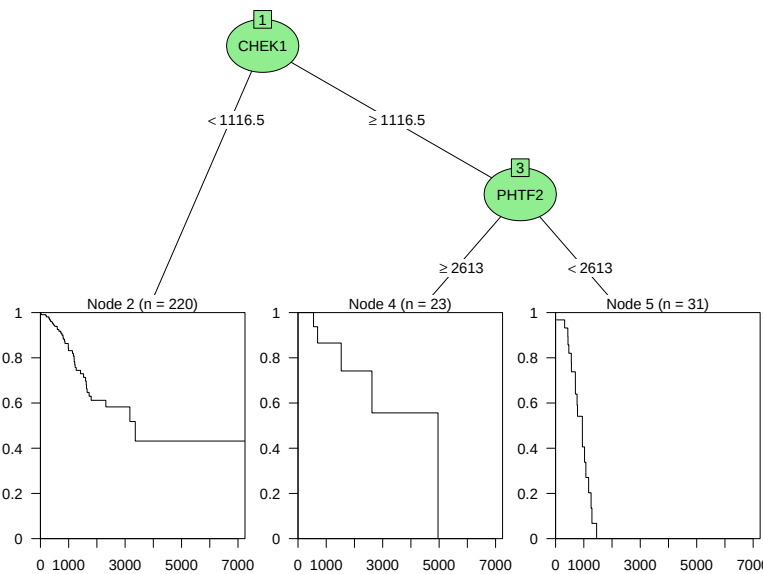
para realizar a poda. A árvore de sobrevivência pós-poda gerada é formada por 2 nós e está ilustrada na Figura 34a.

Tabela 30 – Resumo do modelo da árvore de sobrevivência para os DEGs - LUAD.

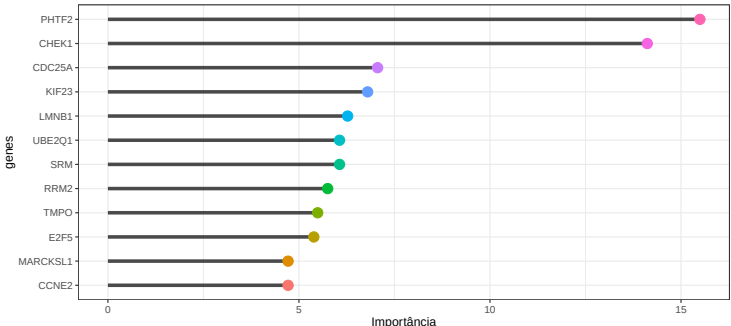
CP	Nº de divisões	Erro relativo	Erro de validação cruzada	Erro padrão
0,058	0	1,000	1,012	0,090
0,041	2	0,883	1,054	0,098
0,039	4	0,800	1,148	0,114
0,038	5	0,761	1,154	0,114
0,036	7	0,686	1,166	0,114
0,033	8	0,650	1,167	0,115
0,032	9	0,617	1,167	0,115
0,025	10	0,585	1,156	0,122
0,020	11	0,560	1,179	0,126
0,019	14	0,499	1,185	0,126
0,010	15	0,480	1,233	0,135

A árvore pós-poda (Figura 34a), contém os DEGs: *CHEK1* e *PHTF2*. As observações que tiverem *CJEK1* < 1116,5 seguem para a esquerda e possuem maior chance de sobrevivência e, caso contrário, vão para a direita. Se as observações tiverem *PHTF2* < 2613 é o pior cenário e os pacientes possuem aproximadamente 1500 dias de sobrevivência (nó 5), caso contrário, vão para a esquerda (nó 4) e possuem até 5000 dias de tempo de sobrevivência.

A Figura 34b apresenta os DEGs mais importantes relacionados a sobrevivência dos pacientes, o destaque está no *PHTF2* e *CHEK1*.



(a) Árvore de sobrevivência pós-poda.



(b) Variáveis importantes.

Figura 34 – Árvore de sobrevivência pós-poda e variáveis mais importantes para os DEGs - LUAD.

Discussões

Após realizar análises descritivas, biológicas e estatísticas nos bancos de dados de miRNA-seq e RNA-seq, foi possível examinar e identificar padrões e comportamentos relevantes. Todas as análises foram conduzidas com o objetivo inicial de compreender o perfil das amostras de pacientes presentes nos bancos de dados. Em seguida, a atenção foi direcionada para identificar DE-miRNAs e DEGs, que desempenham um papel crucial na diferenciação entre tecido T vs. N e na sobrevivência de pacientes com LUAD em estágios precoces.

Os testes estatísticos apresentaram diferenças significativas, considerando com $p < 0,05$, na diferenciação de T vs. N, nas variáveis diagnóstico primário e malignidade anterior para o banco de dados de miRNA-seq. Já para o banco de RNA-seq foi identificado que o tamanho do tumor apresentou diferenças estatisticamente significativas.

A decisão de empregar o teste de Wilcoxon-Mann-Whitney, juntamente com a correção de Bonferroni, para conduzir a análise de expressão diferencial foi motivada por uma abordagem que visa preservar a escala original dos dados e analisar as variáveis individualmente. Ao optar por métodos estatísticos não paramétricos, almejou-se proporcionar interpretações biológicas mais diretas, ao mesmo tempo em que se procurava ampliar a quantidade de DE-miRNAs e DEGs identificados. Essa estratégia não apenas mantém a fidelidade dos dados originais, mas também contribui para a robustez e interpretabilidade dos resultados, permitindo uma compreensão mais profunda das variações biológicas subjacentes. A correção de Bonferroni, por sua vez, desempenha um papel crucial na mitigação do erro tipo I, assegurando a confiabilidade estatística das descobertas. Com esta abordagem, busca-se não apenas quantidade, mas qualidade nas identificações de DE-miRNAs e DEGs, contribuindo para uma análise mais abrangente e significativa no contexto da diferenciação entre T vs. N. Assim, foi possível identificar 254 DE-miRNAs (24% do total) e 7.374 DEGs (18% do total).

Após a identificação dos DE-miRNAs e DEGs, as análises biológicas levaram a identificação de 81 pares potenciais (DE-miRNA - DEG candidatos) que possuem expressões contrárias, fortes correlações negativas e são validados experimentalmente.

Ao analisar o conjunto dos 28 pares de DE-miRNAs *up* e DEGs *down* a rede regulatória (Figura 24) destacou o *TGFBR2*. Este gene sinaliza através da regulação do crescimento celular, diferenciação, apoptose, invasão, angiogênese e resposta imune (DAVID; MASSAGUÉ, 2018). Sardo *et al.* (2021) destacam que este gene codifica *Transforming Growth Factor- β* (*TGF- β*) receptor 2, um membro da sinalização *TGF- β* , que está envolvido no desenvolvimento embrionário, homeostase tecidual e tumorigênese.

Chen *et al.* (2017b) concluem em seu trabalho que o miR-301a promove a proliferação e migração de células NSCLC, visando o *TGFBR2* (*downregulated*) para modular a sinalização do *TGF- β* . Na rede criada neste trabalho, o *TGFBR2* está ligado ao miR-301a-3p. Hirono *et al.* (2019) destacam que o miR-130b contém alta expressão e pode funcionar como um miRNA

oncogênico no NSCLC.

Vale destacar o gene *QKI* na rede de regulação [Zong et al. \(2014\)](#) relatam que este gene desempenha um papel de regulação no processo de *splicing*¹ em casos de câncer de pulmão. Essa descoberta destaca o papel potencialmente relevante do QKI na modulação de eventos moleculares associados à formação do RNA mensageiro em células cancerígenas do pulmão.

Outro importante miRNA identificado foi o miR-96-5p que é considerado um oncomiR, ou seja, é um miRNA que está associado ao câncer, que facilita o desenvolvimento de doenças malignas, promovendo o crescimento, proliferação e sobrevivência de células cancerígenas ([RESS et al., 2015](#)). [Chen et al. \(2015\)](#) concluem que o miR-96-5p promoveu comportamentos malignos nas células, visando o *FHL1* (*Four and a Half LIM Domains 1*).

Além desses importantes resultados, foram encontradas estatisticamente significativas vias enriquecidas relacionadas ao LUAD, identificadas com os DE-miRNAs e DEGs, como: via de ciclo celular, caminhos no câncer, Proteoglicanos no Câncer, Via de sinalização p53, Resistência a Inibidores de Tirosina Quinase do EGFR, Via de Sinalização MAPK, Via de Sinalização ErbB, Via de Sinalização PI3K-Akt, Desregulação Transcricional no Câncer, Via de Sinalização Wnt, Carcinogênese Viral, Senescência celular, entre outras.

Ao analisar o conjunto dos 53 DE-miRNAs *down* e DEGs *up*, a rede regulatória (Figura 25) identificou importantes DE-miRNAs relacionadas a DEGs. O miR-196-5p está relacionado a 14 DEGs [Li et al. \(2020\)](#) relatam que esse miRNA pode atuar como um biomarcador que contribui para o diagnóstico do câncer de pulmão e a detecção de sua população de alto risco. Por sua vez o miR-195-5p é reconhecido como um supressor tumoral ativado no ciclo celular, apoptose e proliferação durante a progressão de diversas doenças cancerígenas ([SUN et al., 2017](#)).

O let-7c-5p aparece relacionado a 8 DEGs, incluindo o CDC25A, [Huang et al. \(2023\)](#) relatam nos seus resultados que let-7c-5p/CDC25A é um alvo terapêutico subjacente para LUAD, fornecendo uma estratégia viável para superar a resistência à cisplatina, substância amplamente utilizada no tratamento de diversos tipos de câncer, incluindo câncer de pulmão.

O miR-30a-5p também aparece em destaque, relacionado a 8 miRs, o trabalho de [Kanthaje et al. \(2021\)](#) traz uma vasta revisão, apresentando uma compreensão mais profunda do papel regulador e do significado clínico do miR-30 e enfatizamos os papéis emergentes no câncer de pulmão. [Świtlik et al. \(2018\)](#) concluem que miR-30a-5p junto com miR-210-3p possui aceitável sensibilidade e especificidade para distinguir tecido canceroso de tecido não canceroso e, portanto, podem se tornar um potencial biomarcador de diagnóstico para NSCLC.

As vias enriquecidas identificadas incluem aquelas que são comuns e já foram relatadas anteriormente, juntamente com outras vias de grande relevância para o LUAD. Destacam-se a Via de Sinalização das Neurotrofinas, o Processamento de Proteínas no Retículo Endoplasmático,

¹ Processo de remoção de segmentos não codificadores (íntrons) do RNA precursor e à união dos segmentos codificadores (éxons), formando assim o RNA mensageiro funcional.

a Adesão Focal, a Via de Sinalização Hedgehog, além de vias relacionadas à neurodegeneração, a Via de Sinalização AMPK, a Via de Sinalização da Insulina, o Câncer de Pulmão de Células Pequenas, entre outras.

A rede de PPI (Figura 26) formada por todos os 81 DEGs identificados no pares, mostrou valor de enriquecimento significativo, $p < 0,0001$, afirmando que as proteínas identificadas pelos DEGs estão biologicamente conectadas de forma significativa. Na rede o *TGFBR2* apareceu conectado, por exemplo, ao *KRAS* (*Kirsten rat sarcoma viral oncogene homolog*) que é um proto-oncogene, que podem levar à ativação descontrolada da via de sinalização RAS, promovendo o crescimento tumoral.

As análises estatísticas foram realizadas para os conjuntos de DE-miRNAs e DEGs que foram identificados validados experimentalmente, 28 DE-miRNAs e 63 DEGs (sem repetições), independentes da direção da expressão.

O miR-9-5p teve destaque na árvore de classificação e Florestas Aleatórias como o mais importante na diferenciação de T vs. N. Este miRNA já havia aparecido na rede regulatória da Figura 24.

Em relação à sobrevivência dos pacientes, o miR-184 foi o mais importante relacionado com LUAD precoce. Lin *et al.* (2015) concluem em seu trabalho que miR-184 é um supressor de tumor que inibe a proliferação celular e a capacidade de invasão através do direcionamento de CDC25A e c-Myc e que baixo nível de miR-184 pode prever pior prognóstico em pacientes com NSCLC, neste trabalho o miR-184 está *downregulated*.

Em relação aos DEGs, foram identificados na árvore de classificação o FHL1 e BZW1 diferenciaram os tecidos T vs. N, esses DEGs também foram identificados em Florestas Aleatórias. O *TGBR2* apareceu como primeiro mais importante para diferenciar T vs. N em Florestas Aleatórias e na árvore de classificação fio na segunda posição.

Para a sobrevivência dos pacientes, CHECK1 e PHTF2 foram os mais importantes. O trabalho de Wu *et al.* (2020) relata que a análise de Kaplan-Meier demonstrou que os pacientes com alta expressão de PHTF2 tiveram menor sobrevida, na árvore de sobrevivência (Figura 34a) o PHTF2 pode acarretar menores chances de vida também. O trabalho também destaca que PHTF2 foi ainda identificado como um alvo direto do miR-30a-5p em células NSCLC e abrigou uma correlação negativa com a expressão de miR-30a-5p em tecidos NSCLC, concordando com os achados da Tabela S1.

Por fim, as análises realizadas nos bancos de dados de miRNA-seq e RNA-seq em estágios iniciais de LUAD utilizando técnicas de bioinformática possibilitaram a identificação de perfis moleculares específicos, ajudaram a compreender a expressão gênica e a regulação dos miRNAs, fornecendo *insights* valiosos para diagnóstico precoce, prognóstico e desenvolvimento de estratégias terapêuticas personalizadas.

4.2.2 Análise de expressão diferencial, biológica e estatística de LUSC

Nesta parte estão os resultados encontrados e discussões para os bancos de dados LUSC miRNA-seq e RNA-seq. O banco de dados de miRNA-seq é formado por 249 amostras de pacientes divididos em dois grupos: 23 N e 226 T, sendo que 23 pacientes possuem amostras tanto de N quanto de T, e 1.061 expressões de miRNA-seq. O banco de dados de RNA-seq é composto por 266 amostras de pacientes também dividido em dois grupos: 25 N e 241 T, sendo que 22 amostras são diferentes do miRNA-seq, e 40.919 expressões de RNA-seq. Ambos os bancos de dados compartilham as mesmas 12 variáveis clínicas, que englobam características demográficas, diagnósticas e informações sobre o tratamento dos pacientes (definições na Tabela 2).

A Tabela 31 apresenta as particularidades dos pacientes presentes nos dados. Pode-se destacar, para ambos os bancos de dados, a predominância do sexo masculino, raça branca e estado vital vivo. O estágio IB ocorreu com mais frequência com o tamanho do tumor T2, o diagnóstico primário de carcinoma de células escamosas predominou e o lobo superior originou mais o câncer. A maioria dos pacientes não possuíam malignidade anterior e não realizavam tratamento ou terapia.

A média das idades dos pacientes com amostras nos tecidos T e N estão entre 68 a 70 anos e os dias para o último acompanhamento estão variando de 827 a 1067 dias, as Figuras 35 e 36 ilustram o comportamento das idades e dos dias do último acompanhamento. Nenhuma variável foi estatisticamente significativa, considerando valor $p < 0,05$, para a separação do tipo de tecido.

Tabela 31 – Características demográficas, diagnóstico e tratamento das amostras dos pacientes dos bancos de dados LUSC miRNA-seq e RNA-seq.

Variáveis	miRNA-seq				Valor p	RNA-seq				Valor p
	T (n=226)		N (n=23)			T (n=241)		N (n=25)		
	n	%	n	%		n	%	n	%	
Sexo										
Feminino	65	89,04%	8	10,96%	0,55	68	90,67%	7	9,33%	0,98*
Masculino	161	91,48%	15	8,52%		173	90,58%	18	9,42%	
Raça										
Branco	164	89,62%	19	10,38%	0,48	176	89,80%	20	10,20%	0,75**
Negro ou afro-americano	15	100,00%	0	0,00%		15	88,24%	2	11,76%	
Asiático	2	100,00%	0	0,00%		2	100,00%	0	0,00%	
Estado Vital										
Vivo	139	92,05%	12	7,95%	0,38	143	93,46%	10	6,54%	0,06*
Morto	87	88,78%	11	11,22%		98	86,73%	15	13,27%	
Estadiamento										
IB	140	88,61%	18	11,39%	0,38	148	89,16%	18	10,84%	0,55**
IA	83	94,32%	5	5,68%		90	92,78%	7	7,22%	
I	3	100,00%	0	0,00%		3	100,00%	0	0,00%	
Tamanho do tumor (T)										
T2	88	91,67%	8	8,33%	0,26	93	90,29%	10	9,71%	0,85**
T2a	52	83,87%	10	16,13%		55	87,30%	8	12,70%	
T1	36	92,31%	3	7,69%		42	93,33%	3	6,67%	
T1b	30	93,75%	2	6,25%		31	91,18%	3	8,82%	
T1a	20	100,00%	0	0,00%		20	95,24%	1	4,76%	
Diagnóstico primário										
Carcinoma de células escamosas	210	91,30%	20	8,70%	0,29	224	91,30%	22	8,70%	0,33**
Carcinoma basoide espinocelular	6	85,71%	1	14,29%		7	85,71%	1	14,29%	
Outros	10	83,33%	2	16,67%		10	83,33%	2	16,67%	
Tecido ou órgão de origem										
Lobo superior	127	91,37%	12	8,63%	0,87	133	90,48%	14	9,52%	0,81**
Lobo inferior	85	89,47%	10	10,53%		87	89,69%	10	10,31%	
Outros	14	93,33%	1	6,67%		21	95,45%	1	4,55%	
Malignidade anterior										
Não	195	62,00%	19	38,00%	0,54	207	90,79%	21	9,21%	0,77**
Sim	31	97,99%	4	2,01%		34	89,47%	4	10,53%	
Tratamento ou terapia										
Não	175	89,29%	21	10,71%	0,71	184	89,76%	21	10,24%	1,00**
Sim	21	95,45%	1	4,55%		24	92,31%	2	7,69%	
Idade										
Média	68,40		70,00		0,60	68,20		70,20		0,44•
Mediana	69		68			69		68		
Dias para o último acompanhamento										
Média	1058,00		827,00		0,18	1067,00		908,00		0,31•
Mediana	700		500			700		562		

*Teste qui-quadrado. **Teste exato de Fisher. •Teste de Wilcoxon-Mann-Whitney.

Legenda: A categoria outros^o contém: carcinoma espinocelular papilífero, carcinoma espinocelular de pequenas células não queratinizante, carcinoma de células escamosas- grandes células- não queratinizante e carcinoma de células escamosas- queratinizante. A categoria outros⁺ contém: brônquio principal, lobo médio e pulmão.

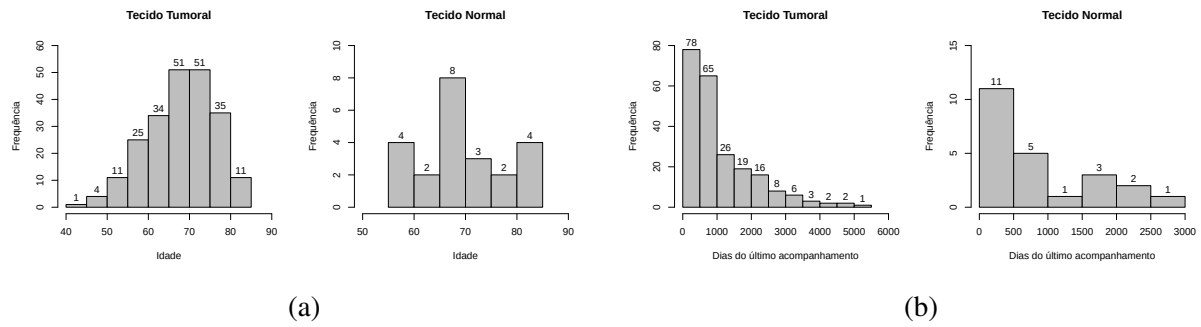


Figura 35 – Histograma da idade e dos dias do último acompanhamento - LUSC miRNA-seq.

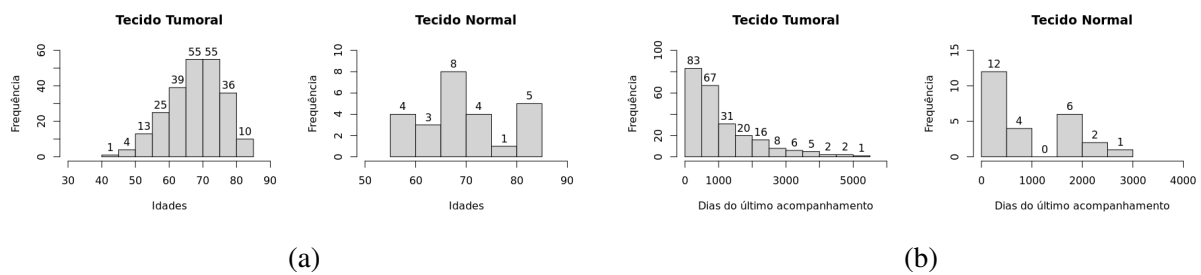


Figura 36 – Histograma da idade e do número de dias do último acompanhamento - LUSC RNA-seq.

Análise de Expressão diferencial

Foi conduzida a análise de expressão diferencial para cada miRNA e RNA, avaliados de forma independente, visando identificar DE-miRNAs e DEGs na diferenciação entre T vs. N.

Realizou-se o teste de Wilcoxon-Mann-Whitney nas 1.061 expressões de miRNA-seq, considerando $p < 0,05$ como estatisticamente significativo. Posteriormente, a correção de Bonferroni foi feita, para isso, adotou-se $\alpha_{corrigido}$:

$$\alpha_{corrigido} = \frac{\alpha_{original}}{n} = \frac{0,05}{1.061} \approx 0,000047.$$

Considerando $p < 0,000047$, identificaram-se 337 DE-miRNAs, sendo 272 *downregulated* e 65 *upregulated* e, ambos em relação à mediana do N e apresentados na Tabela 32 e 33, respectivamente, conforme valores p mais baixos.

Tabela 32 – DE-miRNAs *downregulated* (↓) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUSC.

miRNAs	Mediana		valor <i>p</i>	miRNAs	Mediana		valor <i>p</i>	miRNAs	Mediana		valor <i>p</i>
	N	T			N	T			N	T	
miR-4732-3p	↓ 12,0	0,0	p< 0,000047	miR-374b-5p	↓ 493,0	189,0	p< 0,000047	let-7g-5p	↓ 2988,0	1348,5	p< 0,000047
miR-30a-5p	↓ 542643,0	33284,5	p< 0,000047	miR-30e-3p	↓ 44027,0	13793,5	p< 0,000047	miR-378d	↓ 5,0	1,0	p< 0,000047
miR-30c-2-3p	↓ 1062,0	61,5	p< 0,000047	miR-10a-5p	↓ 255637,0	62832,0	p< 0,000047	miR-330-5p	↓ 214,0	78,5	p< 0,000047
miR-30a-3p	↓ 121349,0	8402,0	p< 0,000047	miR-140-5p	↓ 276,0	97,0	p< 0,000047	miR-6716-3p	↓ 4,0	2,0	p< 0,000047
miR-338-5p	↓ 233,0	12,0	p< 0,000047	miR-223-5p	↓ 40,0	11,0	p< 0,000047	miR-607	↓ 2,0	0,0	p< 0,000047
miR-30d-5p	↓ 178534,0	16821,5	p< 0,000047	miR-3065-5p	↓ 62,0	18,0	p< 0,000047	miR-219a-2-3p	↓ 1,0	0,0	p< 0,000047
miR-490-3p	↓ 28,0	0,0	p< 0,000047	miR-3920	↓ 4,0	1,0	p< 0,000047	miR-6761-5p	↓ 7,0	3,0	p< 0,000047
miR-326	↓ 180,0	11,0	p< 0,000047	miR-125b-5p	↓ 5122,0	1447,0	p< 0,000047	miR-2355-5p	↓ 422,0	194,5	p< 0,000047
miR-139-3p	↓ 242,0	12,5	p< 0,000047	miR-1468-5p	↓ 48,0	14,0	p< 0,000047	miR-6501-5p	↓ 3,0	1,0	p< 0,000047
miR-486-5p	↓ 5824,0	193,5	p< 0,000047	miR-6806-3p	↓ 8,0	2,0	p< 0,000047	miR-6837-3p	↓ 9,0	4,0	p< 0,000047
miR-338-3p	↓ 15327,0	1068,5	p< 0,000047	miR-361-5p	↓ 1589,0	762,0	p< 0,000047	miR-32-5p	↓ 245,0	83,5	p< 0,000047
miR-195-3p	↓ 53,0	9,0	p< 0,000047	miR-1287-5p	↓ 222,0	69,0	p< 0,000047	miR-381-3p	↓ 265,0	89,5	p< 0,000047
miR-146b-3p	↓ 2666,0	399,5	p< 0,000047	miR-552-5p	↓ 11,0	1,0	p< 0,000047	miR-589-5p	↓ 448,0	245,0	p< 0,000047
miR-1976	↓ 194,0	31,5	p< 0,000047	miR-99b-5p	↓ 186478,0	64000,5	p< 0,000047	miR-653-5p	↓ 84,0	28,0	p< 0,000047
miR-29b-2-5p	↓ 352,0	74,0	p< 0,000047	miR-221-3p	↓ 2593,0	884,0	p< 0,000047	miR-516a-5p	↓ 10,0	1,0	p< 0,000047
miR-139-5p	↓ 861,0	68,5	p< 0,000047	miR-1247-5p	↓ 71,0	9,5	p< 0,000047	miR-192-3p	↓ 7,0	3,0	p< 0,000047
miR-3154	↓ 1,0	0,0	p< 0,000047	miR-664b-3p	↓ 27,0	10,0	p< 0,000047	miR-491-5p	↓ 12,0	5,0	p< 0,000047
miR-3065-3p	↓ 1530,0	138,0	p< 0,000047	miR-150-3p	↓ 23,0	5,0	p< 0,000047	miR-582-5p	↓ 91,0	45,0	p< 0,000047
miR-218-1-3p	↓ 21,0	2,0	p< 0,000047	miR-4521	↓ 7,0	1,0	p< 0,000047	miR-676-3p	↓ 7,0	2,0	p< 0,000047
miR-133b	↓ 28,0	1,0	p< 0,000047	miR-374a-5p	↓ 288,0	99,0	p< 0,000047	miR-185-3p	↓ 52,0	20,0	p< 0,000047
miR-126-5p	↓ 4856,0	677,0	p< 0,000047	miR-15a-5p	↓ 1389,0	437,5	p< 0,000047	miR-320a-3p	↓ 2707,0	1285,5	p< 0,000047
miR-4740-5p	↓ 2,0	0,0	p< 0,000047	miR-150-5p	↓ 8666,0	1921,5	p< 0,000047	miR-4999-5p	↓ 11,0	3,0	p< 0,000047
miR-144-5p	↓ 4754,0	191,0	p< 0,000047	miR-146b-5p	↓ 8656,0	2490,0	p< 0,000047	miR-3074-5p	↓ 36,0	16,0	p< 0,000047
miR-181a-5p	↓ 35469,0	5225,5	p< 0,000047	miR-1247-3p	↓ 160,0	34,0	p< 0,000047	miR-4423-5p	↓ 3,0	0,0	p< 0,000047
miR-133a-3p	↓ 156,0	10,0	p< 0,000047	miR-181a-2-3p	↓ 9371,0	2713,0	p< 0,000047	miR-6843-3p	↓ 1,0	0,0	p< 0,000047
miR-4777-3p	↓ 13,0	1,0	p< 0,000047	miR-3912-3p	↓ 6,0	1,0	p< 0,000047	miR-148b-3p	↓ 992,0	528,0	p< 0,000047
miR-218-5p	↓ 1525,0	163,5	p< 0,000047	miR-5000-3p	↓ 14,0	4,0	p< 0,000047	miR-501-5p	↓ 21,0	11,0	p< 0,000047
miR-140-3p	↓ 21235,0	2103,0	p< 0,000047	miR-30c-5p	↓ 4508,0	1436,0	p< 0,000047	miR-199b-5p	↓ 1587,0	641,0	p< 0,000047
miR-101-3p	↓ 281640,0	25732,5	p< 0,000047	miR-23b-3p	↓ 12061,0	4960,0	p< 0,000047	miR-374a-3p	↓ 6463,0	2372,5	p< 0,000047
miR-497-5p	↓ 319,0	55,0	p< 0,000047	miR-500a-3p	↓ 1722,0	707,0	p< 0,000047	miR-3064-3p	↓ 2,0	0,0	p< 0,000047
miR-100-5p	↓ 74525,0	12586,5	p< 0,000047	miR-5586-5p	↓ 35,0	8,0	p< 0,000047	miR-378a-5p	↓ 639,0	295,5	p< 0,000047
miR-6502-5p	↓ 8,0	1,0	p< 0,000047	miR-378c	↓ 46,0	14,0	p< 0,000047	miR-4632-3p	↓ 1,0	0,0	p< 0,000047
miR-30d-3p	↓ 91,0	15,0	p< 0,000047	miR-4781-3p	↓ 3,0	1,0	p< 0,000047	let-7f1-3p	↓ 16,0	8,0	p< 0,000047
miR-451a	↓ 22378,0	940,5	p< 0,000047	miR-194-5p	↓ 1366,0	397,5	p< 0,000047	miR-3688-3p	↓ 1,0	0,0	p< 0,000047
miR-181a-3p	↓ 3736,0	668,5	p< 0,000047	miR-362-3p	↓ 24,0	6,0	p< 0,000047	miR-500b-5p	↓ 29,0	12,0	p< 0,000047
miR-6892-5p	↓ 31,0	5,0	p< 0,000047	miR-6842-3p	↓ 29,0	10,0	p< 0,000047	miR-500a-5p	↓ 29,0	12,0	p< 0,000047
miR-181c-5p	↓ 662,0	116,0	p< 0,000047	miR-23a-3p	↓ 30627,0	12116,0	p< 0,000047	miR-5588-3p	↓ 1,0	0,0	p< 0,000047
miR-101-5p	↓ 89,0	18,0	p< 0,000047	miR-125b-2-3p	↓ 168,0	41,5	p< 0,000047	miR-574-3p	↓ 466,0	255,0	p< 0,000047
miR-548b-3p	↓ 28,0	4,0	p< 0,000047	let-7f1-5p	↓ 52001,0	16911,0	p< 0,000047	miR-217-5p	↓ 148,0	47,5	p< 0,000047
miR-652-3p	↓ 480,0	94,5	p< 0,000047	miR-1294	↓ 3,0	0,0	p< 0,000047	miR-27b-3p	↓ 12472,0	5848,5	p< 0,000047
miR-195-5p	↓ 559,0	71,0	p< 0,000047	miR-6788-3p	↓ 1,0	0,0	p< 0,000047	miR-455-5p	↓ 228,0	93,0	p< 0,000047
miR-30b-3p	↓ 35,0	7,0	p< 0,000047	miR-181c-3p	↓ 394,0	137,5	p< 0,000047	miR-136-3p	↓ 62,0	23,0	p< 0,000047
miR-206	↓ 10,0	0,0	p< 0,000047	let-7d-3p	↓ 2293,0	1159,5	p< 0,000047	miR-28-5p	↓ 689,0	355,0	p< 0,000047
miR-511-5p	↓ 313,0	52,0	p< 0,000047	miR-605-5p	↓ 4,0	1,0	p< 0,000047	miR-618	↓ 6,0	2,0	p< 0,000047
miR-4683	↓ 2,0	0,0	p< 0,000047	miR-125a-3p	↓ 37,0	12,0	p< 0,000047	miR-3130-5p	↓ 45,0	15,5	p< 0,000047
miR-532-5p	↓ 10452,0	2632,5	p< 0,000047	miR-488-5p	↓ 1,0	0,0	p< 0,000047	miR-216a-5p	↓ 4,0	1,0	p< 0,000047
miR-145-5p	↓ 16695,0	3830,5	p< 0,000047	miR-27a-5p	↓ 267,0	55,5	p< 0,000047	miR-5010-3p	↓ 5,0	2,0	p< 0,000047
miR-1258	↓ 15,0	1,0	p< 0,000047	miR-202-5p	↓ 4,0	0,0	p< 0,000047	miR-138-1-3p	↓ 3,0	1,0	p< 0,000047
miR-361-3p	↓ 1931,0	459,5	p< 0,000047	miR-99b-3p	↓ 581,0	189,0	p< 0,000047	miR-664a-5p	↓ 17,0	7,0	p< 0,000047
miR-4709-3p	↓ 24,0	3,0	p< 0,000047	miR-181d-5p	↓ 176,0	53,0	p< 0,000047	miR-486-3p	↓ 3,0	1,0	p< 0,000047
miR-375-3p	↓ 51463,0	4683,0	p< 0,000047	miR-6507-5p	↓ 2,0	0,0	p< 0,000047	miR-6755-5p	↓ 5,0	2,0	p< 0,000047
miR-29a-3p	↓ 80652,0	17045,5	p< 0,000047	miR-342-5p	↓ 54,0	27,0	p< 0,000047	miR-27a-3p	↓ 12990,0	5833,0	p< 0,000047
miR-30b-5p	↓ 7636,0	1011,0	p< 0,000047	miR-4796-5p	↓ 1,0	0,0	p< 0,000047	miR-6733-3p	↓ 2,0	1,0	p< 0,000047
miR-26a-5p	↓ 23808,0	5802,5	p< 0,000047	miR-133a-5p	↓ 1,0	0,0	p< 0,000047	miR-4785	↓ 1,0	0,0	p< 0,000047
miR-143-3p	↓ 825923,0	136831,5	p< 0,000047	miR-624-5p	↓ 14,0	5,0	p< 0,000047	miR-1266-5p	↓ 52,0	17,0	p< 0,000047
miR-99a-5p	↓ 5632,0	1042,0	p< 0,000047	miR-4636	↓ 4,0	1,0	p< 0,000047	miR-98-5p	↓ 332,0	171,0	p< 0,000047
miR-144-3p	↓ 325,0	17,0	p< 0,000047	miR-16-1-3p	↓ 31,0	15,0	p< 0,000047	miR-548j-5p	↓ 4,0	2,0	p< 0,000047
miR-126-3p	↓ 44604,0	8204,5	p< 0,000047	miR-7156-5p	↓ 7,0	1,0	p< 0,000047	miR-5002-5p	↓ 1,0	0,0	p< 0,000047
miR-3926	↓ 7,0	1,0	p< 0,000047	miR-3199	↓ 9,0	3,0	p< 0,000047	miR-342-3p	↓ 918,0	500,5	p< 0,000047
miR-145-3p	↓ 440,0	82,0	p< 0,000047	miR-130a-3p	↓ 632,0	264,5	p< 0,000047	miR-34a-3p	↓ 6,0	3,0	p< 0,000047
miR-4727-5p	↓ 2,0	0,0	p< 0,000047	let-7e-5p	↓ 5897,0	1999,0	p< 0,000047	miR-664b-5p	↓ 2,0	1,0	p< 0,000047
miR-660-5p	↓ 666,0	153,0	p< 0,000047	miR-887-5p	↓ 6,0	1,0	p< 0,000047	miR-4724-5p	↓ 3,0	1,0	p< 0,000047
miR-502-3p	↓ 100,0	31,0	p< 0,000047	miR-584-5p	↓ 538,0	153,0	p< 0,000047	miR-29a-5p	↓ 39,0	24,0	p< 0,000047
miR-4772-3p	↓ 106,0	23,0	p< 0,000047	miR-3193	↓ 6,0	1,0	p< 0,000047	miR-519a-3p	↓ 2,0	0,0	p< 0,000047
miR-99a-3p	↓ 32,0	6,0	p< 0,000047	miR-17-3p	↓ 2893,0	1132,0	p< 0,000047	miR-542-5p	↓ 10,0	4,0	p< 0,000047
miR-1-3p	↓ 189,0	13,0	p< 0,000047	miR-664a-3p	↓ 146,0	67,0	p< 0,000047	miR-335-5p	↓ 64,0	34,0	p< 0,000047
miR-6513-3p	↓ 2,0	0,0	p< 0,000047	miR-504-5p	↓ 7,0	2,0	p< 0,000047	miR-363-3p	↓ 82,0	24,0	p< 0,000047
miR-552-3p	↓ 8,0	0,0	p< 0,000047	miR-3913-5p	↓ 27,0	10,0	p< 0,000047	miR-26b-5p	↓ 3790,0	1878,5	p< 0,000047
miR-125a-5p	↓ 5201,0	1365,5	p< 0,000047	miR-184	↓ 96,0	8,0	p< 0,000047	miR-29b-3p	↓ 4083,0	1986,0	p< 0,000047

continuação

miRNAs	Mediana		valor p	miRNAs	Mediana		valor p	miRNAs	Mediana		valor p
	N	T			N	T			N	T	
let-7c-5p	↓ 17990,0	3509,0	$p < 0,000047$	miR-193a-5p	↓ 1404,0	570,0	$p < 0,000047$	miR-340-5p	↓ 63,0	30,0	$p < 0,000047$
miR-138-5p	↓ 56,0	9,5	$p < 0,000047$	miR-10a-3p	↓ 21,0	6,0	$p < 0,000047$	miR-151a-3p	↓ 15411,0	7350,0	$p < 0,000047$
miR-29c-3p	↓ 33637,0	5632,5	$p < 0,000047$	miR-4529-3p	↓ 3,0	0,0	$p < 0,000047$	miR-3616-3p	↓ 0,0	0,0	$p < 0,000047$
miR-223-3p	↓ 5214,0	1002,0	$p < 0,000047$	miR-185-5p	↓ 597,0	280,0	$p < 0,000047$	miR-24-1-5p	↓ 92,0	50,5	$p < 0,000047$
miR-135a-5p	↓ 15,0	1,0	$p < 0,000047$	let-7i-5p	↓ 3098,0	1603,5	$p < 0,000047$	miR-551a	↓ 7,0	2,0	$p < 0,000047$
let-7a-5p	↓ 305077,0	66045,5	$p < 0,000047$	miR-4536-3p	↓ 2,0	0,0	$p < 0,000047$	miR-501-3p	↓ 246,0	147,5	$p < 0,000047$
miR-4800-3p	↓ 1,0	0,0	$p < 0,000047$	miR-584-3p	↓ 3,0	0,0	$p < 0,000047$	miR-24-3p	↓ 13591,0	7015,5	$p < 0,000047$
let-7b-5p	↓ 164848,0	42747,0	$p < 0,000047$	miR-3117-3p	↓ 7,0	1,0	$p < 0,000047$	miR-25-3p	↓ 73933,0	33757,5	$p < 0,000047$
miR-29c-5p	↓ 203,0	75,0	$p < 0,000047$	miR-4705	↓ 2,0	0,0	$p < 0,000047$	miR-34c-3p	↓ 303,0	53,5	$p < 0,000047$
miR-511-3p	↓ 12,0	2,0	$p < 0,000047$	miR-582-3p	↓ 2184,0	953,5	$p < 0,000047$	miR-5684	↓ 4,0	1,0	$p < 0,000047$
miR-3614-5p	↓ 33,0	8,0	$p < 0,000047$	miR-6503-3p	↓ 2,0	0,0	$p < 0,000047$	miR-132-3p	↓ 548,0	277,0	$p < 0,000047$
miR-598-3p	↓ 240,0	60,0	$p < 0,000047$	miR-100-3p	↓ 3,0	0,0	$p < 0,000047$	miR-34c-5p	↓ 672,0	101,5	$p < 0,000047$
miR-190a-5p	↓ 48,0	7,0	$p < 0,000047$	let-7d-5p	↓ 1422,0	752,5	$p < 0,000047$	miR-514a-3p	↓ 89,0	23,0	$p < 0,000047$
miR-107	↓ 677,0	197,5	$p < 0,000047$	miR-3934-3p	↓ 18,0	7,0	$p < 0,000047$	miR-6514-5p	↓ 2,0	0,0	$p < 0,000047$
miR-151b	↓ 17,0	3,0	$p < 0,000047$	miR-6718-5p	↓ 9,0	2,0	$p < 0,000047$	miR-222-5p	↓ 56,0	28,5	$p < 0,000047$
miR-3614-3p	↓ 16,0	3,0	$p < 0,000047$	miR-330-3p	↓ 25,0	11,0	$p < 0,000047$	miR-4772-5p	↓ 8,0	3,0	$p < 0,000047$
let-7e-3p	↓ 117,0	35,5	$p < 0,000047$	miR-103a-3p	↓ 150566,0	61419,0	$p < 0,000047$	miR-934	↓ 7,0	1,0	$p < 0,000047$
miR-22-3p	↓ 901276,0	237824,0	$p < 0,000047$	miR-193a-3p	↓ 65,0	25,5	$p < 0,000047$	miR-1262	↓ 7,0	3,0	$p < 0,000047$
miR-5680	↓ 4,0	0,0	$p < 0,000047$	miR-676-5p	↓ 2,0	0,0	$p < 0,000047$	miR-5683	↓ 15,0	3,5	$p < 0,000047$
miR-181b-5p	↓ 4618,0	1172,5	$p < 0,000047$	miR-425-3p	↓ 56,0	25,0	$p < 0,000047$	miR-362-5p	↓ 63,0	32,0	$p < 0,000047$
miR-34a-5p	↓ 1454,0	526,0	$p < 0,000047$	miR-500b-3p	↓ 8,0	3,0	$p < 0,000047$	miR-423-5p	↓ 281,0	122,0	$p < 0,000047$
let-7b-3p	↓ 193,0	67,0	$p < 0,000047$	miR-548aw	↓ 2,0	0,0	$p < 0,000047$				

Tabela 33 – DE-miRNAs *upregulated* (↑) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUSC.

miRNAs	Mediana		valor p	miRNAs	Mediana		valor p	miRNAs	Mediana		valor p
	N	T			N	T			N	T	
miR-210-3p	↑ 402,0	6821,0	$p < 0,000047$	miR-708-3p	↑ 203,0	749,0	$p < 0,000047$	miR-20a-3p	↑ 14,0	28,0	$p < 0,000047$
miR-205-5p	↑ 416,0	40707,0	$p < 0,000047$	miR-105-5p	↑ 1,0	92,5	$p < 0,000047$	miR-615-3p	↑ 1,0	5,0	$p < 0,000047$
miR-205-3p	↑ 0,0	19,0	$p < 0,000047$	miR-149-5p	↑ 125,0	519,5	$p < 0,000047$	miR-1277-3p	↑ 2,0	5,0	$p < 0,000047$
miR-708-5p	↑ 40,0	343,5	$p < 0,000047$	miR-9-3p	↑ 1,0	9,0	$p < 0,000047$	miR-4778-3p	↑ 0,0	1,0	$p < 0,000047$
miR-196b-5p	↑ 48,0	1044,0	$p < 0,000047$	miR-767-5p	↑ 1,0	49,0	$p < 0,000047$	miR-4665-5p	↑ 0,0	1,0	$p < 0,000047$
miR-4652-5p	↑ 0,0	13,0	$p < 0,000047$	miR-130b-3p	↑ 30,0	71,5	$p < 0,000047$	miR-4668-3p	↑ 5,0	10,0	$p < 0,000047$
miR-944	↑ 10,0	378,0	$p < 0,000047$	miR-455-3p	↑ 318,0	891,0	$p < 0,000047$	miR-224-5p	↑ 105,0	356,0	$p < 0,000047$
miR-301b-5p	↑ 0,0	4,0	$p < 0,000047$	miR-31-3p	↑ 2,0	28,0	$p < 0,000047$	miR-767-3p	↑ 0,0	1,0	$p < 0,000047$
miR-9-5p	↑ 547,0	9274,5	$p < 0,000047$	miR-744-3p	↑ 31,0	78,0	$p < 0,000047$	miR-1248	↑ 0,0	2,0	$p < 0,000047$
miR-130b-5p	↑ 24,0	71,5	$p < 0,000047$	miR-6510-3p	↑ 2,0	32,0	$p < 0,000047$	miR-573	↑ 0,0	1,0	$p < 0,000047$
miR-6499-3p	↑ 0,0	5,0	$p < 0,000047$	miR-3662	↑ 0,0	3,0	$p < 0,000047$	miR-93-3p	↑ 25,0	53,0	$p < 0,000047$
miR-629-3p	↑ 9,0	30,0	$p < 0,000047$	miR-1910-5p	↑ 0,0	2,0	$p < 0,000047$	miR-4449	↑ 0,0	2,0	$p < 0,000047$
miR-577	↑ 2,0	28,0	$p < 0,000047$	miR-196a-5p	↑ 8,0	89,5	$p < 0,000047$	miR-4713-5p	↑ 0,0	1,0	$p < 0,000047$
miR-183-3p	↑ 1,0	6,0	$p < 0,000047$	miR-345-5p	↑ 33,0	82,0	$p < 0,000047$	miR-141-3p	↑ 2018,0	3672,0	$p < 0,000047$
miR-1269a	↑ 4,0	589,5	$p < 0,000047$	miR-1293	↑ 0,0	2,0	$p < 0,000047$	miR-365a-5p	↑ 0,0	3,0	$p < 0,000047$
miR-301a-5p	↑ 7,0	25,0	$p < 0,000047$	miR-323b-3p	↑ 1,0	10,0	$p < 0,000047$	miR-5579-3p	↑ 0,0	2,0	$p < 0,000047$
miR-31-5p	↑ 2,0	42,0	$p < 0,000047$	miR-503-5p	↑ 5,0	18,0	$p < 0,000047$	miR-193b-5p	↑ 8,0	17,0	$p < 0,000047$
miR-183-5p	↑ 11809,0	38341,5	$p < 0,000047$	miR-33b-5p	↑ 4,0	12,0	$p < 0,000047$	miR-7-5p	↑ 2,0	5,0	$p < 0,000047$
miR-96-5p	↑ 23,0	84,0	$p < 0,000047$	miR-4766-3p	↑ 0,0	2,0	$p < 0,000047$	miR-200b-5p	↑ 33,0	64,5	$p < 0,000047$
miR-616-5p	↑ 2,0	10,0	$p < 0,000047$	miR-33a-5p	↑ 49,0	128,0	$p < 0,000047$	miR-579-5p	↑ 1,0	3,0	$p < 0,000047$
miR-301b-3p	↑ 1,0	6,0	$p < 0,000047$	miR-877-5p	↑ 2,0	8,0	$p < 0,000047$	miR-671-5p	↑ 9,0	20,0	$p < 0,000047$
miR-6499-5p	↑ 0,0	6,0	$p < 0,000047$	miR-7974	↑ 0,0	1,0	$p < 0,000047$				

A análise de expressão diferencial também foi realizada para identificar os DEGs entre os grupos T vs. N nas 40.919 expressões de RNA-seq. O teste de Wilcoxon-Mann-Whitney com correção de Bonferroni foi realizado, para isso, é necessário calcular o valor de $\alpha_{\text{corrigido}}$:

$$\alpha_{\text{corrigido}} = \frac{\alpha_{\text{original}}}{n} = \frac{0,05}{40.919} \approx 0,0000012.$$

Considerando a correção de Bonferroni com valor $p < 0,0000012$, identificou-se 9.877 DEGs, sendo 6.980 *downregulated* e 2897 *upregulated*, ambos em relação ao N. Os 200 DEGs com menores valores de p estão apresentados na Tabela 34, o *UBE2T* apresentou o menor valor p e foi único *upregulated*, em comparado aos 200 apresentados.

Tabela 34 – 200 DEGs mais importantes identificados no teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUSC.

RNAs	Mediana N	T	valor p	RNAs	Mediana N	T	valor p	RNAs	Mediana N	T	valor p
UBE2T	↑ 119	1376	p<0,0000012	GGTLC3	↓ 10	0	p<0,0000012	CNTN6	↓ 1039	10	p<0,0000012
AL355499.2	↓ 8	0	p<0,0000012	AL031681.1	↓ 3	0	p<0,0000012	EDNRB	↓ 9973	393	p<0,0000012
LINC01070	↓ 10	0	p<0,0000012	AC236972.3	↓ 171	1	p<0,0000012	AQP4	↓ 37651	683	p<0,0000012
AL356095.1	↓ 6	0	p<0,0000012	GP9	↓ 27	1	p<0,0000012	CAV1	↓ 88866	4976	p<0,0000012
AC010776.3	↓ 11	0	p<0,0000012	AL035670.1	↓ 10	0	p<0,0000012	TNNC1	↓ 3951	55	p<0,0000012
AC105914.2	↓ 12	0	p<0,0000012	ACSM5P1	↓ 3	0	p<0,0000012	GPIHBP1	↓ 3241	47	p<0,0000012
AC022325.2	↓ 8	0	p<0,0000012	AC226101.1	↓ 4	0	p<0,0000012	AL161457.1	↓ 3	0	p<0,0000012
LINC01584	↓ 3	0	p<0,0000012	CELF2.AS2	↓ 6	0	p<0,0000012	PEBP4	↓ 8831	129	p<0,0000012
AL133466.1	↓ 3	0	p<0,0000012	AL355499.1	↓ 32	1	p<0,0000012	ARHGAP44	↓ 3445	276	p<0,0000012
TBC1D26.AS1	↓ 5	0	p<0,0000012	RPL31P40	↓ 6	0	p<0,0000012	TNS1	↓ 73429	6262	p<0,0000012
AC010776.2	↓ 19	0	p<0,0000012	AC073862.5	↓ 21	1	p<0,0000012	HCG21	↓ 3	0	p<0,0000012
AL136369.1	↓ 35	0	p<0,0000012	LINC02346	↓ 8	0	p<0,0000012	TMEM125	↓ 8699	665	p<0,0000012
GYPB	↓ 3	0	p<0,0000012	AC135012.3	↓ 71	1	p<0,0000012	MTARC2	↓ 2745	555	p<0,0000012
AL136369.2	↓ 2	0	p<0,0000012	ARHGEF2.AS1	↓ 49	1	p<0,0000012	CLEC3B	↓ 3354	70	p<0,0000012
HBM	↓ 5	0	p<0,0000012	CYP1A2	↓ 81	1	p<0,0000012	CD302	↓ 1213	118	p<0,0000012
Y_RNA.513	↓ 6	0	p<0,0000012	AL008733.1	↓ 22	1	p<0,0000012	ADAMTS9.AS2	↓ 220	15	p<0,0000012
NCAPGP2	↓ 21	0	p<0,0000012	AC073149.1	↓ 12	0	p<0,0000012	GRK5	↓ 8195	817	p<0,0000012
AC092944.2	↓ 4	0	p<0,0000012	RS1	↓ 255	2	p<0,0000012	RPS29P11	↓ 146	4	p<0,0000012
KHDRBS2.OT1	↓ 20	0	p<0,0000012	AC006159.1	↓ 10	0	p<0,0000012	FHL5	↓ 1084	47	p<0,0000012
FAM138B	↓ 2	0	p<0,0000012	AC105254.1	↓ 5	0	p<0,0000012	FILIP1	↓ 1268	87	p<0,0000012
AC078899.3	↓ 8	0	p<0,0000012	AC008268.1	↓ 1671	2	p<0,0000012	CACNA2D2	↓ 14311	276	p<0,0000012
AC093787.1	↓ 27	0	p<0,0000012	AC012103.1	↓ 13	0	p<0,0000012	CYP4B1	↓ 40570	580	p<0,0000012
LINC01785	↓ 10	0	p<0,0000012	GGTLC2	↓ 36	1	p<0,0000012	ESAM	↓ 10894	1085	p<0,0000012
AC073862.4	↓ 6	0	p<0,0000012	HELT	↓ 26	0	p<0,0000012	AGER	↓ 121127	588	p<0,0000012
AC026780.2	↓ 3	0	p<0,0000012	AC010998.1	↓ 40	1	p<0,0000012	ACVRL1	↓ 12234	1086	p<0,0000012
AC095050.1	↓ 35	0	p<0,0000012	AL135841.1	↓ 8	0	p<0,0000012	PALM2AKAP2	↓ 34232	2547	p<0,0000012
DPPA3P2	↓ 32	0	p<0,0000012	PGM5P4	↓ 45	1	p<0,0000012	C13orf46	↓ 774	17	p<0,0000012
LINC01082	↓ 13	0	p<0,0000012	LINC01645	↓ 37	1	p<0,0000012	PLA2G1B	↓ 909	11	p<0,0000012
LINC01985	↓ 8	0	p<0,0000012	AL138826.1	↓ 2	0	p<0,0000012	SFTA2	↓ 9435	183	p<0,0000012
RXFP2	↓ 39	0	p<0,0000012	AL606469.1	↓ 138	2	p<0,0000012	SCN7A	↓ 5540	162	p<0,0000012
AC053503.5	↓ 7	0	p<0,0000012	LINC01996	↓ 169	1	p<0,0000012	JAM2	↓ 3991	337	p<0,0000012
LINC01616	↓ 7	0	p<0,0000012	AC105053.1	↓ 28	1	p<0,0000012	NPR1	↓ 5074	249	p<0,0000012
AC024257.1	↓ 6	0	p<0,0000012	RNU6.917P	↓ 6	0	p<0,0000012	RAMP2	↓ 5173	449	p<0,0000012
GUCA2A	↓ 27	0	p<0,0000012	AL136452.1	↓ 82	2	p<0,0000012	CDH5	↓ 18430	1290	p<0,0000012
LINC00710	↓ 9	0	p<0,0000012	RNU1.38P	↓ 6	0	p<0,0000012	AP000866.2	↓ 199	15	p<0,0000012
AL161740.1	↓ 3	0	p<0,0000012	LINC02858	↓ 8	0	p<0,0000012	LRRC36	↓ 1232	33	p<0,0000012
AC093151.5	↓ 7	0	p<0,0000012	ECELIP2	↓ 530	6	p<0,0000012	AC112206.2	↓ 310	11	p<0,0000012
AC005998.1	↓ 3	0	p<0,0000012	GKN2	↓ 2885	18	p<0,0000012	SFTA3	↓ 3865	76	p<0,0000012
AL773573.1	↓ 9	0	p<0,0000012	AC110014.1	↓ 6	0	p<0,0000012	TEK	↓ 5888	264	p<0,0000012
AL357146.1	↓ 12	0	p<0,0000012	OR6K3	↓ 17	0	p<0,0000012	RTKN2	↓ 18121	600	p<0,0000012
LINC01863	↓ 42	0	p<0,0000012	EMP2	↓ 85201	7837	p<0,0000012	PTPRB	↓ 11343	705	p<0,0000012
AC091305.1	↓ 19	0	p<0,0000012	CLDN18	↓ 68990	153	p<0,0000012	GPX3	↓ 64804	3955	p<0,0000012
AL445430.2	↓ 9	0	p<0,0000012	CLIC5	↓ 26443	280	p<0,0000012	CAT	↓ 20949	4116	p<0,0000012
AC104260.1	↓ 2	0	p<0,0000012	ADGRF5	↓ 32350	1520	p<0,0000012	AC104984.4	↓ 47	1	p<0,0000012
AL354714.3	↓ 29	0	p<0,0000012	LINC02817	↓ 12	1	p<0,0000012	NAGPA.AS1	↓ 90	4	p<0,0000012
AC092944.3	↓ 8	0	p<0,0000012	HID1.AS1	↓ 65	3	p<0,0000012	CA4	↓ 2783	26	p<0,0000012
KIAA0408	↓ 32	0	p<0,0000012	PGM5P4.AS1	↓ 25	1	p<0,0000012	ABCA8	↓ 4408	122	p<0,0000012
GUCA2B	↓ 7	0	p<0,0000012	TEX49	↓ 4	0	p<0,0000012	AK1	↓ 1235	171	p<0,0000012
AC005908.2	↓ 4	0	p<0,0000012	ADGRD1	↓ 6137	195	p<0,0000012	FAM189A2	↓ 4823	195	p<0,0000012
AL772337.2	↓ 18	0	p<0,0000012	LMO7	↓ 36183	3328	p<0,0000012	HYAL1	↓ 4205	355	p<0,0000012
AC073862.2	↓ 21	1	p<0,0000012	TGFBR2	↓ 37000	3947	p<0,0000012	F11	↓ 530	6	p<0,0000012
BEND2	↓ 6	0	p<0,0000012	CTSLP8	↓ 16	0	p<0,0000012	SEC14L6	↓ 1361	63	p<0,0000012
LINC02016	↓ 179	1	p<0,0000012	AC243772.3	↓ 21	1	p<0,0000012	MYCT1	↓ 2415	227	p<0,0000012
CABP5	↓ 2	0	p<0,0000012	OTC	↓ 55	1	p<0,0000012	INMT	↓ 13425	283	p<0,0000012
AC025431.1	↓ 3	0	p<0,0000012	AC093110.1	↓ 699	38	p<0,0000012	LDB2	↓ 5242	411	p<0,0000012
AC007861.1	↓ 10	0	p<0,0000012	AC104211.1	↓ 76	3	p<0,0000012	C1orf116	↓ 22736	1994	p<0,0000012
RNU5B.4P	↓ 12	1	p<0,0000012	SLC39A8	↓ 41416	1754	p<0,0000012	LHFPL3.AS2	↓ 1424	33	p<0,0000012
AC067752.1	↓ 25	0	p<0,0000012	PECAM1	↓ 61923	6130	p<0,0000012	TNXB	↓ 11173	247	p<0,0000012
PGM5.AS1	↓ 20	0	p<0,0000012	GGTLC1	↓ 2271	16	p<0,0000012	ROBO4	↓ 8738	518	p<0,0000012
AL133355.1	↓ 10	1	p<0,0000012	AC020895.2	↓ 2	0	p<0,0000012	DPYSL2	↓ 30483	3085	p<0,0000012
CHIAP2	↓ 763	1	p<0,0000012	HSD17B6	↓ 6044	210	p<0,0000012	EPAS1	↓ 134797	10667	p<0,0000012
LINC00670	↓ 10	0	p<0,0000012	SFTPC	↓ 1554969	1573	p<0,0000012	AC079467.1	↓ 241	6	p<0,0000012
AL445307.1	↓ 29	1	p<0,0000012	WWC2	↓ 9016	683	p<0,0000012	AC027288.3	↓ 550	20	p<0,0000012
AL445493.3	↓ 31	1	p<0,0000012	C16orf89	↓ 15808	436	p<0,0000012	AP001528.2	↓ 368	20	p<0,0000012
AC092691.1	↓ 42	0	p<0,0000012	PTPN21	↓ 5587	596	p<0,0000012	SFTA1P	↓ 4798	92	p<0,0000012
LDLRAD4.AS1	↓ 16	1	p<0,0000012	SELENBP1	↓ 26873	1387	p<0,0000012	SHE	↓ 3413	185	p<0,0000012
AL353152.1	↓ 9	0	p<0,0000012	TMEM252	↓ 29	1	p<0,0000012				

↑ upregulated e ↓ downregulated.

Análises Biológicas

Foram conduzidas análises biológicas visando integrar e examinar os DE-miRNAs ($n = 337$, sendo 272 *downregulated* e 65 *upregulated*) e os DEGs ($n = 9877$, sendo 6980 *downregulated* e 2897 *upregulated*) identificados.

Primeiramente foi realizada a predição de genes-alvos dos DE-miRNAs, por meio da ferramenta mirDIP. Identificaram-se 60.664 genes-alvos dos DE-miRNAs (pares de combinações miRNA - gene-alvo), o qual sem duplicatas contém 10.634 genes-alvos. O Diagrama de Venn (Figura 37) mostra haver uma sobreposição de 23,3% entre os genes-alvos dos DE-miRNAs e dos DEGs, os genes que estão na sobreposição serão nomeados de DEGs candidatos e seguiram nas próximas análises.

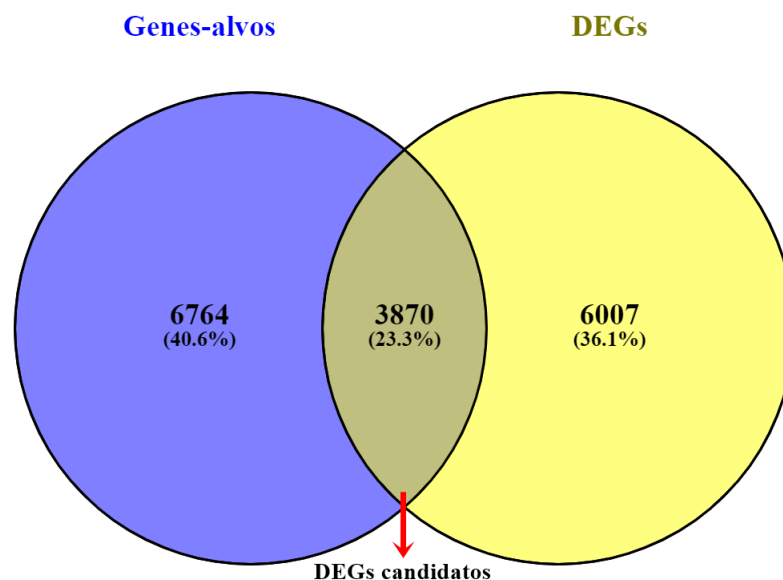


Figura 37 – Diagrama de Venn entre genes-alvos dos DE-miRNAs e DEGs - LUSC.

A lista gerada pelo mirDIP, contendo os genes-alvos, foi filtrada com base nos DEGs candidatos, resultando na identificação de 22.354 pares formados pelos DEGs candidatos e pelos DE-miRNAs que os regulam. Foram selecionados DEGs candidatos cuja expressão era oposta à dos DE-miRNAs. Dessa maneira, identificaram-se 6.872 pares com expressões opostas.

Para realizar a análise de correlação entre os 6.872 pares com expressões opostas, foi necessário combinar os bancos de dados de miRNA-seq ($n = 249$) e RNA-seq ($n = 266$), utilizando o ID da amostra do paciente como referência. Isso resultou em um banco de dados combinado contendo $n = 244$ amostras, que estavam presentes em ambos os conjuntos de dados.

Calculou-se a correlação linear de Spearman entre os 6.872 pares, considerando como estatisticamente significativos aqueles que apresentavam correlações negativas com $\rho < -0,3$ e $p < 0,05$. Assim, foram identificados 766 pares significantes (Tabela S4, material suplementar). A

Tabela 35 exibe os 50 pares com as maiores correlações negativas, a menor correlação negativa foi de $\rho = -0,55$ entre miR-205-5p e *ENPP4* (*Ectonucleotide Pyrophosphatase/Phosphodiesterase 4*).

Tabela 35 – As 50 maiores correlações negativas entre os pares - LUSC.

miRNA	Gene	ρ	p
miR-205-5p	ENPP4	-0,55	$p < 0,05$
miR-205-5p	MGAT4A	-0,55	$p < 0,05$
miR-141-3p	ZEB2	-0,54	$p < 0,05$
miR-205-5p	CFL2	-0,53	$p < 0,05$
miR-140-3p	OLA1	-0,53	$p < 0,05$
miR-100-5p	BIRC5	-0,51	$p < 0,05$
miR-944	FGL2	-0,50	$p < 0,05$
miR-944	FLI1	-0,50	$p < 0,05$
miR-205-3p	CELF2	-0,50	$p < 0,05$
miR-146b-5p	PGK1	-0,49	$p < 0,05$
miR-205-3p	CPED1	-0,49	$p < 0,05$
miR-205-3p	RAB8B	-0,49	$p < 0,05$
miR-30c-2-3p	PAK1	-0,49	$p < 0,05$
miR-205-3p	CYRIA	-0,49	$p < 0,05$
miR-195-3p	PDCD10	-0,48	$p < 0,05$
miR-708-5p	RAB8B	-0,48	$p < 0,05$
miR-29b-3p	TDG	-0,48	$p < 0,05$
miR-141-3p	ZCCHC24	-0,48	$p < 0,05$
miR-150-5p	PFN2	-0,48	$p < 0,05$
miR-96-5p	RAB8B	-0,48	$p < 0,05$
miR-30c-5p	TWF1	-0,47	$p < 0,05$
miR-497-5p	CDCA4	-0,47	$p < 0,05$
miR-210-3p	CNRIP1	-0,47	$p < 0,05$
miR-877-5p	HSD11B1	-0,47	$p < 0,05$
miR-205-5p	MFNG	-0,46	$p < 0,05$
miR-205-5p	HSD17B11	-0,46	$p < 0,05$
miR-361-5p	CKS1B	-0,46	$p < 0,05$
miR-944	UBL3	-0,46	$p < 0,05$
miR-141-3p	ATP6V1B2	-0,46	$p < 0,05$
miR-660-5p	SINHCAF	-0,46	$p < 0,05$
miR-141-3p	ZEB1	-0,46	$p < 0,05$
miR-141-3p	SDC2	-0,45	$p < 0,05$
let-7e-5p	BZW2	-0,45	$p < 0,05$
miR-30d-5p	TWF1	-0,45	$p < 0,05$
miR-195-5p	CDCA4	-0,45	$p < 0,05$
miR-210-3p	ISCU	-0,45	$p < 0,05$
miR-30d-5p	DBF4	-0,45	$p < 0,05$
let-7e-3p	Nup37	-0,45	$p < 0,05$
miR-210-3p	KCNA3	-0,45	$p < 0,05$
miR-181a-3p	Nup37	-0,45	$p < 0,05$
miR-30a-5p	YWHAZ	-0,45	$p < 0,05$
miR-140-5p	TTK	-0,45	$p < 0,05$
miR-31-5p	STK40	-0,45	$p < 0,05$
miR-708-5p	DRAM1	-0,45	$p < 0,05$
miR-205-5p	LYSMD3	-0,45	$p < 0,05$
miR-30c-5p	YWHAZ	-0,44	$p < 0,05$
miR-141-3p	QKI	-0,44	$p < 0,05$
miR-29a-3p	TDG	-0,44	$p < 0,05$
miR-30d-3p	CACYBP	-0,44	$p < 0,05$
miR-100-5p	CDC25A	-0,44	$p < 0,05$

Os 766 pares foram submetidos à análise no miRTarBase, a fim de selecionar exclusivamente as interações validadas experimentalmente entre DE-miRNAs e DEGs candidatos.

Foram identificados 161 pares, dos quais 46 consistiam em DE-miRNA *upregulated* e DEGs *downregulated* (Tabela 36), enquanto 115 eram compostos por DE-miRNAs *downregulated* e DEGs *upregulated* (Tabela 39).

Tabela 36 – DE-miRNAs *up* e DEGs *down* validados experimentalmente - LUSC.

ID	miRNA		Gene	
MIRT440724	miR-205-5p	↑	ENPP4	↓
MIRT000316	miR-141-3p	↑	ZEB2	↓
MIRT000317	miR-141-3p	↑	ZEB1	↓
MIRT003163	miR-210-3p	↑	ISCU	↓
MIRT007319	miR-31-5p	↑	STK40	↓
MIRT440298	miR-205-5p	↑	LYSMD3	↓
MIRT099339	miR-141-3p	↑	QKI	↓
MIRT054210	miR-210-3p	↑	SH3BGR1	↓
MIRT003666	miR-205-5p	↑	ZEB2	↓
MIRT502520	miR-301b-3p	↑	ESR1	↓
MIRT224764	miR-944	↑	DPYSL2	↓
MIRT548121	miR-944	↑	GATA6	↓
MIRT001989	miR-96-5p	↑	MITF	↓
MIRT007191	miR-205-5p	↑	LRRK2	↓
MIRT502518	miR-130b-3p	↑	ESR1	↓
MIRT020143	miR-130b-3p	↑	TGFBR2	↓
MIRT052905	miR-141-3p	↑	STAT4	↓
MIRT080800	miR-767-5p	↑	SH3GLB1	↓
MIRT091701	miR-301b-3p	↑	TGFBR2	↓
MIRT439886	miR-210-3p	↑	PSAP	↓
MIRT130344	miR-205-5p	↑	PTPRJ	↓
MIRT549458	miR-301b-3p	↑	ACSL4	↓
MIRT003165	miR-210-3p	↑	GPD1L	↓
MIRT002478	miR-205-5p	↑	ZEB1	↓
MIRT731301	miR-301b-3p	↑	PTEN	↓
MIRT191923	miR-196a-5p	↑	CALM1	↓
MIRT672611	miR-301b-3p	↑	IRF1	↓
MIRT554511	miR-210-3p	↑	RUNX1T1	↓
MIRT003281	miR-141-3p	↑	PTEN	↓
MIRT549459	miR-130b-3p	↑	ACSL4	↓
MIRT719385	miR-6499-5p	↑	SECTM1	↓
MIRT672613	miR-130b-3p	↑	IRF1	↓
MIRT546287	miR-205-3p	↑	TMEM245	↓
MIRT027934	miR-96-5p	↑	PROK2	↓
MIRT658778	miR-96-5p	↑	EIF4EBP2	↓
MIRT054654	miR-96-5p	↑	ZEB1	↓
MIRT440647	miR-210-3p	↑	FCHSD2	↓
MIRT099326	miR-301b-3p	↑	QKI	↓
MIRT718805	miR-4778-3p	↑	PRKCB	↓
MIRT054104	miR-205-5p	↑	PTPRM	↓
MIRT053520	miR-224-5p	↑	DPYSL2	↓
MIRT736999	miR-183-5p	↑	PTEN	↓
MIRT657497	miR-4778-3p	↑	HBEGF	↓
MIRT224763	miR-301b-3p	↑	DPYSL2	↓
MIRT439720	miR-205-5p	↑	RTN3	↓
MIRT054460	miR-130b-3p	↑	PTEN	↓

↑ *upregulated* e ↓ *downregulated*.

A rede de regulatória entre DE-miRNAs *upregulated* e DEGs *downregulated* foi construída e está apresentada na Figura 38. A rede possui destaque para alguns miRNAs, como:

miR-210-3p ligado a *ISCU*, *SH3BGRL*, *PSAP*, *GPD1L*, *RUNX1T1* e *FCHSD2*, miR-205-5p ligado a *ZEB1*, *ENPP4*, *LYSMD*, *PTPRJ*, *PTPRM*, *RTN3*, *LRRK2* e *ZEB2*, além dos miR-141-3p, miR-301b-3p e miR-130b-3p relacionados a outros DEGs.

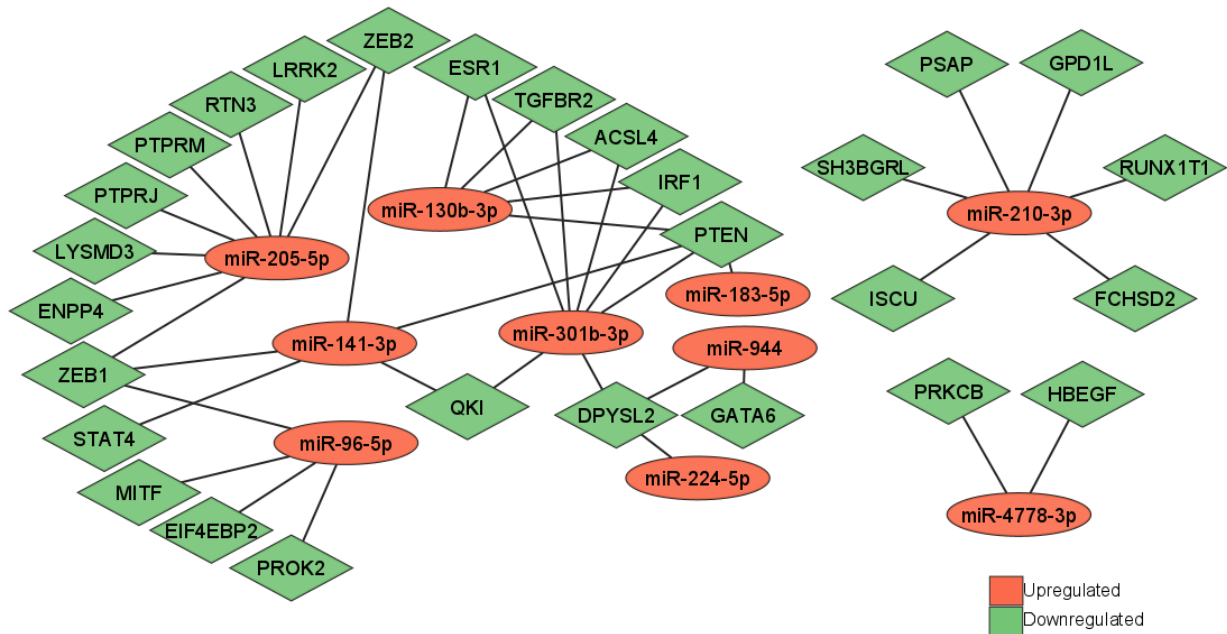


Figura 38 – Rede regulatória entre DE-miRNAs *up* e DEGs *down* - LUSC.

Realizou-se a análise de enriquecimento das vias para os DE-miRNAs e DEGs da Tabela 36. Para os DE-miRNAs, utilizou-se a ferramenta miRPath, e para os DEGs, o Enrichr. A Tabela 37 apresenta as principais vias identificadas com os DE-miRNAs (lista completa na Tabela S5), e a Tabela 38 apresentam todas as vias identificadas com DEGs, considerando valor $p < 0,05$. Essas vias estão envolvidas em diversos processos biológicos, como regulação do ciclo celular, sinalização celular, apoptose, proliferação celular, e resposta a fatores externos. A ativação ou não dessas vias pode desempenhar um papel importante na progressão e desenvolvimento do LUSC.

Tabela 37 – Principais vias enriquecidas para DE-miRNAs *up* - LUSC.

<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Adherens junction	79	45	12	p<0,01
FoxO signaling pathway	139	66	13	p<0,01
Focal adhesion	213	87	13	p<0,01
Cell cycle	129	59	13	p<0,01
Regulation of actin cytoskeleton	224	87	13	p<0,01
Ubiquitin mediated proteolysis	142	62	11	p<0,01
Pathways in cancer	555	176	13	p<0,01
Proteoglycans in cancer	220	85	13	p<0,01
p53 signaling pathway	75	39	12	p<0,01
PI3K-Akt signaling pathway	372	123	13	p<0,01
Hippo signaling pathway	164	65	12	p<0,01
EGFR tyrosine kinase inhibitor resistance	82	39	12	p<0,01
Endocytosis	311	105	13	p<0,01
HIF-1 signaling pathway	112	48	13	p<0,01
Small cell lung cancer	100	42	13	p<0,01
Non-small cell lung cancer	81	36	12	p<0,01
MAPK signaling pathway	329	105	13	p<0,01
RNA transport	199	69	12	p<0,01
mTOR signaling pathway	177	62	12	p<0,01
Breast cancer	163	55	13	0,001
Thyroid cancer	43	20	13	0,001
Ras signaling pathway	241	72	13	0,003
Apoptosis	151	48	13	0,005
Wnt signaling pathway	173	50	11	0,030

Tabela 38 – Vias enriquecidas para DEGs *down* - LUSC.

<i>Term</i>	<i>Overlap</i>	<i>Adjusted P value</i>	<i>Genes</i>
Pathways in cancer	8/531	0,000	PRKCB;PTEN;STAT4;MITF;CALM1;ESR1;RUNX1T1;TGFB2
Adherens junction	3/71	0,010	PTPRM;PTPRJ;TGFB2
Glioma	3/75	0,010	PRKCB;PTEN;CALM1
Transcriptional misregulation in cancer	4/192	0,010	ZEB1;MITF;RUNX1T1;TGFB2
GnRH signaling pathway	3/93	0,013	PRKCB;CALM1;HBEGF
Phosphatidylinositol signaling system	3/97	0,013	PRKCB;PTEN;CALM1
Melanogenesis	3/101	0,013	PRKCB;MITF;CALM1
Estrogen signaling pathway	3/137	0,026	CALM1;ESR1;HBEGF
MicroRNAs in cancer	4/310	0,026	ZEB2;ZEB1;PRKCB;PTEN
Cellular senescence	3/156	0,031	PTEN;CALM1;TGFB2
Hepatitis B	3/162	0,031	PRKCB;STAT4;TGFB2
Hepatocellular carcinoma	3/168	0,032	PRKCB;PTEN;TGFB2
Endocrine and other factor-regulated calcium reabsorption	2/53	0,003	PRKCB;ESR1
Diabetic cardiomyopathy	3/203	0,045	PRKCB;PTEN;TGFB2
Proteoglycans in cancer	3/205	0,045	PRKCB;ESR1;HBEGF
Long-term potentiation	2/67	0,048	PRKCB;CALM1
Amphetamine addiction	2/69	0,048	PRKCB;CALM1
Prolactin signaling pathway	2/70	0,048	IRF1;ESR1
Melanoma	2/72	0,048	PTEN;MITF
Gastric acid secretion	2/76	0,048	PRKCB;CALM1
Pertussis	2/76	0,048	IRF1;CALM1

Os 115 pares de DE-miRNAs *down* e DEGs candidatos *up* estão apresentados na Tabela 39. A rede regulatória foi construída e está apresentada na Figura 39. Apesar da rede regulatória estar complexa, o maior destaque foi a presença da família miR-30: miR-30a-5p, miR-30b-5p, miR-30c-5p e miR-30d-5p com as maiores quantidades de ligações. Mas também há outros miRNAs com bastante ligações, como os miR-26a-5p e o miR-26b-5p. O *LRRC8D* (*Leucine Rich Repeat Containing 8 Family Member D*) também está relacionado aos miR-181d-5p, 181c-5p, 181b-5p, 30c,5p, 181a-5p, 30a-5p e 30d-5p.

Tabela 39 – DE-miRNAs *down* e DEGs *up* validados experimentalmente - LUSC.

ID	miRNA		Gene		ID	miRNA		Gene	
MIRT053293	miR-29b-3p	↓	TDG	↑	MIRT168588	let-7i-5p	↓	HMGA1	↑
MIRT007203	miR-30c-5p	↓	TWF1	↑	MIRT732968	miR-206	↓	TWF1	↑
MIRT502918	miR-497-5p	↓	CDCA4	↑	MIRT067384	miR-29a-3p	↓	TMTC3	↑
MIRT726209	miR-30d-5p	↓	TWF1	↑	MIRT561386	miR-107	↓	TWF1	↑
MIRT502917	miR-195-5p	↓	CDCA4	↑	MIRT067976	miR-30c-5p	↓	TFDP1	↑
MIRT171550	miR-30d-5p	↓	DBF4	↑	MIRT032297	let-7b-5p	↓	IGF2BP3	↑
MIRT028537	miR-30a-5p	↓	YWHAZ	↑	MIRT467969	miR-29a-3p	↓	SLC16A1	↑
MIRT053292	miR-29a-3p	↓	TDG	↑	MIRT003976	miR-1-3p	↓	HSPD1	↑
MIRT088445	miR-30c-5p	↓	LCLAT1	↑	MIRT695702	miR-4683	↓	OLA1	↑
MIRT732469	miR-145-5p	↓	MCM2	↑	MIRT005164	miR-30a-5p	↓	TWF1	↑
MIRT053107	miR-223-3p	↓	ECT2	↑	MIRT735058	miR-125b-5p	↓	SUV39H1	↑
MIRT052385	let-7a-5p	↓	RRM2	↑	MIRT088446	miR-30d-5p	↓	LCLAT1	↑
MIRT561387	miR-103a-3p	↓	TWF1	↑	MIRT193070	miR-26a-5p	↓	MAPK6	↑
MIRT733026	miR-30a-5p	↓	CCNE2	↑	MIRT297014	let-7f-5p	↓	RRM2	↑
MIRT736272	miR-30a-5p	↓	CCNE2	↑	MIRT007177	miR-26a-5p	↓	E2F7	↑
MIRT737208	miR-101-5p	↓	GIN51	↑	MIRT535653	miR-3926	↓	NLN	↑
MIRT025186	miR-181a-5p	↓	METAP1	↑	MIRT029831	miR-26b-5p	↓	RCN2	↑
MIRT538618	miR-139-5p	↓	CCT5	↑	MIRT206887	miR-30c-5p	↓	XPO1	↑
MIRT043700	miR-342-3p	↓	RRM2	↑	MIRT346489	miR-26a-5p	↓	SSX2IP	↑
MIRT002296	let-7b-5p	↓	CDC25A	↑	MIRT003025	miR-29c-3p	↓	DNMT3B	↑
MIRT001921	miR-29c-3p	↓	TDG	↑	MIRT363746	miR-138-5p	↓	EIF4EBP1	↑
MIRT726210	miR-30b-5p	↓	TWF1	↑	MIRT005010	miR-125b-5p	↓	PPAT	↑
MIRT502922	miR-15a-5p	↓	CDCA4	↑	MIRT149119	miR-181c-5p	↓	LRRC8D	↑
MIRT028546	miR-30a-5p	↓	LCLAT1	↑	MIRT726441	miR-181a-5p	↓	SSX2IP	↑
MIRT067385	miR-29b-3p	↓	TMTC3	↑	MIRT002082	let-7b-5p	↓	HMGA2	↑
MIRT147275	miR-26b-5p	↓	KPNA2	↑	MIRT193104	let-7g-5p	↓	MAPK6	↑
MIRT000768	miR-34a-5p	↓	CCNE2	↑	MIRT556267	miR-664a-3p	↓	MAPK6	↑
MIRT474332	miR-221-3p	↓	KPNA2	↑	MIRT006679	miR-126-3p	↓	SLC7A5	↑
MIRT025431	miR-34a-5p	↓	LDHA	↑	MIRT732551	miR-490-3p	↓	HMGA2	↑
MIRT149117	miR-181a-5p	↓	LRRC8D	↑	MIRT739340	miR-140-3p	↓	HS6ST2	↑
MIRT297017	let-7i-5p	↓	RRM2	↑	MIRT000965	miR-126-3p	↓	CCNE2	↑
MIRT437797	miR-497-5p	↓	CDC25A	↑	MIRT507483	miR-26b-5p	↓	E2F7	↑
MIRT021534	miR-145-5p	↓	ABRACL	↑	MIRT727588	miR-30c-5p	↓	GCLC	↑
MIRT067975	miR-30a-5p	↓	TFDP1	↑	MIRT020400	miR-29c-3p	↓	TMTC3	↑
MIRT021307	miR-125a-5p	↓	EIF4EBP1	↑	MIRT193101	let-7e-5p	↓	MAPK6	↑
MIRT467971	miR-29b-3p	↓	SLC16A1	↑	MIRT731901	miR-139-5p	↓	TPD52	↑
MIRT149118	miR-181b-5p	↓	LRRC8D	↑	MIRT520906	miR-221-3p	↓	STMN1	↑
MIRT438685	miR-223-3p	↓	PAX6	↑	MIRT023815	miR-1-3p	↓	PDCD10	↑
MIRT502569	miR-195-5p	↓	E2F7	↑	MIRT193071	miR-26b-5p	↓	MAPK6	↑
MIRT003026	miR-29b-3p	↓	DNMT3B	↑	MIRT726356	miR-15a-5p	↓	TBL1XR1	↑
MIRT506855	miR-195-5p	↓	KIF23	↑	MIRT481851	let-7e-5p	↓	AP1S1	↑
MIRT028476	miR-30a-5p	↓	XPO1	↑	MIRT734567	miR-26a-5p	↓	PLOD2	↑
MIRT727590	miR-30a-5p	↓	GCLC	↑	MIRT168581	let-7c-5p	↓	HMGA1	↑
MIRT149112	miR-30a-5p	↓	LRRC8D	↑	MIRT500532	miR-22-3p	↓	YWHAZ	↑
MIRT265078	miR-195-5p	↓	CHEK1	↑	MIRT726354	miR-195-5p	↓	TBL1XR1	↑
MIRT479225	miR-26b-5p	↓	CKS2	↑	MIRT005395	miR-146b-5p	↓	IRAK1	↑
MIRT467970	miR-29c-3p	↓	SLC16A1	↑	MIRT149123	miR-181d-5p	↓	LRRC8D	↑
MIRT045362	miR-185-5p	↓	RAE1	↑	MIRT067977	miR-30d-5p	↓	TFDP1	↑
MIRT006303	miR-30b-5p	↓	CCNE2	↑	MIRT088447	miR-30b-5p	↓	LCLAT1	↑
MIRT193105	let-7i-5p	↓	MAPK6	↑	MIRT626821	miR-4772-3p	↓	PRR11	↑
MIRT054859	miR-497-5p	↓	CHEK1	↑	MIRT149116	miR-30d-5p	↓	LRRC8D	↑
MIRT741394	miR-342-5p	↓	GIN53	↑	MIRT727945	miR-181a-5p	↓	CPOX	↑
MIRT000110	miR-26a-5p	↓	HMGA1	↑	MIRT437628	miR-150-5p	↓	PDIA6	↑
MIRT046007	miR-125b-5p	↓	TDG	↑	MIRT733289	miR-150-5p	↓	PLP2	↑
MIRT149115	miR-30c-5p	↓	LRRC8D	↑	MIRT265076	miR-15a-5p	↓	CHEK1	↑
MIRT000416	let-7a-5p	↓	CDC25A	↑	MIRT081649	miR-497-5p	↓	CCNE1	↑
MIRT003027	miR-29a-3p	↓	DNMT3B	↑	MIRT000397	let-7g-5p	↓	CDC25A	↑
MIRT506857	miR-497-5p	↓	KIF23	↑					

↑ upregulated e ↓ downregulated.

Tabela 40 – Principais vias enriquecidas para DE-miRNAs *down* - LUSC.

<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Proteoglycans in cancer	220	156	45	p>0,01
Cell cycle	129	101	45	p>0,01
Pathways in cancer	555	333	46	p>0,01
Ubiquitin mediated proteolysis	142	107	44	p>0,01
Focal adhesion	213	147	45	p>0,01
Autophagy - animal	146	108	44	p>0,01
FoxO signaling pathway	139	102	45	p>0,01
Regulation of actin cytoskeleton	224	148	44	p>0,01
EGFR tyrosine kinase inhibitor resistance	82	65	45	p>0,01
Adherens junction	79	61	45	p>0,01
Pancreatic cancer	78	60	44	p>0,01
HIF-1 signaling pathway	112	79	44	p>0,01
PI3K-Akt signaling pathway	372	214	45	p>0,01
p53 signaling pathway	75	56	44	p>0,01
Neurotrophin signaling pathway	124	84	45	p>0,01
AMPK signaling pathway	130	87	45	p>0,01
Hepatocellular carcinoma	177	112	45	p>0,01
Small cell lung cancer	100	69	45	p>0,01
Hippo signaling pathway	164	103	44	p>0,01
Transcriptional misregulation in cancer	206	125	45	p>0,01
Tight junction	182	112	45	p>0,01
MAPK signaling pathway	329	187	46	p>0,01
Insulin resistance	124	81	44	p>0,01
TGF-beta signaling pathway	103	69	43	p>0,01
Endometrial cancer	61	45	44	p>0,01
Glioma	79	55	44	p>0,01
mRNA surveillance pathway	108	71	43	p>0,01
Non-small cell lung cancer	81	56	45	p>0,01
mTOR signaling pathway	177	107	44	p>0,01
RNA transport	199	118	44	p>0,01
Apoptosis	151	93	44	p>0,01
Breast cancer	163	95	45	0,001
Wnt signaling pathway	173	98	44	0,002

Tabela 41 – Vias enriquecidas para DEGs *up* - LUSC.

<i>Term</i>	<i>Overlap</i>	<i>Adjusted p value</i>	<i>Genes</i>
Cell cycle	8/124	0,000	DBF4;TFDP1;CCNE2;CCNE1;CHEK1;YWHAZ;CDC25A;MCM2
MicroRNAs in cancer	7/310	0,001	CCNE2;CCNE1;STMN1;DNMT3B;HMGA2;KIF23;CDC25A
p53 signaling pathway	4/73	0,002	RRM2;CCNE2;CCNE1;CHEK1
Cellular senescence	5/156	0,002	CCNE2;CCNE1;CHEK1;EIF4EBP1;CDC25A
Cysteine and methionine metabolism	3/50	0,006	LDHA;GCLC;DNMT3B
Hepatitis B	4/162	0,016	CCNE2;IRAK1;CCNE1;YWHAZ
Small cell lung cancer	3/92	0,027	CCNE2;CCNE1;CKS2
Viral carcinogenesis	4/203	0,027	CCNE2;CCNE1;CHEK1;YWHAZ
Human T-cell leukemia virus 1 infection	4/219	0,032	XPO1;CCNE2;CCNE1;CHEK1
Oocyte meiosis	3/129	0,049	CCNE2;CCNE1;YWHAZ

A fim de compreender a relação funcional entre os 90 DEGs (*ups* e *downs*) e os fenômenos biológicos associados, foram exploradas as interações funcionais das proteínas codificadas por esses genes por meio da construção de uma rede de conectividade, utilizando o banco de dados STRING. A rede de genes associada contém 90 nós e 195 arestas (Figura 40).

A média do coeficiente de clusterização local é de 0,431 e o grau médio é de 4,33. O valor de enriquecimento do PPI foi $p < 0,0001$, indicando que as proteínas estão biológica e significativamente conectadas. A interação com a pontuação combinada mais alta ocorreu

entre as proteínas *MCM2* e *GINS1* (pontuação combinada = 0,999). Destaca-se o *TGFBR2* relacionado ao *QKI*.

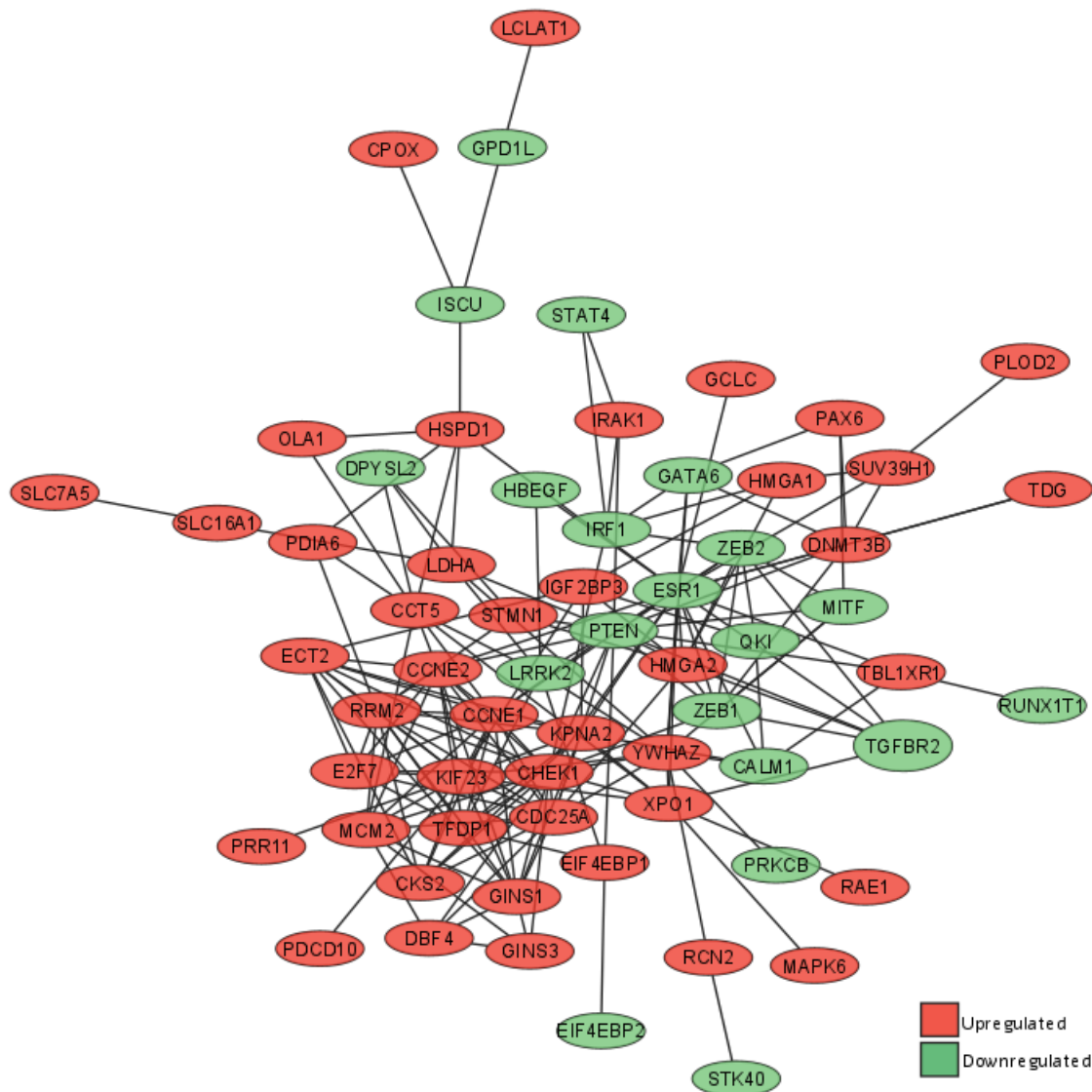


Figura 40 – Rede de interação proteína-proteína - LUSC.

A Tabela 42 apresenta os 30 processos biológicos mais fortes relacionados a PPI (lista completa na Tabela S7), todos considerando $p < 0,05$ como estatisticamente significativo. Os processos biológicos destacados estão predominantemente associados à senescência celular induzida por oncogenes, ao processo mitótico do ciclo celular, à regulação da divisão celular e à regulação positiva do ciclo celular. Estes eventos biológicos são particularmente relevantes no contexto do LUSC em estágio inicial, pois podem desempenhar papéis cruciais na supressão tumoral, na regulação da replicação celular e no controle adequado do ciclo celular.

Tabela 42 – Processos biológicos relacionados a Rede PPI - LUSC.

GO-termo	Descrição	Contagem na rede	Força	FDR
GO:0090402	Oncogene-induced cell senescence	2 de 4	2,04	0,04
GO:0060736	Prostate gland growth	3 de 12	1,74	0,01
GO:1902969	Mitotic DNA replication	3 de 12	1,74	0,01
GO:0006268	DNA unwinding involved in DNA replication	4 de 22	1,60	0,00
GO:0006270	DNA replication initiation	4 de 29	1,48	0,00
GO:0090329	Regulation of DNA-templated DNA replication	5 de 57	1,28	0,00
GO:0000082	G1/S transition of mitotic cell cycle	7 de 86	1,25	0,00
GO:0044772	Mitotic cell cycle phase transition	10 de 173	1,10	0,00
GO:0032465	Regulation of cytokinesis	5 de 94	1,07	0,02
GO:0035265	Organ growth	5 de 98	1,05	0,02
GO:0006260	DNA replication	9 de 203	0,99	0,00
GO:0000079	Regulation of cyclin-dependent protein serine/threonine kinase activity	5 de 111	0,99	0,03
GO:0006261	DNA-templated DNA replication	6 de 137	0,98	0,01
GO:2000278	Regulation of DNA biosynthetic process	5 de 124	0,95	0,04
GO:0032147	Activation of protein kinase activity	7 de 203	0,88	0,01
GO:1903047	Mitotic cell cycle process	18 de 537	0,87	0,00
GO:0051302	Regulation of cell division	6 de 188	0,84	0,04
GO:0071900	Regulation of protein serine/threonine kinase activity	12 de 386	0,83	0,00
GO:0071902	Positive regulation of protein serine/threonine kinase activity	7 de 226	0,83	0,02
GO:0045787	Positive regulation of cell cycle	10 de 349	0,80	0,00
GO:0090068	Positive regulation of cell cycle process	7 de 251	0,79	0,03
GO:0022402	Cell cycle process	21 de 835	0,74	0,00
GO:0045860	Positive regulation de protein kinase activity	10 de 416	0,72	0,01
GO:0045859	Regulation of protein kinase activity	15 de 663	0,69	0,00
GO:0051347	Positive regulation of transferase activity	13 de 586	0,69	0,00
GO:0048589	Developmental growth	9 de 412	0,68	0,02
GO:0007049	Cell cycle	25 de 1246	0,64	0,00
GO:0010564	Regulation od cell cycle process	14 de 716	0,63	0,00
GO:0031667	Response to nutrient levels	9 de 461	0,63	0,04
GO:0051726	Regulation of cell cycle	21 de 1108	0,62	0,00

Análises estatísticas

Foram geradas árvores de classificação, floresta aleatória e árvores de sobrevivência visando identificar potenciais DE-miRNAs e DEGs de relevância significativa na distinção de T vs. N, bem como na análise da sobrevida dos pacientes diagnosticados com LUSC precocemente.

O banco de dados combinado, contendo miRNA-seq e RNA-seq ($n = 244$), foi filtrado para incluir apenas os 64 DE-miRNAs e 90 DEGs (sem repetições) validados experimentalmente. As análises foram realizadas separadamente para miRNA-seq e RNA-seq, independentemente da direção da expressão. Neste contexto relacionado à doença LUSC, não foram identificadas variáveis clínicas significativas. Portanto, nenhuma dessas variáveis foi considerada nesta etapa das análises.

Para conduzir as análises de árvore de classificação e Florestas Aleatórias (construído a fim de complementar os resultados obtidos na árvore), buscando a distinção entre T vs. N, o banco de dados foi dividido de maneira aleatória, reservando 80% ($n = 195$) para o treinamento e 20% ($n = 49$) para o teste. Isso possibilitou a avaliação do desempenho do banco de treino em novos conjuntos de dados (banco de teste). A fim de lidar com o desequilíbrio de classes nos conjuntos de treinamento, foi realizada uma subamostragem nos bancos de dados.

Árvore de classificação

Foram realizadas simulações com 100 mil árvores de classificação tanto para miRNA-seq quanto para RNA-seq, visando identificar a árvore de classificação que apresentasse as maiores acurácias. Para essas árvores, foram considerados os hiperparâmetros: *minsplit*= 2, *minbucket*= 1 e *cp*= 0,00001.

A Figura 41 exibe as acurácias obtidas nessas simulações com os miRNAs, a grande maioria dos valores estão acima de 90% e a média ficou em torno de 94% de acurácia.

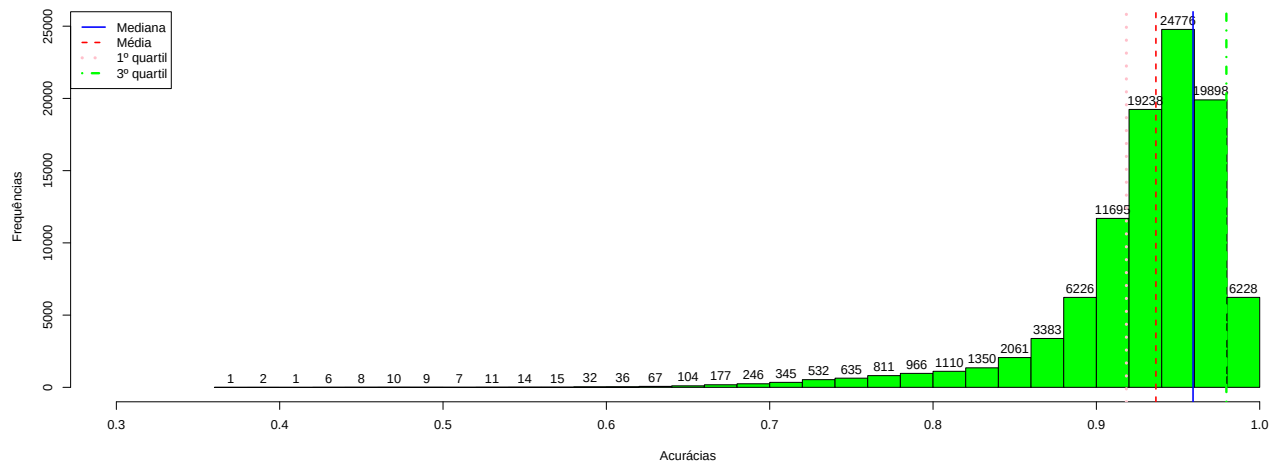


Figura 41 – Histograma das 100 mil acurácias para árvores de classificação dos DE-miRNAs - LUSC.

A Figura 42a ilustra a árvore de classificação dos miRNAs, apresentando uma acurácia de 100% ao ser testada no banco de dados de teste. A árvore inclui apenas o miR-30d-5p, o qual $\text{miR-30d-5p} \geq 57127$ as observações são classificadas como tecido normal, caso contrário, são classificadas como tumor primário (azul). As variáveis mais importantes para o modelo de árvore de classificação estão ilustradas na Figura 42b. O miR-30d-5p e miR-30a-5p apresentaram a mesma importância.

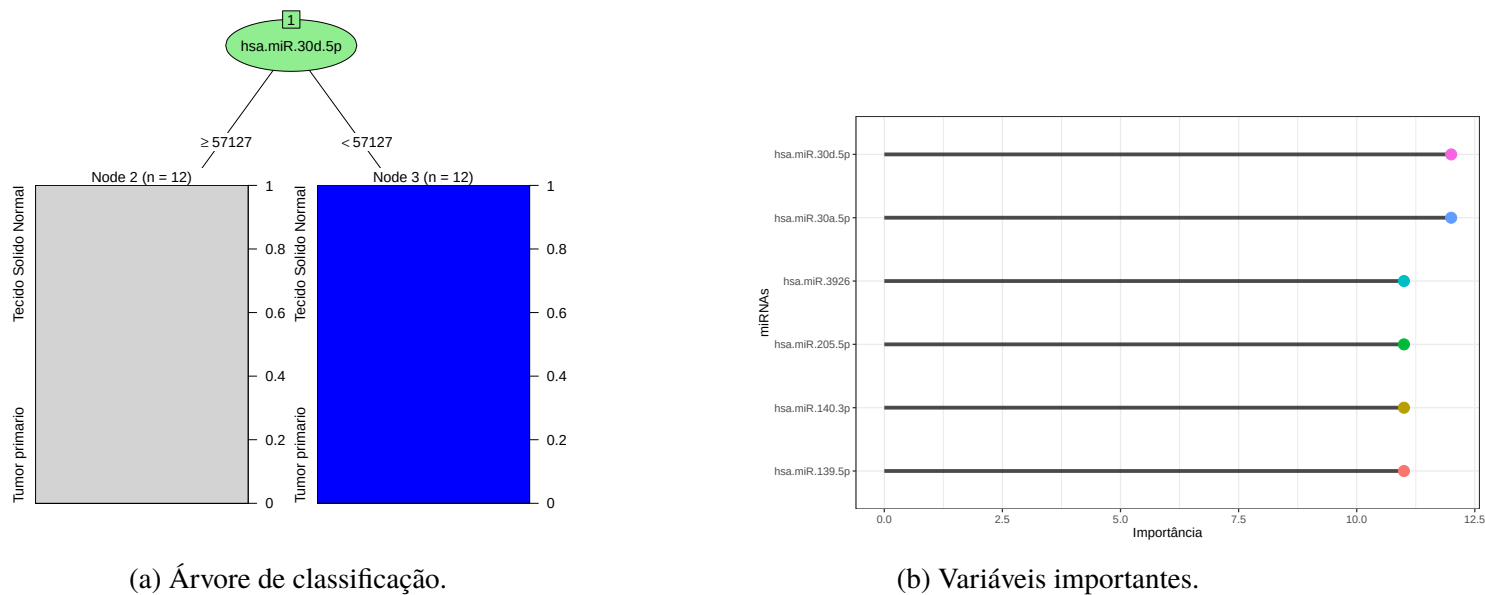


Figura 42 – Árvore de classificação e variáveis importantes para os DE-miRNAs - LUSC.

A Figura 43 apresenta as acurácias derivadas do conjunto de DEGs, o primeiro quartil está em aproximadamente 92% e 11.695 observações estão entre 98 a 100 de acurácia.

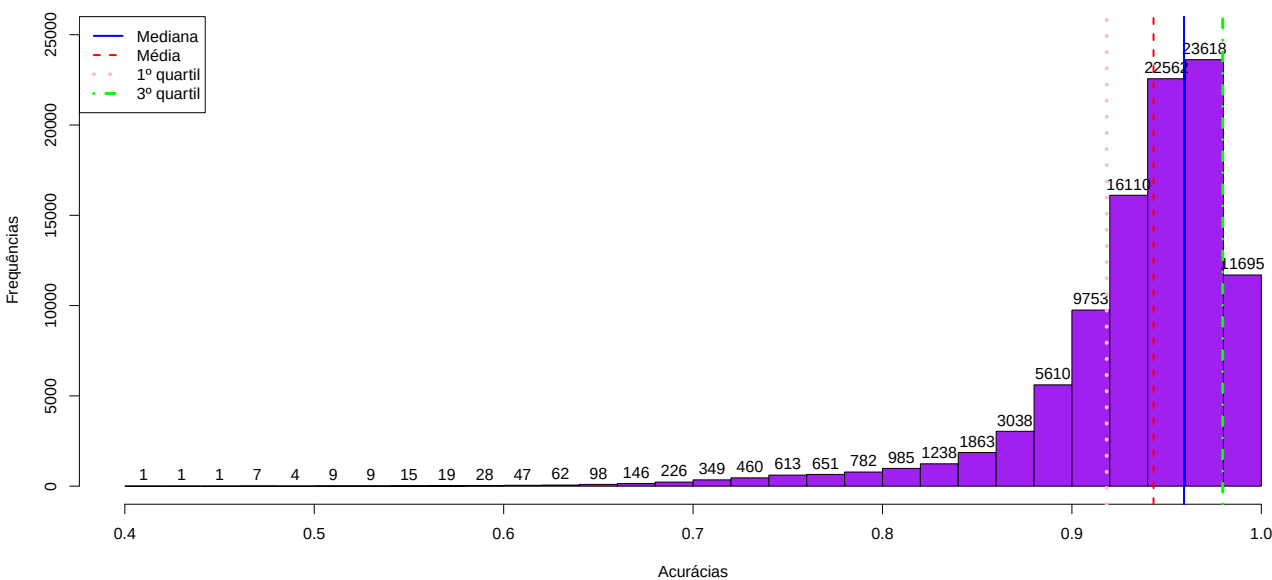
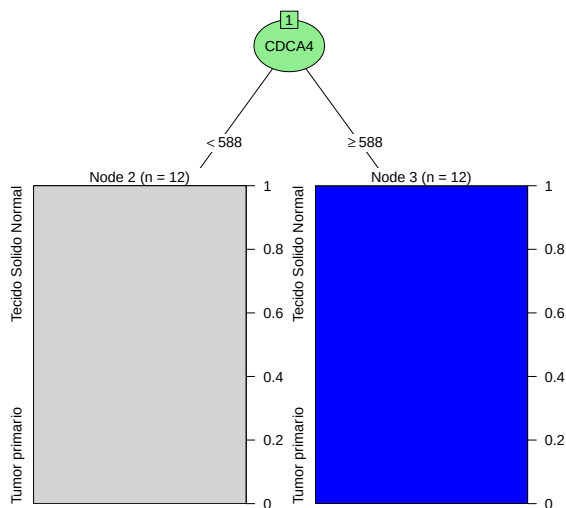


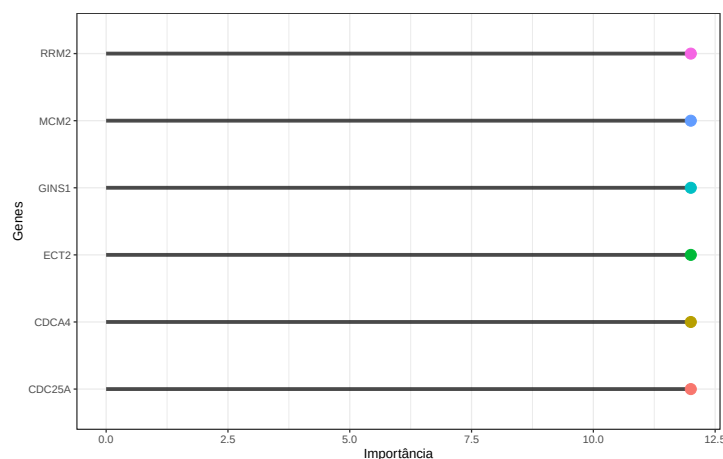
Figura 43 – Histograma das 100 mil acurácias para árvores de classificação dos DEGs - LUSC.

A Figura 44a exibe a árvore de classificação formada pelos DEGs, alcançando uma acurácia de 100%. Essa árvore é composta apenas pelo *CDCA4* (*Cell Division Cycle Associated 4*), que quando $CDCA4 < 588$ é classificado como tecido normal e, caso contrário, como tumor primário (azul).

A Figura 44b apresenta os genes mais importantes para o modelo de árvore de classificação na separação dos tecidos (T vs. N), vale ressaltar que os 5 genes: *RRM2*, *MCM2*, *GINS1*, *ECT2*, *CDCA4* e *CDC25A* possuem a mesma importância.



(a) Árvore de classificação.



(b) Variáveis importantes.

Figura 44 – Árvore de classificação e variáveis importantes para os DEGs - LUSC.

Florestas Aleatórias

A análise do modelo de Florestas Aleatórias foi conduzida usando os DE-miRNAs ou DEGs para distinguir entre T vs. N. Foram geradas 5000 árvores de classificação e o modelo identificou os 10 DE-miRNAs e DEGs mais importantes para essa diferenciação, conforme representado nas Figuras 45a e 46a. Os DE-miRNAs de maior importância são: miR-30d-5p, mesmo que a árvore de classificação identificou, miR-30a-5p seguido do miR-210-3p. Os DEGs mais relevantes são: CDCA4, mesmo identificado pela árvore, além dos *TGFBR2* e *GINS1*.

Os erros dos modelos do Florestas Aleatórias estão representados nas Figuras 45b e 46b, onde os pontos vermelhos indicam o tumor primário e os pontos verdes, o tecido normal. É notável que os erros oscilaram entre 0 e 0,25, para os DE-miRNAs, eles atingem uma estabilização após 3.000 árvores, enquanto para os DEGs isso ocorre após 1.000 árvores.

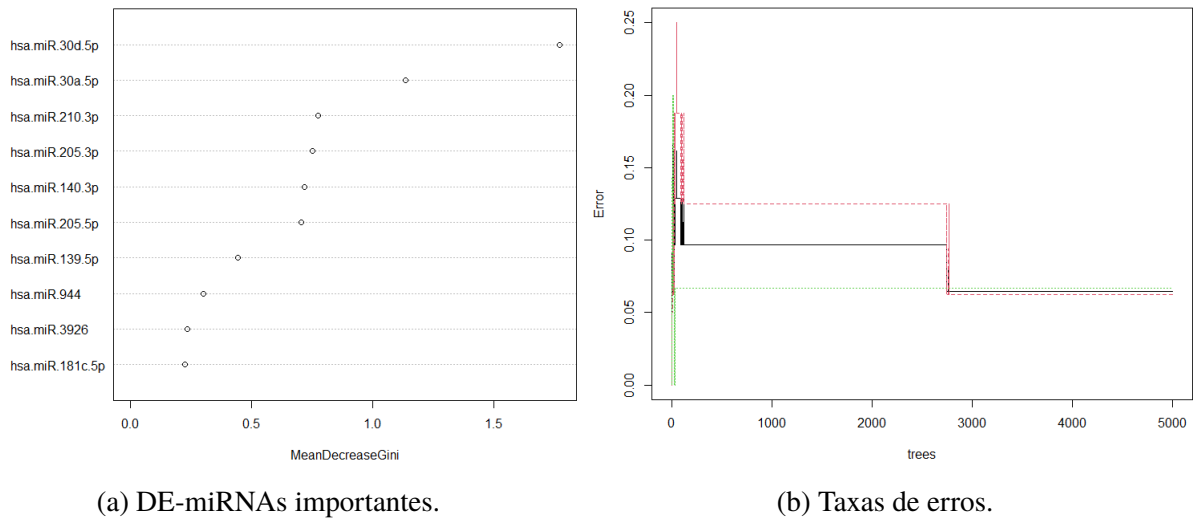


Figura 45 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DE-miRNAs - LUSC.

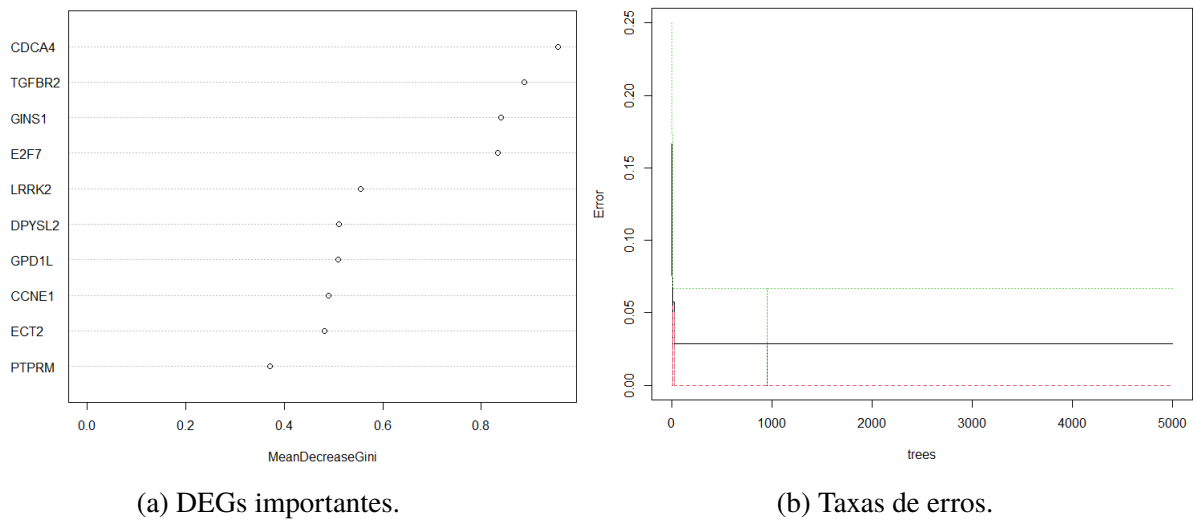


Figura 46 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DEGs - LUSC.

Árvore de sobrevivência

Foram criadas árvores de sobrevivência com o propósito de identificar os DE-miRNAs e DEGs associados à sobrevivência de pacientes em estágios iniciais de LUSC. A variável temporal utilizada na construção da árvore foi o tempo em dias até o último acompanhamento do paciente, e a variável de censura representou o *status* vital do paciente: se faleceu devido ao LUSC, *status* = 1 (composto por 97 casos); caso contrário, *status* = 0 (composto por 147 casos).

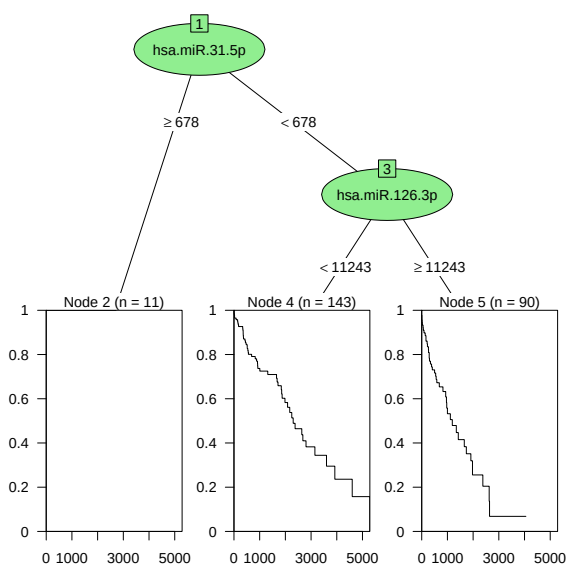
A árvore de sobrevivência construída para o conjunto de DE-miRNAs revelou-se complexa, apresentando 20 nós, conforme evidenciado na coluna *nsplit* da Tabela 43. Diante dessa complexidade, torna-se necessário realizar uma pós-poda. Para isso, foi identificado o menor erro

de validação cruzada, que é 1, 175, e, por conseguinte, determinar o valor de corte $cp = 0,036$. Esse procedimento resultará em uma árvore simplificada, composta por apenas dois nós.

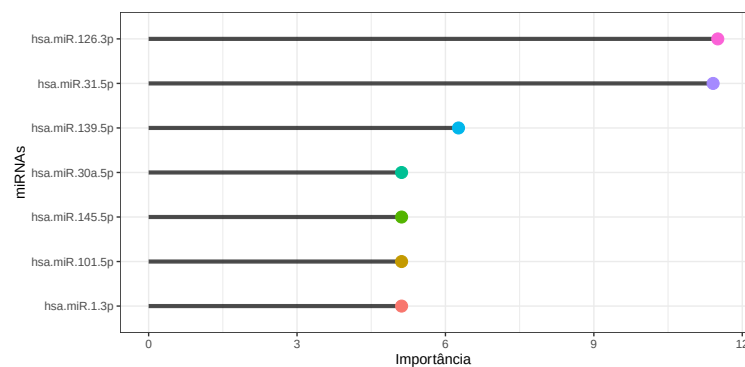
Tabela 43 – Resumo do modelo da árvore de sobrevivência para os DE-miRNAs - LUSC.

CP	Nº de divisões	Erro relativo	Erro de validação cruzada	Erro padrão
0,037	0	1,000	1,008	0,072
0,036	2	0,926	1,175	0,087
0,036	5	0,817	1,175	0,087
0,022	7	0,744	1,195	0,095
0,020	10	0,678	1,342	0,109
0,020	12	0,639	1,399	0,113
0,018	13	0,619	1,408	0,115
0,018	14	0,601	1,404	0,115
0,017	15	0,583	1,417	0,116
0,017	16	0,566	1,416	0,116
0,013	18	0,533	1,428	0,119
0,011	19	0,520	1,430	0,121
0,010	20	0,509	1,430	0,121

A árvore de sobrevivência pós-poda está apresentada na Figura 47a, identificou 2 miRNAs: miR-31-5p e miR-126-3p como importantes para a sobrevivência dos pacientes. As observações que possuem $\text{miR-31-5p} \geq 678$ vão para o nó 2 e possui grande chance de sobrevivência, caso contrário seguem para: $\text{miR-126-3p} < 11243$ os pacientes possuem 50% de chance próxima aos 3.000 dias e caso contrário é o pior cenário e o menor tempo de sobrevivência. A Figura 47b mostra os mesmos DE-miRNAs com mais importância: miR-126-3p e miR-31-5p.



(a) Árvore de sobrevivência pós-poda.



(b) Variáveis importantes.

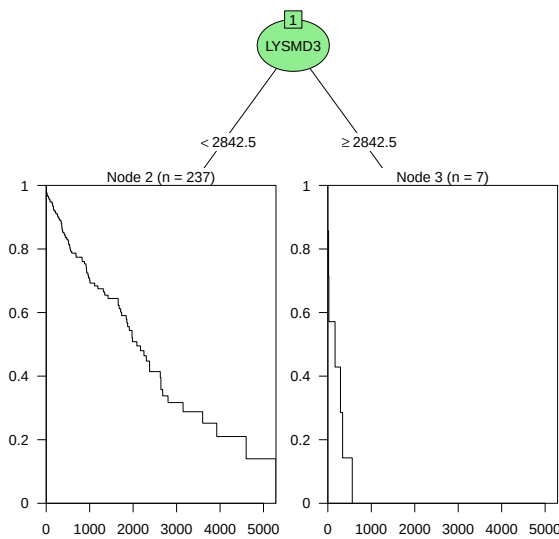
Figura 47 – Árvore de sobrevivência pós-poda e variáveis importantes para os DE-miRNAs - LUSC.

A árvore de sobrevivência gerada para o conjunto de DEGs apresentou uma complexidade notável, com 18 nós, conforme demonstrado na Tabela 44. Diante dessa complexidade, tornou-se crucial realizar a pós-poda da árvore. Nesse processo, identificou-se o menor erro de validação cruzada, resultando em um valor de $cp = 0,049$ para a poda, a árvore gerada pós-poda possui apenas 1 nó.

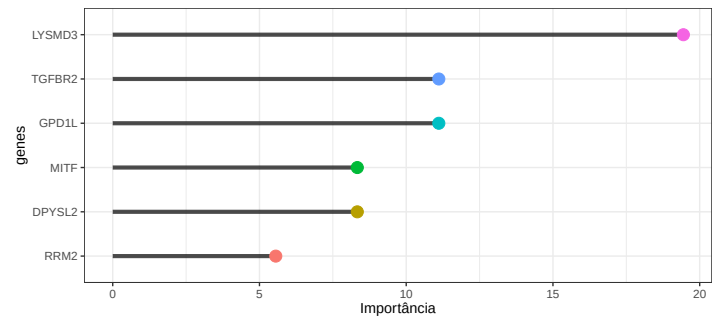
Tabela 44 – Resumo do modelo da árvore de sobrevivência para os DEGs - LUSC.

CP	Nº de divisões	Erro relativo	Erro de validação cruzada	Erro padrão
0,057	0	1,000	1,006	0,072
0,049	1	0,943	1,048	0,082
0,041	2	0,893	1,067	0,086
0,031	4	0,811	1,084	0,088
0,030	6	0,750	1,198	0,099
0,029	7	0,720	1,203	0,099
0,027	10	0,630	1,220	0,099
0,026	11	0,603	1,259	0,102
0,023	12	0,577	1,339	0,112
0,022	13	0,554	1,341	0,113
0,022	14	0,533	1,342	0,113
0,019	15	0,511	1,338	0,113
0,015	16	0,492	1,348	0,115
0,015	17	0,477	1,353	0,114
0,010	18	0,462	1,360	0,115

A árvore de sobrevivência pós-poda, Figura 48a, identificou o LYSDMD3 (*LysM Domain Containing 3*) como importante para a sobrevivência dos pacientes. As observações que possuem $LYSDMD3 < 2842.5$ seguem para a esquerda e possuem mais tempo de sobrevivência e caso contrário vão para a direita e as estimativas da curva de Kaplan-Meier indicam sobrevivência com menos de 1000 dias de vida. A Figura 48b os DEGs mais importantes para a sobrevivência, sendo LYSDMD3 seguido do *TGFBR2* e *GPD1L*.



(a) Árvore de sobrevivência pós-poda.



(b) Variáveis importantes.

Figura 48 – Árvore de sobrevivência pós-poda e variáveis mais importantes para os RNAs - LUSC.

Discussões

Após conduzir análises descritivas, biológicas e estatísticas nos bancos de dados de miRNA-seq e RNA-seq, foi possível examinar e reconhecer padrões e comportamentos relevantes. O propósito inicial dessas análises era compreender o perfil das amostras de pacientes presentes nos bancos de dados. Posteriormente, a atenção foi direcionada para a identificação de DE-miRNAs e DEGs, desempenhando papéis cruciais na diferenciação entre tecido T vs. N e na sobrevivência de pacientes com LUSC em estágios precoces. As análises descritivas das variáveis clínicas não revelaram diferenças significativas, considerando um nível de significância de $p < 0,05$, na diferenciação de T vs. N.

A análise de expressão diferencial utilizando o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni foi realizada visando manter os dados originais e assim utilizar estatísticas não-paramétricas, diferenciando das estratégias convencionais. Foram identificados 337 DE-miRNAs e os 9.877 DEGS.

Depois de identificar os DE-miRNAs e DEGs, as análises biológicas resultaram na identificação de 161 pares potenciais (candidatos a DE-miRNA - DEG) que exibem expressões opostas, com correlações negativas significativas sendo validadas experimentalmente.

Ao analisar o conjunto de 46 pares de DE-miRNAs *up* e DEGs *down* a rede regulatória (Figura 38) destacou o miR-205-5p, com 8 ligações. [Lebanony et al. \(2009\)](#) concluem que o miR-205 é um marcador altamente preciso para LUSC, durante o ensaio de diagnóstico padronizado no estudo, podendo fornecer uma subclassificação altamente precisa de pacientes com LUSC.

Outro destaque foi o miR-210-3p. [Chen et al. \(2021\)](#) apontam que o miR-210-3p é regulado positivamente em tecidos de câncer de pulmão e consideram ser um biomarcador promissor para o tratamento do câncer de pulmão.

O *TGFBR2* apareceu relacionado ao miR-130b-3p e ao miR-301-3p. Os genes *ZEB1* (*Zinc Finger E-Box Binding Homeobox 1*) e *ZEB2* (*Zinc Finger E-Box Binding Homeobox 2*) também apareceram e são fatores de transcrição que desempenham papéis importantes em diversos processos biológicos, incluindo a regulação da transição epitélio-mesenquimal (EMT).

Outro gene em destaque foi o *PTEN*, relacionado ao miR-130b-3p, miR-141-3p e STAT4, atua como um supressor de tumor, regulando negativamente a via de sinalização PI3K/AKT/mTOR, que desempenha um papel crucial no controle do crescimento celular, sobrevivência e metabolismo ([LIM; CROWE; YANG, 2015](#)). As vias *PI3K-Akt signaling pathway* e *mTOR signaling pathway* foram identificadas (Tabela 37).

Além dessas importantes identificações que a rede regulatória proporcionou, foram encontradas, estatisticamente significativas, vias enriquecidas relacionadas ao LUSC identificadas com os DE-miRNAs e DEGs, os principais processos relacionados a algumas das vias são: envolvimento na adesão celular, regulação de genes envolvidos em apoptose, interação célula-matriz extracelular, regulação do ciclo celular, organização do citoesqueleto celular, diversas vias associadas ao câncer, sobrevivência celular, entre outras.

Os 115 pares de DE-miRNAs *down* e DEGs *up* estão apresentados na Tabela 40. A rede regulatória gerada (Figura 39) ficou mais complexa, contudo o destaque maior foi a família da miR-30 e miR-29.

O trabalho de [Kanthaje et al. \(2021\)](#) trás uma vasta revisão, apresentando uma compreensão mais profunda do papel regulador e do significado clínico do miR-30, enfatizamos os papéis emergentes no câncer de pulmão. Eles relatam que a família regula várias vias de sinalização imperativas como p53, PI3K/AKT, resultando na modulação de eventos carcinogênicos importantes envolvendo proliferação celular, apoptose, metástase, transição epitelial-mesenquimal e resistência a medicamentos.

[Tan, Wu e Cai \(2013\)](#) destacam que a expressão de miR-29a e miR-29b foi parcialmente regulada por DNMT3A e DNMT3B num ciclo de feedback positivo. [Chen et al. \(2018\)](#) destacam um circuito regulador TP53/ miR-29s/SETDB1 e atribuem um papel de regulador de metilação H3K9 aos miR-29s, que pode ser um potencial alvo terapêutico no tratamento de NSCLC.

A rede de PPI (Figura 40) formada por todos os 90 DEGs identificados nos pares, mostrou valor de enriquecimento significativo, $p < 0,0001$, afirmando que as proteínas identificadas pelos DEGs estão biológica e significativamente conectadas. Na rede o *TGFBR2*, *QKI*, *ZEB1*, *ZEB2* aparecem todos relacionados.

Foram conduzidas análises estatísticas nos conjuntos de DE-miRNAs e DEGs, previamente validados experimentalmente, consistindo em 64 DE-miRNAs e 90 DEGs (sem duplicatas),

independentemente da direção da expressão.

O miR-30d-5p foi identificado como um destaque tanto na árvore de classificação quanto na análise de Florestas Aleatórias, sendo considerado o mais relevante na diferenciação entre tecido T vs. N. Esse miRNA também foi observado na rede regulatória apresentada na Figura 39.

Em relação à sobrevivência dos pacientes, o miR-31-5p e miR-126-3p foram os mais importantes relacionados com LUSC precoce. Chi *et al.* (2019) destacam que a expressão é mais elevada de miR-31-5p em pacientes com LUSC e concluem que esse miRNA serve como um biomarcador molecular para LUSC no ambiente clínico.

Em relação aos DEGs, foi identificado na árvore de classificação que o *CDCA4* diferenciaram os tecidos T vs. N, esse DEG também foi identificado em Florestas Aleatórias como o mais importante. O trabalho de Hu *et al.* (2021) relatam que ao visar o *CDCA4*, o miR-497-5p em baixo nível restringiu o desenvolvimento do LUSC, a rede da Figura 39 encontrou essa relação DE-miRNA e DEG. Para a sobrevivência dos pacientes, *LYSMD3* (*LysM Domain Containing 3*) foi o mais importante.

Em síntese, estas análises integrativas proporcionaram *insights* valiosos sobre os mecanismos moleculares subjacentes ao LUSC, destacando potenciais biomarcadores e alvos terapêuticos. Estas descobertas contribuem para a compreensão mais profunda da biologia de LUSC, apontando para direções promissoras para futuras pesquisas e aplicações clínicas.

4.2.3 Análise de expressão diferencial, biológica e estatística de LUSC vs. LUAD

Nesta parte do estudo, utilizaram-se os bancos de dados miRNA-seq e RNA-seq de LUAD-LUSC a fim de distinguir a doença em estágio inicial no qual os pacientes se encontram (LUSC vs. LUAD). O banco de dados LUAD-LUSC miRNA-seq contém um total de 496 amostras de pacientes, sendo que 54% são classificadas como LUAD e 46% como LUSC. Este banco de dados possui 988 perfis de expressão de miRNA-seq. Já o banco de dados LUAD-LUSC RNA-seq é composto por 511 amostras de pacientes, com 53% representando o subtipo LUAD e 47% o subtipo LUSC. Este banco de dados possui 39.275 perfis de expressão de RNA-seq. Em ambos os casos as variáveis clínicas não foram analisadas.

Análise de expressão diferencial

A análise de expressão diferencial foi conduzida de forma independente para cada miRNA e RNA, visando identificar DE-miRNAs e DEGs na distinção de LUSC vs. LUAD, por meio do teste de Wilcoxon-Mann-Whitney. A expressão das medianas dos miRNAs e RNAs significativos foram avaliadas seguindo a regra: se $LUSC > LUAD$ então é classificado como *upregulated* e caso contrário como *downregulated*.

A análise foi realizada no banco de dados LUAD-LUSC miRNA-seq, considerando critério de significância estatística de $p < 0,05$. Posteriormente, a correção de Bonferroni foi aplicada, assim, identificou-se o valor de $\alpha_{\text{corrigido}}$:

$$\alpha_{\text{corrigido}} = \frac{\alpha_{\text{original}}}{n} = \frac{0,05}{988} \approx 0,000051.$$

Considerando $p < 0,000051$, identificou-se 301 DE-miRNAs, sendo 85 *upregulated* e 216 *downregulated*, estão todos apresentados nas Tabelas 45 e 46, respectivamente.

Tabela 45 – DE-miRNAs *upregulated* (↑) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD-LUSC.

miRNA	Mediana		valor p	miRNA	Mediana		valor p	miRNA	Mediana		valor p
	LUAD	LUSC			LUAD	LUSC			LUAD	LUSC	
miR-205-3p	↑ 0,0	19,0	$p < 0,000051$	miR-6720-3p	↑ 3,0	9,0	$p < 0,000051$	miR-3651	↑ 2,0	4,0	$p < 0,000051$
miR-205-5p	↑ 317,5	40707,0	$p < 0,000051$	miR-31-5p	↑ 9,5	42,0	$p < 0,000051$	miR-18a-5p	↑ 42,0	64,5	$p < 0,000051$
miR-944	↑ 5,0	378,0	$p < 0,000051$	miR-31-3p	↑ 5,0	28,0	$p < 0,000051$	miR-3928-3p	↑ 3,0	5,5	$p < 0,000051$
miR-196b-5p	↑ 71,5	1044,0	$p < 0,000051$	miR-455-3p	↑ 512,0	891,0	$p < 0,000051$	miR-942-5p	↑ 20,5	30,0	$p < 0,000051$
miR-149-5p	↑ 99,0	519,5	$p < 0,000051$	miR-203b-3p	↑ 155,0	283,0	$p < 0,000051$	miR-296-3p	↑ 2,0	5,0	$p < 0,000051$
miR-1910-5p	↑ 0,0	2,0	$p < 0,000051$	miR-301b-5p	↑ 2,0	4,0	$p < 0,000051$	miR-301a-5p	↑ 16,0	25,0	$p < 0,000051$
miR-203a-3p	↑ 16536,5	68021,5	$p < 0,000051$	miR-767-3p	↑ 0,0	1,0	$p < 0,000051$	miR-539-5p	↑ 8,0	12,0	$p < 0,000051$
miR-6510-3p	↑ 4,0	32,0	$p < 0,000051$	miR-1269b	↑ 0,0	1,0	$p < 0,000051$	miR-651-5p	↑ 11,0	18,0	$p < 0,000051$
miR-708-5p	↑ 131,3	343,5	$p < 0,000051$	miR-7974	↑ 0,0	1,0	$p < 0,000051$	miR-675-3p	↑ 17,0	30,0	$p < 0,000051$
miR-4652-5p	↑ 1,0	13,0	$p < 0,000051$	miR-149-3p	↑ 0,0	1,0	$p < 0,000051$	miR-4746-5p	↑ 7,0	11,0	$p < 0,000051$
miR-767-5p	↑ 1,0	49,0	$p < 0,000051$	miR-4491	↑ 2,0	4,0	$p < 0,000051$	miR-7112-3p	↑ 0,0	1,0	$p < 0,000051$
miR-708-3p	↑ 313,5	749,0	$p < 0,000051$	miR-629-3p	↑ 18,0	30,0	$p < 0,000051$	miR-23a-5p	↑ 6,0	8,0	$p < 0,000051$
miR-105-5p	↑ 2,0	92,5	$p < 0,000051$	miR-6783-3p	↑ 0,0	1,0	$p < 0,000051$	let-7c-3p	↑ 24,5	36,5	$p < 0,000051$
miR-3662	↑ 0,0	3,0	$p < 0,000051$	miR-584-5p	↑ 85,0	153,0	$p < 0,000051$	miR-323b-3p	↑ 5,0	10,0	$p < 0,000051$
miR-5579-3p	↑ 0,0	2,0	$p < 0,000051$	miR-15b-3p	↑ 102,0	154,0	$p < 0,000051$	miR-2277-5p	↑ 3,0	4,0	$p < 0,000051$
miR-1269a	↑ 10,0	589,5	$p < 0,000051$	miR-577	↑ 11,0	28,0	$p < 0,000051$	miR-106a-5p	↑ 43,0	62,0	$p < 0,000051$
miR-203a-5p	↑ 1,0	5,0	$p < 0,000051$	miR-378a-3p	↑ 1075,0	1617,0	$p < 0,000051$	miR-17-5p	↑ 1416,0	1870,5	$p < 0,000051$
miR-4797-3p	↑ 1,0	3,0	$p < 0,000051$	miR-20a-3p	↑ 19,0	28,0	$p < 0,000051$	miR-16-2-3p	↑ 32,5	42,5	$p < 0,000051$
miR-33a-5p	↑ 62,5	128,0	$p < 0,000051$	miR-590-5p	↑ 45,0	68,5	$p < 0,000051$	miR-1911-5p	↑ 1,0	2,0	$p < 0,000051$
miR-1271-5p	↑ 7,5	14,0	$p < 0,000051$	miR-1305	↑ 0,0	0,5	$p < 0,000051$	miR-134-5p	↑ 479,0	669,5	$p < 0,000051$
miR-9-5p	↑ 2397,0	9274,5	$p < 0,000051$	miR-378a-5p	↑ 173,0	295,5	$p < 0,000051$	miR-579-5p	↑ 2,0	3,0	$p < 0,000051$
miR-4662a-5p	↑ 41,5	84,5	$p < 0,000051$	miR-512-3p	↑ 0,0	0,5	$p < 0,000051$	miR-509-3p	↑ 12,0	20,5	$p < 0,000051$
miR-744-3p	↑ 40,5	78,0	$p < 0,000051$	miR-152-5p	↑ 2,0	3,0	$p < 0,000051$	miR-127-5p	↑ 375,0	526,5	$p < 0,000051$
miR-4766-3p	↑ 1,0	2,0	$p < 0,000051$	miR-1295a	↑ 2,0	4,0	$p < 0,000051$	miR-323a-3p	↑ 4,0	7,0	$p < 0,000051$
miR-452-3p	↑ 12,5	27,0	$p < 0,000051$	miR-548f-3p	↑ 0,0	1,0	$p < 0,000051$	miR-93-5p	↑ 16511,0	22278,5	$p < 0,000051$
miR-224-5p	↑ 128,5	356,0	$p < 0,000051$	miR-873-5p	↑ 0,0	1,0	$p < 0,000051$	miR-3653-3p	↑ 11,0	16,0	$p < 0,000051$
miR-210-3p	↑ 3517,0	6821,0	$p < 0,000051$	miR-769-5p	↑ 63,5	88,0	$p < 0,000051$	miR-452-5p	↑ 282,5	424,0	$p < 0,000051$
miR-1248	↑ 1,0	2,0	$p < 0,000051$	miR-570-3p	↑ 1,0	3,0	$p < 0,000051$				
miR-4665-5p	↑ 0,0	1,0	$p < 0,000051$	miR-4668-3p	↑ 7,0	10,0	$p < 0,000051$				

Tabela 46 – DE-miRNAs *downregulated* (↓) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD-LUSC.

miRNA	Mediana		valor <i>p</i>	miRNA	Mediana		valor <i>p</i>	miRNA	Mediana		valor <i>p</i>
	LUAD	LUSC			LUAD	LUSC			LUAD	LUSC	
miR-375-3p	↓ 69901,0	4683	p<0,000051	miR-135b-3p	↓ 7,0	3	p<0,000051	miR-6715a-3p	↓ 0,0	0	p<0,000051
miR-326	↓ 68,0	11	p<0,000051	miR-542-5p	↓ 8,0	4	p<0,000051	miR-133b	↓ 3,0	1	p<0,000051
miR-5698	↓ 11,0	2	p<0,000051	miR-4727-5p	↓ 1,0	0	p<0,000051	miR-625-3p	↓ 988,0	719,5	p<0,000051
miR-29a-3p	↓ 49404,0	17045,5	p<0,000051	miR-29a-5p	↓ 38,0	24	p<0,000051	miR-3934-3p	↓ 11,0	7	p<0,000051
miR-30d-5p	↓ 47957,0	16821,5	p<0,000051	miR-5010-3p	↓ 4,0	2	p<0,000051	miR-101-5p	↓ 25,0	18	p<0,000051
miR-3189-3p	↓ 2,0	0	p<0,000051	miR-99b-5p	↓ 94666,5	64000,5	p<0,000051	miR-6808-3p	↓ 1,0	1	p<0,000051
miR-4728-3p	↓ 8,0	2	p<0,000051	miR-652-3p	↓ 165,0	94,5	p<0,000051	miR-7846-3p	↓ 0,8	0	p<0,000051
miR-3617-3p	↓ 1,0	0	p<0,000051	miR-937-3p	↓ 13,0	5	p<0,000051	miR-4802-3p	↓ 0,0	0	p<0,000051
miR-4709-3p	↓ 10,0	3	p<0,000051	miR-99b-3p	↓ 349,5	189	p<0,000051	miR-147b-3p	↓ 4,0	2	p<0,000051
miR-1287-5p	↓ 180,0	69	p<0,000051	miR-372-3p	↓ 1,0	0	p<0,000051	miR-100-3p	↓ 1,0	0	p<0,000051
miR-1266-5p	↓ 53,0	17	p<0,000051	miR-140-3p	↓ 3993,5	2103	p<0,000051	miR-186-5p	↓ 1254,0	905,5	p<0,000051
miR-26a-5p	↓ 11876,5	5802,5	p<0,000051	miR-4775	↓ 0,0	0	p<0,000051	miR-217-5p	↓ 77,5	47,5	p<0,000051
let-7g-5p	↓ 2600,5	1348,5	p<0,000051	miR-5091	↓ 1,0	1	p<0,000051	miR-218-1-3p	↓ 4,0	2,0	p<0,000051
miR-151b	↓ 9,0	3	p<0,000051	miR-425-3p	↓ 42,5	25	p<0,000051	miR-624-5p	↓ 7,0	5,0	p<0,000051
miR-92b-5p	↓ 3,0	1	p<0,000051	miR-151a-3p	↓ 10917,5	7350	p<0,000051	miR-4802-5p	↓ 0,0	0,0	p<0,000051
miR-181b-5p	↓ 2734,0	1172,5	p<0,000051	miR-32-5p	↓ 141,0	83,5	p<0,000051	miR-502-3p	↓ 42,0	31,0	p<0,000051
let-7e-5p	↓ 5333,0	1999	p<0,000051	miR-101-3p	↓ 1,0	0	p<0,000051	miR-2355-3p	↓ 19,5	14,0	p<0,000051
miR-4724-5p	↓ 3,0	1	p<0,000051	miR-424-3p	↓ 10,0	5	p<0,000051	miR-181b-3p	↓ 91,0	59,0	p<0,000051
miR-340-5p	↓ 71,5	30	p<0,000051	miR-6810-3p	↓ 0,0	0	p<0,000051	miR-362-3p	↓ 10,0	6,0	p<0,000051
miR-30b-5p	↓ 2454,5	1011	p<0,000051	miR-3064-3p	↓ 1,0	0	p<0,000051	miR-489-3p	↓ 1,0	0,0	p<0,000051
miR-30d-3p	↓ 31,5	15	p<0,000051	miR-101-3p	↓ 47794,0	25732,5	p<0,000051	miR-30c-2-3p	↓ 106,0	61,5	p<0,000051
miR-139-3p	↓ 42,0	12,5	p<0,000051	miR-653-3p	↓ 2,0	1	p<0,000051	miR-3191-5p	↓ 0,0	0,0	p<0,000051
miR-30b-3p	↓ 15,0	7	p<0,000051	miR-6892-5p	↓ 8,0	5	p<0,000051	miR-30a-3p	↓ 14830,5	8402,0	p<0,000051
miR-181a-5p	↓ 10964,5	5225,5	p<0,000051	miR-664b-5p	↓ 2,0	1	p<0,000051	miR-3127-5p	↓ 21,0	14,0	p<0,000051
miR-34a-5p	↓ 958,5	526	p<0,000051	miR-5680	↓ 1,0	0	p<0,000051	miR-3934-5p	↓ 7,0	4,0	p<0,000051
miR-92b-3p	↓ 759,5	344	p<0,000051	miR-181a-2-3p	↓ 3984,0	2713	p<0,000051	miR-132-3p	↓ 425,0	277,0	p<0,000051
miR-548b-3p	↓ 9,5	4	p<0,000051	miR-330-3p	↓ 18,0	11	p<0,000051	miR-4529-3p	↓ 0,0	0,0	p<0,000051
miR-664b-3p	↓ 19,0	10	p<0,000051	miR-95-3p	↓ 12,0	7	p<0,000051	miR-6514-5p	↓ 1,0	0,0	p<0,000051
miR-338-5p	↓ 29,0	12	p<0,000051	miR-5090	↓ 1,0	0	p<0,000051	miR-30a-5p	↓ 55658,0	33284,5	p<0,000051
miR-3065-3p	↓ 360,0	138	p<0,000051	miR-34a-3p	↓ 5,0	3	p<0,000051	miR-4758-3p	↓ 1,0	0,0	p<0,000051
miR-107	↓ 401,5	197,5	p<0,000051	miR-6733-3p	↓ 2,0	1	p<0,000051	miR-766-3p	↓ 39,5	28,0	p<0,000051
miR-625-5p	↓ 16,5	8	p<0,000051	miR-98-5p	↓ 238,5	171	p<0,000051	miR-3605-5p	↓ 0,0	0,0	p<0,000051
miR-664a-5p	↓ 12,5	7	p<0,000051	miR-150-3p	↓ 10,0	5	p<0,000051	miR-10a-5p	↓ 93769,5	62832,0	p<0,000051
miR-29c-3p	↓ 13084,5	5632,5	p<0,000051	miR-212-3p	↓ 24,0	15	p<0,000051	miR-653-5p	↓ 42,0	28,0	p<0,000051
miR-423-5p	↓ 223,0	122	p<0,000051	miR-150-5p	↓ 3847,5	1921,5	p<0,000051	miR-642a-5p	↓ 16,0	10,0	p<0,000051
miR-339-3p	↓ 52,0	32	p<0,000051	miR-552-5p	↓ 3,0	1	p<0,000051	miR-135a-5p	↓ 2,0	1,0	p<0,000051
miR-195-5p	↓ 149,0	71	p<0,000051	miR-511-5p	↓ 86,5	52	p<0,000051	miR-574-3p	↓ 349,0	255,0	p<0,000051
miR-29b-3p	↓ 3595,0	1986	p<0,000051	miR-29c-5p	↓ 107,5	75	p<0,000051	miR-361-5p	↓ 947,5	762,0	p<0,000051
let-7f-5p	↓ 35700,5	16911	p<0,000051	miR-3614-5p	↓ 12,5	8	p<0,000051	miR-4999-5p	↓ 5,0	3,0	p<0,000051
miR-146b-3p	↓ 834,0	399,5	p<0,000051	let-7e-3p	↓ 51,5	35,5	p<0,000051	miR-1304-3p	↓ 0,0	0,0	p<0,000051
miR-181c-5p	↓ 240,0	116	p<0,000051	let-7i-5p	↓ 2071,5	1603,5	p<0,000051	miR-30e-5p	↓ 29253,0	21599,5	p<0,000051
miR-361-3p	↓ 953,0	459,5	p<0,000051	miR-145-3p	↓ 127,5	82	p<0,000051	miR-192-3p	↓ 5,0	3,0	p<0,000051
miR-2110	↓ 7,0	3	p<0,000051	miR-511-3p	↓ 4,0	2	p<0,000051	miR-4732-3p	↓ 1,0	0,0	p<0,000051
miR-143-3p	↓ 280713,5	136831,5	p<0,000051	miR-5708	↓ 0,0	0	p<0,000051	miR-4777-3p	↓ 2,0	1,0	p<0,000051
miR-181a-3p	↓ 1296,5	668,5	p<0,000051	miR-30e-3p	↓ 20501,0	13793,5	p<0,000051	miR-3690	↓ 2,5	2,0	p<0,000051
miR-139-5p	↓ 136,0	68,5	p<0,000051	miR-5193	↓ 0,0	0	p<0,000051	miR-551a	↓ 4,0	2,0	p<0,000051
miR-100-5p	↓ 24565,5	12586,5	p<0,000051	miR-148a-5p	↓ 256,5	153,5	p<0,000051	miR-3913-5p	↓ 14,0	10,0	p<0,000051
miR-1-3p	↓ 34,5	13	p<0,000051	miR-4683	↓ 1,0	0	p<0,000051	miR-3174	↓ 1,0	0,0	p<0,000051
miR-29b-2-5p	↓ 142,0	74	p<0,000051	miR-338-3p	↓ 1995,5	1068,5	p<0,000051	let-7f-2-3p	↓ 3,0	2,0	p<0,000051
miR-1275	↓ 1,0	0	p<0,000051	miR-133a-3p	↓ 22,0	10	p<0,000051	miR-1270	↓ 4,0	3,0	p<0,000051
miR-1976	↓ 60,0	31,5	p<0,000051	miR-526b-5p	↓ 0,0	0	p<0,000051	miR-374b-5p	↓ 231,0	189,0	p<0,000051
miR-215-5p	↓ 15,0	10	p<0,000051	miR-202-5p	↓ 1,0	0	p<0,000051	miR-10a-3p	↓ 10,0	6,0	p<0,000051
miR-21-5p	↓ 1485338,0	919922	p<0,000051	miR-191-5p	↓ 1959,0	1350	p<0,000051	miR-5684	↓ 2,0	1,0	p<0,000051
miR-491-5p	↓ 9,0	5	p<0,000051	miR-4660	↓ 2,0	1	p<0,000051	miR-449a	↓ 4,0	2,0	p<0,000051
miR-3912-3p	↓ 3,0	1	p<0,000051	miR-216a-5p	↓ 2,0	1	p<0,000051	miR-3065-5p	↓ 23,0	18,0	p<0,000051
miR-6806-3p	↓ 5,0	2	p<0,000051	miR-331-3p	↓ 147,5	100	p<0,000051	miR-500b-3p	↓ 5,0	3,0	p<0,000051
miR-628-3p	↓ 7,0	3	p<0,000051	miR-26b-5p	↓ 2628,0	1878,5	p<0,000051	miR-7705	↓ 4,0	3,0	p<0,000051
miR-146b-5p	↓ 4557,0	2490	p<0,000051	miR-501-5p	↓ 16,0	11	p<0,000051	miR-6868-3p	↓ 1,0	0,0	p<0,000051
miR-664a-3p	↓ 110,0	67	p<0,000051	miR-6501-5p	↓ 1,0	1	p<0,000051	miR-3177-3p	↓ 0,0	0,0	p<0,000051
miR-1294	↓ 1,0	0	p<0,000051	miR-3191-3p	↓ 1,0	0	p<0,000051	miR-556-3p	↓ 1,0	1,0	p<0,000051
let-7a-5p	↓ 113621,0	66045,5	p<0,000051	miR-146a-5p	↓ 657,5	415,5	p<0,000051	let-7f-1-3p	↓ 11,0	8,0	p<0,000051
miR-589-5p	↓ 370,5	245	p<0,000051	miR-450b-5p	↓ 55,0	34	p<0,000051	miR-1287-3p	↓ 27,0	19,0	p<0,000051
miR-125a-5p	↓ 2433,0	1365,5	p<0,000051	miR-181c-3p	↓ 235,0	137,5	p<0,000051	miR-4726-5p	↓ 1,0	0,0	p<0,000051
miR-125a-3p	↓ 22,5	12	p<0,000051	miR-6772-3p	↓ 0,0	0	p<0,000051	miR-3194-3p	↓ 0,0	0,0	p<0,000051
miR-504-5p	↓ 4,0	2	p<0,000051	miR-6755-5p	↓ 3,0	2	p<0,000051	miR-5187-5p	↓ 2,0	1,0	p<0,000051
miR-330-5p	↓ 142,0	78,5	p<0,000051	miR-486-5p	↓ 386,5	193,5	p<0,000051	miR-450a-5p	↓ 33,0	20,5	p<0,000051
miR-181d-5p	↓ 102,5	53	p<0,000051	miR-3616-3p	↓ 0,0	0	p<0,000051	miR-27b-3p	↓ 8286,0	5848,5	p<0,000051
miR-148b-3p	↓ 809,0	528	p<0,000051	miR-497-5p	↓ 81,0	55	p<0,000051	miR-451a	↓ 1578,0	940,5	p<0,000051
miR-6761-5p	↓ 6,0	3	p<0,000051	miR-3909	↓ 2,0	1	p<0,000051	miR-3920	↓ 1,0	1,0	p<0,000051
miR-135b-5p	↓ 223,5	115	p<0,000051	miR-744-5p	↓ 127,5	80	p<0,000051	miR-1304-5p	↓ 1,0	1,0	p<0,000051
miR-194-5p	↓ 649,5	397,5	p<0,000051	miR-4788	↓ 1,0	0	p<0,000051	miR-340-3p	↓ 28,0	22,0	p<0,000051
miR-542-3p	↓ 935,5	496,5	p<0,000051	miR-195-3p	↓ 13,0	9	p<0,000051	miR-4685-3p	↓ 1,0	1,0	p<0,000051

O teste de Wilcoxon-Mann-Whitney, com correção de Bonferroni, foi executado nas expressões do RNA-seq. Para este propósito, foi adotado $\alpha_{\text{corrigido}}$:

$$\alpha_{\text{corrigido}} = \frac{\alpha_{\text{original}}}{n} = \frac{0,05}{39.275} \approx 0,0000013.$$

Considerando $p < 0,0000013$, identificou-se 12.816 DEGs para a separação do subtipo de doença, sendo 7.408 *upregulated* e 5.408 *downregulated*. Os 200 DEGs com menores valores de p estão apresentados nas Tabelas 47 e 48.

Tabela 47 – DEGs *downregulated* (↓) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD-LUSC.

RNA	Mediana		valor p	RNA	Mediana		valor p	RNA	Mediana		valor p
	LUAD	LUSC			LUAD	LUSC			LUAD	LUSC	
ARHGEF38	↓ 767,0	100,0	$p < 0,0000013$	ACSL5	↓ 6044,5	832,0	$p < 0,0000013$	BCL2L15	↓ 543,5	49,0	$p < 0,0000013$
HNF1B	↓ 1393,5	49,0	$p < 0,0000013$	MLPH	↓ 7977,5	1075,0	$p < 0,0000013$	LRRC31	↓ 102,0	4,0	$p < 0,0000013$
TMEM125	↓ 4170,0	665,0	$p < 0,0000013$	LONRF3	↓ 340,0	63,0	$p < 0,0000013$	MUC1	↓ 37896,0	4602,0	$p < 0,0000013$
RORC	↓ 2272,0	206,0	$p < 0,0000013$	KCNK5	↓ 2666,5	380,0	$p < 0,0000013$	ABCC6P1	↓ 94,5	11,0	$p < 0,0000013$
DDAH1	↓ 4728,0	1060,0	$p < 0,0000013$	SMPDL3B	↓ 1460,8	251,0	$p < 0,0000013$	GOLT1A	↓ 643,5	59,0	$p < 0,0000013$
NKX2.1	↓ 6489,5	224,0	$p < 0,0000013$	FMO5	↓ 2158,5	216,0	$p < 0,0000013$	TMC5	↓ 4757,0	630,0	$p < 0,0000013$
NKX2.1.AS1	↓ 365,8	16,0	$p < 0,0000013$	SFTA3	↓ 1801,5	76,0	$p < 0,0000013$	ARRB1	↓ 2713,5	726,0	$p < 0,0000013$
RGL3	↓ 1326,0	142,0	$p < 0,0000013$	ST3GAL5	↓ 4157,0	750,0	$p < 0,0000013$	AC004982.2	↓ 183,0	42,0	$p < 0,0000013$
SFTA2	↓ 5827,5	183,0	$p < 0,0000013$	OCLN	↓ 1771,2	391,0	$p < 0,0000013$	ICA1	↓ 1974,0	479,0	$p < 0,0000013$
GPR39	↓ 383,5	54,0	$p < 0,0000013$	CAPN8	↓ 3641,5	209,0	$p < 0,0000013$	ALDH3B1	↓ 4479,5	732,0	$p < 0,0000013$

Tabela 48 – DEGs *upregulated* (↑) identificados com o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni - LUAD-LUSC.

RNA	Mediana		valor <i>p</i>	RNA	Mediana		valor <i>p</i>	RNA	Mediana		valor <i>p</i>
	LUAD	LUSC			LUAD	LUSC			LUAD	LUSC	
KRT74	↑ 0,0	52,0	p<0,0000013	AC010280.1	↑ 0,0	21,0	p<0,0000013	DQX1	↑ 18,8	362,0	p<0,0000013
KRT5	↑ 98,5	248724,0	p<0,0000013	GPC1	↑ 2368,5	20306,0	p<0,0000013	DSG1	↑ 1,0	98,0	p<0,0000013
DSG3	↑ 3,0	14490,0	p<0,0000013	KRT13	↑ 14,6	6964,0	p<0,0000013	CAPNS2	↑ 4,0	599,0	p<0,0000013
DSC3	↑ 25,0	14311,0	p<0,0000013	COL7A1	↑ 505,0	12172,0	p<0,0000013	ADAM23	↑ 72,0	2468,0	p<0,0000013
CERS3	↑ 2,0	1143,0	p<0,0000013	HOXD11	↑ 0,0	60,0	p<0,0000013	ZNF488	↑ 8,0	203,0	p<0,0000013
BNC1	↑ 5,0	1785,0	p<0,0000013	TENM2	↑ 14,5	1056,0	p<0,0000013	WFDC5	↑ 1,0	123,0	p<0,0000013
S1PR5	↑ 29,0	753,0	p<0,0000013	PTHLH	↑ 61,0	3531,0	p<0,0000013	KCTD15	↑ 673,5	2917,0	p<0,0000013
KRT6A	↑ 81,0	159046,0	p<0,0000013	GABRQ	↑ 3,0	273,0	p<0,0000013	TINCR	↑ 53,5	819,0	p<0,0000013
KRT6B	↑ 7,0	9211,0	p<0,0000013	SLC2A1	↑ 4470,0	45065,0	p<0,0000013	CDHR1	↑ 20,0	377,0	p<0,0000013
KRT31	↑ 0,0	127,0	p<0,0000013	S100A7	↑ 2,0	533,0	p<0,0000013	SLAH2	↑ 1238,5	3825,0	p<0,0000013
SERPINB13	↑ 1,0	3540,0	p<0,0000013	A2ML1	↑ 8,0	1576,0	p<0,0000013	AC090004.2	↑ 10,0	572,0	p<0,0000013
KRT6C	↑ 1,0	1150,0	p<0,0000013	KRT17P1	↑ 1,0	18,0	p<0,0000013	ST6GALNAC2	↑ 112,3	783,5	p<0,0000013
IFITM3P9	↑ 0,0	46,5	p<0,0000013	AC022031.2	↑ 1,0	143,0	p<0,0000013	FBXO27	↑ 170,0	1311,0	p<0,0000013
CALML3	↑ 0,0	356,0	p<0,0000013	ANXA8L1	↑ 10,5	562,0	p<0,0000013	GJB6	↑ 30,2	3393,0	p<0,0000013
FAM83C	↑ 0,0	509,0	p<0,0000013	OR7E62P	↑ 0,0	19,0	p<0,0000013	ITGA6	↑ 2488,3	16428,0	p<0,0000013
C10orf99	↑ 0,0	245,0	p<0,0000013	KRT17P3	↑ 1,0	31,0	p<0,0000013	ATP5MC1P6	↑ 0,0	9,0	p<0,0000013
MIR205HG	↑ 30,0	3886,0	p<0,0000013	NTRK2	↑ 119,0	9728,0	p<0,0000013	IGSF11	↑ 13,0	340,0	p<0,0000013
AC011374.1	↑ 2,0	110,0	p<0,0000013	LINC01451	↑ 8,0	140,0	p<0,0000013	SERPINB2	↑ 5,0	286,0	p<0,0000013
AL512363.1	↑ 0,0	59,0	p<0,0000013	SLC6A8	↑ 1228,5	14522,0	p<0,0000013	FSCN1	↑ 3306,0	18954,0	p<0,0000013
TP63	↑ 173,5	23069,0	p<0,0000013	LINC00519	↑ 8,5	170,0	p<0,0000013	LINC01752	↑ 1,0	19,0	p<0,0000013
FAT2	↑ 47,0	6344,0	p<0,0000013	SDR9C7	↑ 1,0	23,0	p<0,0000013	NRARP	↑ 285,5	1487,0	p<0,0000013
LINC00640	↑ 3,0	46,0	p<0,0000013	SPRR2G	↑ 0,0	35,0	p<0,0000013	AP2M1	↑ 12732,0	34771,0	p<0,0000013
RPL26P13	↑ 0,0	9,0	p<0,0000013	S100A7A	↑ 0,0	70,0	p<0,0000013	SLC39A2	↑ 2,0	55,0	p<0,0000013
TRIM29	↑ 278,0	26654,0	p<0,0000013	GBP6	↑ 46,5	3884,0	p<0,0000013	ABCC5	↑ 2113,5	13307,0	p<0,0000013
CLCA2	↑ 17,5	12035,0	p<0,0000013	RHCG	↑ 9,0	1109,0	p<0,0000013	TFRC	↑ 5283,5	26819,0	p<0,0000013
KRT77	↑ 0,0	21,0	p<0,0000013	DLX5	↑ 23,5	559,0	p<0,0000013	LINC01503	↑ 194,0	755,0	p<0,0000013
DAPL1	↑ 8,5	1551,0	p<0,0000013	HAP1	↑ 21,5	519,0	p<0,0000013	LINC01416	↑ 0,0	14,0	p<0,0000013
FAM83B	↑ 31,0	1498,0	p<0,0000013	SNAI2	↑ 358,0	2361,0	p<0,0000013	IRX4	↑ 0,2	151,0	p<0,0000013
SPRR2A	↑ 1,0	1469,0	p<0,0000013	PARD6G	↑ 291,5	1414,0	p<0,0000013	NMNAT3	↑ 140,0	585,0	p<0,0000013
KRT14	↑ 14,5	6437,0	p<0,0000013	BMP7	↑ 116,5	4036,0	p<0,0000013	USH1G	↑ 4,0	219,0	p<0,0000013
NECTIN1	↑ 2062,5	25676,0	p<0,0000013	KRT34	↑ 0,0	20,0	p<0,0000013	UGT1A7	↑ 0,0	51,0	p<0,0000013
SPRR2E	↑ 0,0	312,0	p<0,0000013	ADAD2	↑ 2,0	35,0	p<0,0000013	S100A2	↑ 517,0	22316,0	p<0,0000013
AC009123.1	↑ 0,0	12,0	p<0,0000013	IRF6	↑ 3234,0	15629,0	p<0,0000013	LINC02457	↑ 1,0	24,0	p<0,0000013
CERNA2	↑ 2,0	493,0	p<0,0000013	FAM181B	↑ 4,0	118,0	p<0,0000013	CLCA4	↑ 2,5	134,0	p<0,0000013
ATP1B3	↑ 3471,5	19084,0	p<0,0000013	KCNQ1	↑ 8,0	263,0	p<0,0000013	DSC1	↑ 1,0	20,0	p<0,0000013
PERP	↑ 9857,0	72210,0	p<0,0000013	SPRR1A	↑ 2,0	816,0	p<0,0000013	LY6G6C	↑ 3,0	55,0	p<0,0000013
RAET1L	↑ 1,0	307,0	p<0,0000013	C17orf99	↑ 3,0	37,0	p<0,0000013	FOXE1	↑ 18,0	1683,0	p<0,0000013
KRT9	↑ 0,0	15,0	p<0,0000013	LYPD3	↑ 240,0	5647,0	p<0,0000013	FANCE	↑ 391,5	1319,0	p<0,0000013
KRT17	↑ 1158,5	99437,0	p<0,0000013	TMEM117	↑ 171,0	857,0	p<0,0000013	AL358332.1	↑ 1,0	6,0	p<0,0000013
CSTA	↑ 255,0	6604,0	p<0,0000013	MIR205	↑ 0,0	10,0	p<0,0000013	ZIC5	↑ 1,0	203,0	p<0,0000013
PKP1	↑ 189,5	39769,0	p<0,0000013	KRT17P2	↑ 1,0	17,0	p<0,0000013	MICALL1	↑ 1504,5	5701,0	p<0,0000013
KRT16	↑ 91,0	14972,0	p<0,0000013	PNCK	↑ 24,0	929,0	p<0,0000013	FRRS1	↑ 460,5	1801,0	p<0,0000013
ANXA8	↑ 19,0	1566,0	p<0,0000013	AL139095.1	↑ 0,0	6,0	p<0,0000013	AC011092.3	↑ 1,0	12,0	p<0,0000013
SERPINB5	↑ 47,5	9696,0	p<0,0000013	PAX6	↑ 1,0	119,0	p<0,0000013	AL033384.1	↑ 4,0	86,0	p<0,0000013
AC068831.5	↑ 1,0	15,0	p<0,0000013	AC108925.1	↑ 2,0	45,0	p<0,0000013	HES2	↑ 85,0	1154,0	p<0,0000013
TPRXL	↑ 4,0	250,0	p<0,0000013	ACKR3	↑ 676,0	6752,0	p<0,0000013	AP003390.1	↑ 8,0	56,0	p<0,0000013
GJB5	↑ 17,0	1503,0	p<0,0000013	SNX31	↑ 3,0	90,0	p<0,0000013	SOX15	↑ 50,5	1132,0	p<0,0000013
VSNL1	↑ 63,0	1634,0	p<0,0000013	SLC16A1	↑ 548,5	4620,0	p<0,0000013	AC027329.1	↑ 0,0	6,0	p<0,0000013
TMEM40	↑ 49,0	934,0	p<0,0000013	C12orf54	↑ 6,0	137,0	p<0,0000013	NDUFA4L2	↑ 711,0	9662,0	p<0,0000013
LINC02541	↑ 35,0	543,0	p<0,0000013	P2RY1	↑ 234,0	2128,0	p<0,0000013	BICD2	↑ 1779,0	5605,0	p<0,0000013
TMPRSS11A	↑ 1,0	264,0	p<0,0000013	AL033530.1	↑ 0,0	131,0	p<0,0000013	UCN2	↑ 5,0	85,0	p<0,0000013
JAG1	↑ 1871,5	14771,0	p<0,0000013	AC008011.2	↑ 0,0	9,0	p<0,0000013	KRT84	↑ 0,0	3,0	p<0,0000013
TMPRSS11F	↑ 0,0	63,0	p<0,0000013	DENND2C	↑ 145,0	947,0	p<0,0000013	AC008556.1	↑ 23,0	356,0	p<0,0000013
XXYL1	↑ 752,0	3168,0	p<0,0000013	GAST	↑ 0,0	12,0	p<0,0000013	FRMD6	↑ 922,5	4740,0	p<0,0000013
HR	↑ 88,0	1678,0	p<0,0000013	AC009271.1	↑ 3,0	195,0	p<0,0000013	ARTN	↑ 59,0	831,0	p<0,0000013
KRT42P	↑ 5,0	160,0	p<0,0000013	LINC02607	↑ 0,0	9,0	p<0,0000013	G2E3.AS1	↑ 0,0	63,0	p<0,0000013
KRT17P6	↑ 0,0	15,0	p<0,0000013	COL4A6	↑ 38,0	1424,0	p<0,0000013				

Análises biológicas

A análise Biológica foi realizada com o intuito de integrar e analisar os DE-miRNAs ($n = 301$, sendo 85 *upregulated* e 216 *downregulated*) e DEGS ($n = 12.816$, sendo 7408 *upregulated* e 5408 *downregulated*) identificados.

Primeiramente foi realizada a predição de genes-alvos dos DE-miRNAs, por meio da ferramenta mirDIP. Identificaram-se 50.693 genes-alvos dos DE-miRNAs (pares de combinações miRNA - gene-alvo), o qual sem duplicatas contém 10.015 genes-alvos. A Figura 49 mostra haver uma sobreposição de 24% (4.285) entre os genes-alvos dos DE-miRNAs e dos DEGs, os genes que estão na sobreposição serão nomeados de DEGs candidatos e seguiram nas próximas análises.

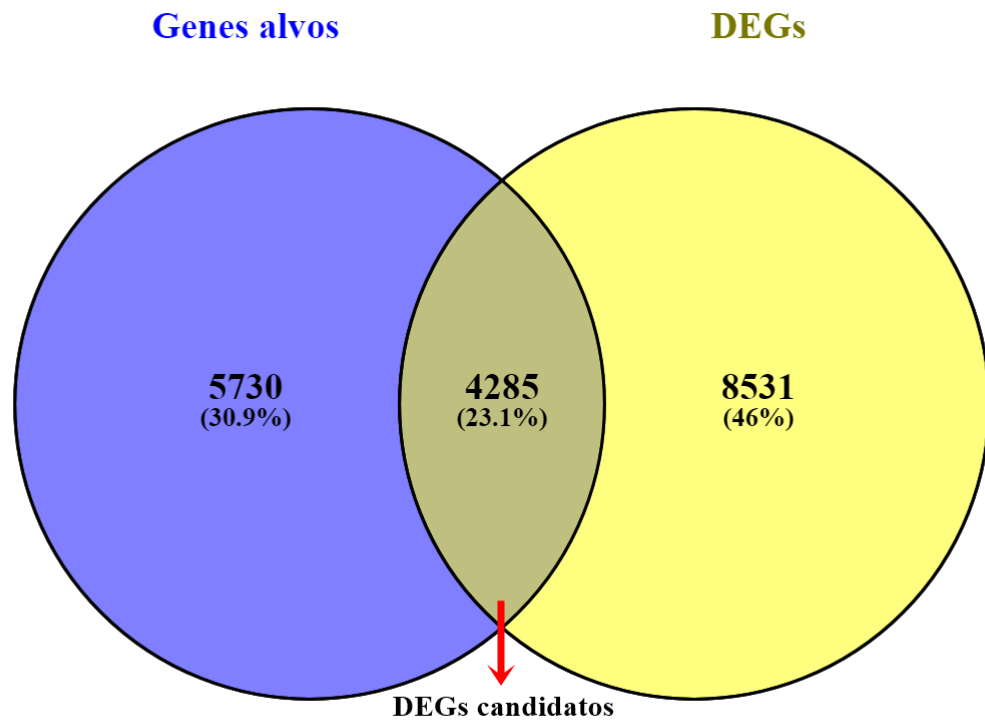


Figura 49 – Diagrama de Venn entre os genes-alvos dos DE-miRNAs e DEGs - LUAD-LUSC.

A lista gerada pelo mirDIP com os genes-alvos foi filtrado com base nos DEGs candidatos, assim identificaram-se 21.489 pares formados pelos DEGs candidatos e pelos DE-miRNAs que os regula. Foram selecionados DEGs candidatos com expressão contrária aos DE-miRNAs, assim, identificaram-se 12.690 pares com expressões contrárias.

Para realizar a análise de correlação, foi necessário combinar os conjuntos de dados de miRNA-seq ($n = 496$) e RNA-seq ($n = 511$) usando o ID da amostra do paciente. Isso resultou em um banco de dados unificado com $n = 488$ amostras de pacientes que eram compartilhadas em ambos os conjuntos de dados.

A correlação linear de Spearman foi calculada entre os pares, considerando correlações negativas, $\text{corr.}\rho < -0,4$ e $p < 0,05$ como estatisticamente significativo. Identificaram-se 288 pares (Tabela S8), a Tabela 49 apresenta as 50 menores correlações.

Tabela 49 – As 50 menores correlações - LUAD-LUSC.

miRNA	Gene	corr	p
miR-375-3p	CLCA2	-0,65	p<0,05
miR-944	SLC41A2	-0,63	p<0,05
miR-205-5p	ENPP4	-0,62	p<0,05
miR-375-3p	CDIN1	-0,61	p<0,05
miR-205-5p	C11orf86	-0,61	p<0,05
miR-205-5p	HSD17B11	-0,60	p<0,05
miR-375-3p	SHOX2	-0,59	p<0,05
miR-205-3p	TOX3	-0,59	p<0,05
miR-205-5p	LPCAT1	-0,58	p<0,05
miR-30d-5p	DBF4	-0,58	p<0,05
miR-29a-3p	TMTC3	-0,58	p<0,05
miR-29a-3p	SLC16A1	-0,57	p<0,05
miR-29a-3p	TDG	-0,56	p<0,05
miR-375-3p	PAX6	-0,56	p<0,05
miR-205-3p	CELF2	-0,55	p<0,05
miR-30d-5p	TWF1	-0,55	p<0,05
miR-30d-5p	ITGA6	-0,55	p<0,05
miR-205-5p	RBM47	-0,54	p<0,05
miR-205-5p	ERBB3	-0,54	p<0,05
miR-30d-5p	B3GNT5	-0,54	p<0,05
miR-29a-3p	POLE3	-0,53	p<0,05
miR-4709-3p	KPNA4	-0,53	p<0,05
miR-196b-5p	CPM	-0,52	p<0,05
miR-205-5p	MGAT4A	-0,51	p<0,05
miR-26a-5p	MAPK6	-0,51	p<0,05
miR-944	SGPP1	-0,51	p<0,05
miR-375-3p	YBX1	-0,51	p<0,05
miR-181b-5p	KPNA4	-0,51	p<0,05
miR-195-5p	CDCA4	-0,51	p<0,05
miR-30d-5p	PFN2	-0,51	p<0,05
miR-29a-3p	PXYLP1	-0,50	p<0,05
miR-29a-3p	MLLT11	-0,50	p<0,05
miR-140-3p	OLA1	-0,50	p<0,05
miR-205-5p	SLC4A4	-0,50	p<0,05
miR-30b-5p	ITGA6	-0,50	p<0,05
miR-181a-5p	KPNA4	-0,50	p<0,05
miR-4709-3p	NMD3	-0,49	p<0,05
miR-29c-3p	TDG	-0,49	p<0,05
miR-196b-5p	CELF2	-0,49	p<0,05
miR-26a-5p	PAK2	-0,49	p<0,05
miR-30b-5p	TWF1	-0,49	p<0,05
miR-26a-5p	E2F7	-0,49	p<0,05
miR-26a-5p	RCN2	-0,49	p<0,05
miR-26a-5p	SENP5	-0,49	p<0,05
miR-195-5p	KIF23	-0,49	p<0,05
miR-29b-3p	TMTC3	-0,49	p<0,05
miR-30d-5p	ATG5	-0,49	p<0,05
miR-29a-3p	CCSAP	-0,49	p<0,05
miR-29a-3p	TMEM183A	-0,49	p<0,05
miR-30d-5p	R3HDM1	-0,48	p<0,05

Os 288 pares foram analisados no miRTarBase, para selecionar apenas as interações entre os DE-miRNAs e DEGs candidatos validadas experimentalmente. Identificaram-se 86 pares, sendo 79 com DE-miRNA *downregulated* e DEGs candidatos *upregulated* (Tabela 50) e 7 DE-miRNAs *upregulated* e DEGs candidatos *downregulated* (Tabela 53).

Tabela 50 – DE-miRNAs *down* e DEGs *up* validados experimentalmente - LUAD-LUSC.

ID	miRNA		Gene		ID	miRNA		Gene	
MIRT171550	miR-30d-5p	↓	DBF4	↑	MIRT727331	miR-181b-5p	↓	KPNA1	↑
MIRT067384	miR-29a-3p	↓	TMTC3	↑	MIRT733335	miR-29a-3p	↓	HMGCR	↑
MIRT467969	miR-29a-3p	↓	SLC16A1	↑	MIRT733523	miR-30b-5p	↓	HOXA1	↑
MIRT053292	miR-29a-3p	↓	TDG	↑	MIRT727796	miR-30d-5p	↓	EED	↑
MIRT726209	miR-30d-5p	↓	TWF1	↑	MIRT047301	miR-181b-5p	↓	C2orf69	↑
MIRT193070	miR-26a-5p	↓	MAPK6	↑	MIRT467970	miR-29c-3p	↓	SLC16A1	↑
MIRT735232	miR-181b-5p	↓	KPNA4	↑	MIRT528220	miR-664a-5p	↓	METTL8	↑
MIRT502917	miR-195-5p	↓	CDCA4	↑	MIRT448783	miR-30d-5p	↓	GNA13	↑
MIRT001921	miR-29c-3p	↓	TDG	↑	MIRT346489	miR-26a-5p	↓	SSX2IP	↑
MIRT726210	miR-30b-5p	↓	TWF1	↑	MIRT736658	miR-139-5p	↓	TMPO	↑
MIRT007177	miR-26a-5p	↓	E2F7	↑	MIRT054563	miR-181b-5p	↓	RAP1B	↑
MIRT506855	miR-195-5p	↓	KIF23	↑	MIRT137722	miR-26a-5p	↓	EIF5	↑
MIRT067385	miR-29b-3p	↓	TMTC3	↑	MIRT168587	let-7g-5p	↓	HMGA1	↑
MIRT438870	miR-30d-5p	↓	ATG5	↑	MIRT193101	let-7e-5p	↓	MAPK6	↑
MIRT751269	miR-5698	↓	AP2M1	↑	MIRT727586	miR-30d-5p	↓	GCLC	↑
MIRT193104	let-7g-5p	↓	MAPK6	↑	MIRT207248	miR-29a-3p	↓	TET3	↑
MIRT548238	miR-29a-3p	↓	FEM1B	↑	MIRT217747	miR-195-5p	↓	TBPL1	↑
MIRT727110	miR-181b-5p	↓	NAA50	↑	MIRT134087	miR-26a-5p	↓	KLHL42	↑
MIRT023460	miR-30b-5p	↓	B3GNT5	↑	MIRT079940	miR-29a-3p	↓	RNF138	↑
MIRT067977	miR-30d-5p	↓	TFDP1	↑	MIRT000682	miR-29a-3p	↓	CDK6	↑
MIRT020400	miR-29c-3p	↓	TMTC3	↑	MIRT732280	miR-26a-5p	↓	JAG1	↑
MIRT053293	miR-29b-3p	↓	TDG	↑	MIRT728254	miR-181b-5p	↓	ATP2B1	↑
MIRT000110	miR-26a-5p	↓	HMGA1	↑	MIRT726354	miR-195-5p	↓	TBL1XR1	↑
MIRT502569	miR-195-5p	↓	E2F7	↑	MIRT054562	miR-181a-5p	↓	RAP1B	↑
MIRT265078	miR-195-5p	↓	CHEK1	↑	MIRT052385	let-7a-5p	↓	RRM2	↑
MIRT188273	miR-30d-5p	↓	RAP1B	↑	MIRT552620	miR-29a-3p	↓	ZBTB5	↑
MIRT025186	miR-181a-5p	↓	METAP1	↑	MIRT050138	miR-26a-5p	↓	ADM	↑
MIRT728263	miR-195-5p	↓	ATP13A3	↑	MIRT764617	miR-30b-3p	↓	SEN5	↑
MIRT467971	miR-29b-3p	↓	SLC16A1	↑	MIRT003976	miR-1-3p	↓	HSPD1	↑
MIRT727746	miR-30d-5p	↓	ERLIN1	↑	MIRT501877	miR-29a-3p	↓	MORF4L1	↑
MIRT088446	miR-30d-5p	↓	LCLAT1	↑	MIRT547822	miR-29a-3p	↓	ISG20L2	↑
MIRT000397	let-7g-5p	↓	CDC25A	↑	MIRT149118	miR-181b-5p	↓	LRRC8D	↑
MIRT727747	miR-30b-5p	↓	ERLIN1	↑	MIRT001771	miR-26a-5p	↓	EZH2	↑
MIRT556267	miR-664a-3p	↓	MAPK6	↑	MIRT067978	miR-30b-5p	↓	TFDP1	↑
MIRT479150	miR-4709-3p	↓	CNBP	↑	MIRT023441	miR-30b-5p	↓	GNA13	↑
MIRT001021	miR-26a-5p	↓	CDK8	↑	MIRT726535	miR-26a-5p	↓	SLC38A2	↑
MIRT149116	miR-30d-5p	↓	LRRC8D	↑	MIRT086584	miR-181a-5p	↓	C2orf69	↑
MIRT188274	miR-30b-5p	↓	RAP1B	↑	MIRT002315	miR-26a-5p	↓	SERBP1	↑
MIRT545976	miR-195-5p	↓	YWHAQ	↑	MIRT088447	miR-30b-5p	↓	LCLAT1	↑
MIRT548237	miR-29b-3p	↓	FEM1B	↑					

↑ upregulated e ↓ downregulated.

A rede de regulatória entre DE-miRNAs e DEGs foi construída e está apresentada na Figura 50. Na rede a esquerda é possível observar o miR-26a-5p ligado a *MAPK6*, *CDK8*, *SSX2IP*, *EIF5*, *KLHL42*, *JAG1*, *ADM*, *SLC38A2*, *SERBP1*, *EZH2*, *E2F7* e *HMGA1* e também o miR-195-5p ligado a *TBL1XR1*, *TBPL1*, *YWHAQ*, *ATP13A3*, *CHEK1*, *KIF23* e *CDCA4*. Além disso, observa-se os miR-30d-5p, miR-181b-5p, miR-30b-5p e miR-29a-3p com bastante ligações.

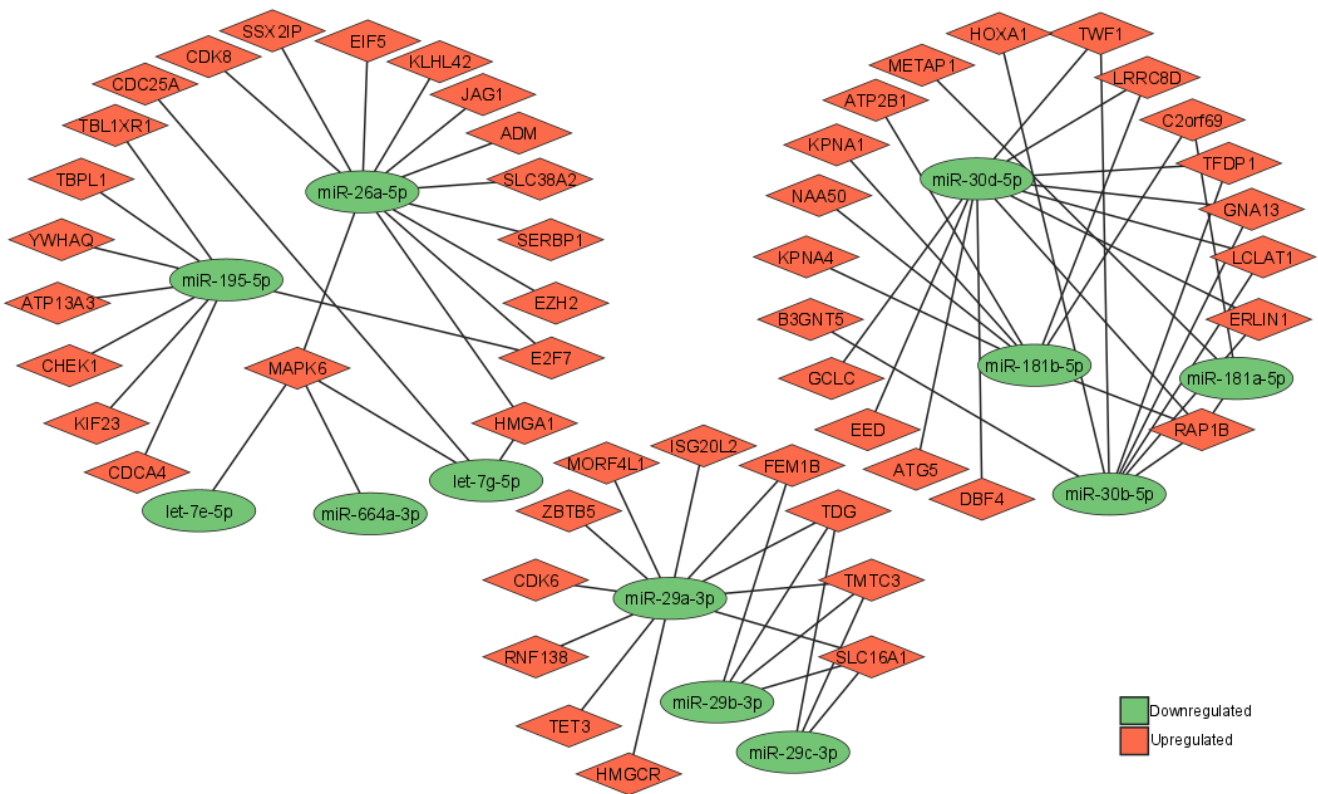


Figura 50 – Rede regulatória entre DE-miRNA *down* e DEGs *up* - LUAD-LUSC.

Realizou-se a análise de enriquecimento das vias para os DE-miRNAs e DEGs da Tabela 50. Para os DE-miRNAs, utilizou-se a ferramenta miRPath, e para os DEGs, a ferramenta Enrichr. A Tabela 51 apresenta as principais vias identificadas com os miRs (Tabela S9 completa), e a Tabela 52 apresenta a única via identificada com os DEGs, considerando valor $p < 0,05$.

As vias moleculares destacadas, como *Proteoglycans in Cancer*, *Pathways in Cancer*, *FoxO Signaling Pathway*, e outras, revelam a complexidade dos processos envolvidos no desenvolvimento e progressão do câncer de pulmão, abrangendo desde a regulação do ciclo celular até a resistência a inibidores de tirosina quinase, fornecendo *insights* valiosos para investigações futuras e estratégias terapêuticas direcionadas em LUAD e LUSC.

Tabela 51 – Principais Vias enriquecidas para os DE-miRNAs -LUAD-LUSC.

KEGG pathway	Term Genes	Target Genes (n)	miRNAs (n)	FDR
Proteoglycans in cancer	220	130	16	p<0,01
Pathways in cancer	555	270	16	p<0,01
FoxO signaling pathway	139	88	16	p<0,01
Cell cycle	129	79	16	p<0,01
p53 signaling pathway	75	52	16	p<0,01
EGFR tyrosine kinase inhibitor resistance	82	52	16	p<0,01
Small cell lung cancer	100	59	16	p<0,01
mTOR signaling pathway	177	92	16	p<0,01
PI3K-Akt signaling pathway	372	170	16	p<0,01
Apoptosis - multiple species	32	24	16	p<0,01
AMPK signaling pathway	130	68	16	p<0,01
Non-small cell lung cancer	81	46	16	p<0,01
TGF-beta signaling pathway	103	54	15	p<0,01

Tabela 52 – Vias enriquecidas para os DEGs - LUAD-LUSC

Term	Overlap	Adjusted P value	Genes
Cell cycle	6/124	$p < 0,001$	DBF4;TFDP1;CDK6;YWHAQ;CHEK1;CDC25A

Os 7 DE-miRNAs *upregulated* e DEGs *downregulated*, estão apresentados na Tabela 53. A rede regulatória está ilustrada na Figura 51, destaca-se os miR-205-5p ligado a ENPP4, ERBB3, LYSMD3 e LPCAT1 e o miR-944 ligado a DPYSL2 e MACC1.

Tabela 53 – DE-miRNAs up e DEGs downs validados experimentalmente - LUAD-LUSC.

miRTarBase ID	miRNA		Gene	
MIRT440724	hsa-miR-205-5p	↑	ENPP4	↓
MIRT440320	hsa-miR-205-5p	↑	LPCAT1	↓
MIRT003659	hsa-miR-205-5p	↑	ERBB3	↓
MIRT440298	hsa-miR-205-5p	↑	LYSMD3	↓
MIRT224764	hsa-miR-944	↑	DPYSL2	↓
MIRT782980	hsa-miR-944	↑	MACC1	↓
MIRT726426	hsa-miR-93-5p	↑	STK17B	↓

↑ *upregulated* e ↓ *downregulated*.

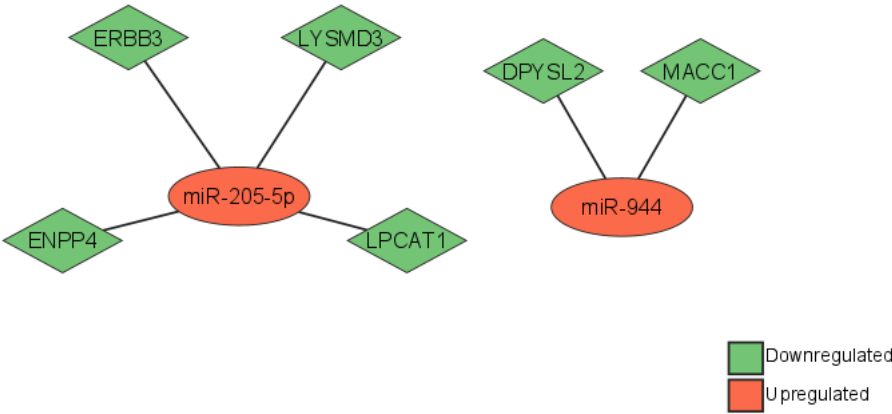


Figura 51 – Rede regulatória entre DE-miRNA *up* e DEGs *down*.

A Tabela 54 apresenta as principais vias identificadas e a lista completa está na Tabela S10. Não foram identificadas vias com os DEGs. Também foram identificados importantes vias, descritas anteriormente, relacionadas ao câncer de pulmão.

considerando especialmente o contexto da regulação do ciclo celular, que desempenha um papel crucial no desenvolvimento e na progressão do câncer.

Tabela 55 – Processos biológicos relacionados a Rede PPI - LUAD-LUSC.

GO-termo	Descrição	Contagem na rede	Força	FDR
GO:0000082	transition of mitotic cell cycle	6 de 86	1,330	0,008
GO:0044772	Mitotic cell cycle phase transition	7 de 173	1,100	0,010
GO:0090068	Positive regulation of cell cycle process	7 de 251	0,930	0,047
GO:0045787	Positive regulation of cell cycle	8 de 349	0,850	0,047
GO:0006259	DNA metabolic process	11 de 785	0,630	0,047
GO:0007049	Cell cycle	14 de 1246	0,540	0,047
GO:0044260	Cellular macromolecule metabolic process	21 de 2512	0,410	0,047
GO:0043412	Macromolecule modification	23 de 2898	0,390	0,047
GO:0080090	Regulation of primary metabolic process	35 de 5899	0,260	0,047
GO:0044238	Primary metabolic process	41 de 7156	0,250	0,020
GO:0044237	Cellular metabolic process	38 de 6568	0,250	0,047
GO:0006807	Nitrogen compound metabolic process	38 de 6643	0,250	0,047
GO:0071704	Organic substance metabolic process	41 de 7522	0,220	0,047
GO:0008152	Metabolic process	42 de 7988	0,210	0,047
GO:0009987	Cellular process	61 de 14826	0,100	0,047

Análises estatísticas

Foram realizadas análises de árvore de classificação e Florestas Aleatórias para a identificação de DE-miRNAs e DEGs na separação dos subtipos LUAD e LUSC. O banco de dados combinado composto por miRNA-seq e RNA-seq ($n = 488$) passou por um processo de filtragem, resultando em 22 miRNAs e 64 genes (sem duplicatas) que, foram previamente validados experimentalmente. A análise foi conduzida separadamente para os miRNAs-seq e RNAs-seq.

Árvore de classificação e Florestas Aleatórias (construído a fim de complementar os resultados obtidos na árvore), foram construídas para discriminar os subtipos LUAD e LUSC. O banco de dados foi aleatoriamente dividido em 80% ($n = 390$) para treinamento e 20% ($n = 98$) para teste. Isso possibilitou a avaliação comparativa do desempenho do banco de treino em novos conjuntos de dados (banco de teste). Neste caso não houve a necessidade de balancear as duas classes de subtipos.

Árvore de classificação

Foram simuladas 100 mil árvores de classificação, tanto para o conjunto de DE-miRNAs quanto para de DEGs, a fim de identificar a árvore de classificação com maiores acurácias na separação de LUSC vs. LUAD. Para essas árvores consideraram-se os hiperparâmetros: *minsplit*= 2, *minbucket*= 1 e *cp*= 0, 00001.

Primeiramente, foram realizadas as simulações para o banco de DE-miRNAs, a Figura 53 apresenta o histograma dessas acurácias, variando de 76% a 99% com média acima de 90%.

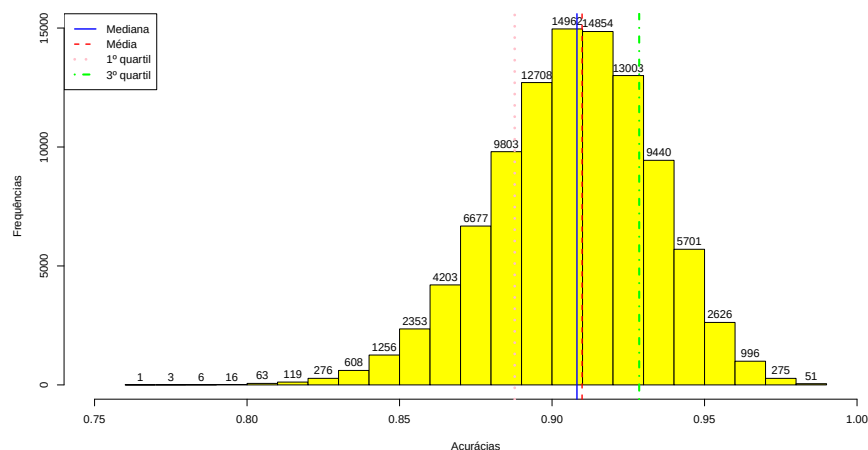


Figura 53 – Histograma das 100 mil acurácias para árvores de classificação dos DE-miRNAs - LUAD-LUSC.

A árvore de classificação com 99% de acurácia foi gerada, contudo, como pode-se observar na Tabela 56, formou-se com 25 nós e ficou inviável de ser utilizada. Dessa forma houve a necessidade de realizar a pós-poda. Para podar a árvore foi selecionado o valor cp que tem o menor erro de validação cruzada, sendo assim, o valor do $cp = 0,005$ irá gerar uma árvore com 6 divisões.

Tabela 56 – Resumo do modelo de árvore de classificação para os DE-miRNAs - LUAD-LUSC.

CP	Nº de divisões	Erro relativo	Erro de validação cruzada	Erro padrão
0,795	0	1,000	1,000	0,053
0,054	1	0,205	0,249	0,034
0,022	2	0,151	0,184	0,030
0,016	4	0,108	0,195	0,031
0,005	6	0,076	0,162	0,028
0,004	13	0,038	0,178	0,030
0,003	19	0,016	0,189	0,031
0,000	25	0,000	0,222	0,033

A árvore podada pode ser visualizada na Figura 54, o primeiro nó contém o miR-944, seguido do miR-29a-3p, let-7e-5p, miR-205-5p, miR-944 e miR-5698. Observa-se a presença do miR-944 duas vezes na árvore e a maior quantidade de observações das amostras se concentram mais nas pontas para LUAD (nó 3) e para LUSC (nó 13).

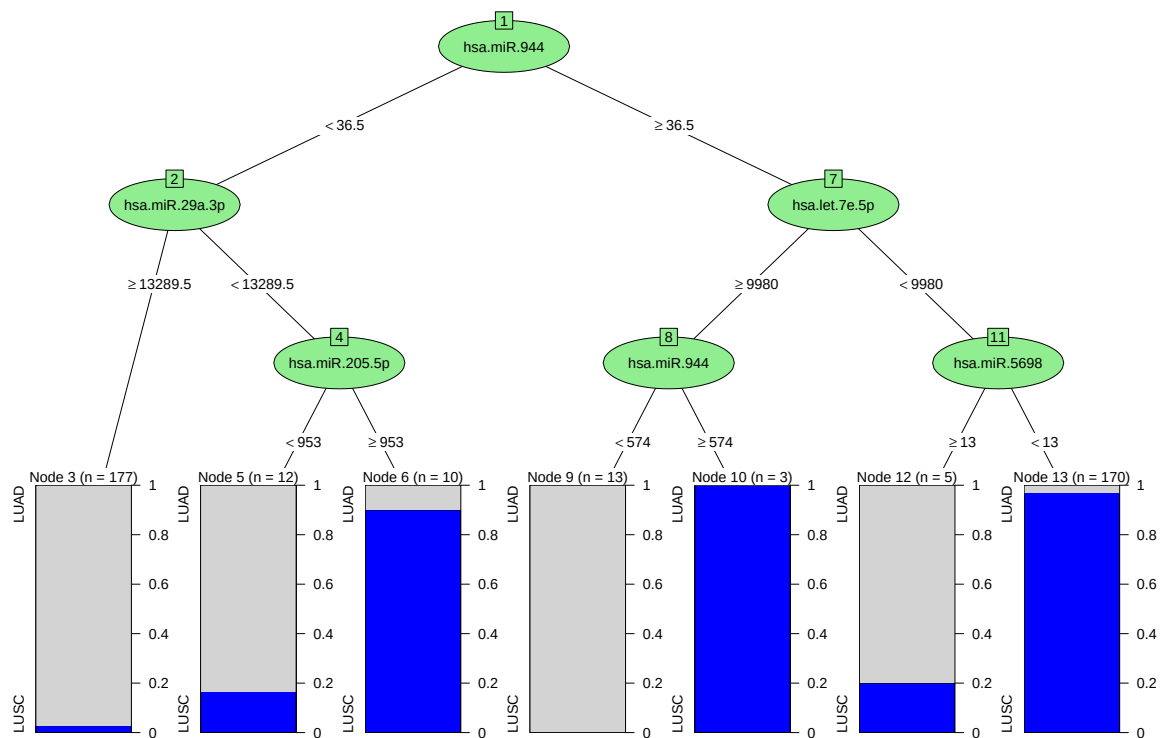


Figura 54 – Árvore de classificação pós-poda para os DE-miRNAs - LUAD-LUSC.

A Figura 55 apresenta os DE-miRNAs mais importantes para o modelo de árvore de classificação pós-poda, com destaque de maior importância para miR-944 e miR-205-5p, com importância acima de 100.

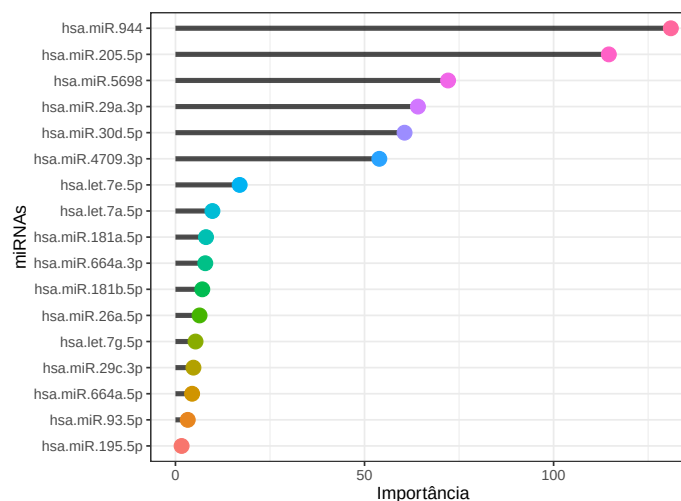


Figura 55 – Variáveis importantes para a árvore pós-poda de DE-miRNAs- LUAD-LUSC.

A seguir tem-se a matriz de confusão ao testar o banco de teste no modelo da árvore pós-poda, destacando-se apenas duas classificações erradas. Logo, tem-se as medidas de desempenho

do modelo, com especificidade de 95% e sensibilidade 100% e a Figura 56 apresenta a curva ROC com $AUC = 0,9828$.

Predição	Referência	
	LUAD	LUSC
LUAD	56	2
LUSC	0	40

Medida	Valor
Acurácia	98%
Sensibilidade	100%
Especificidade	95%

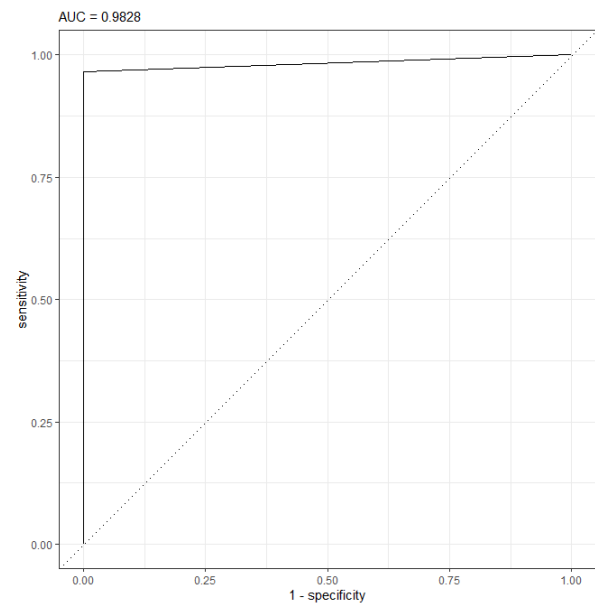


Figura 56 – Curva ROC para DE-miRNAs - LUAD-LUSC.

Em relação aos DEGS, a Figura 57, apresenta o histograma das 100 mil simulações das acurácias, estão variando de 71% a 99%, com média de aproximadamente 88%.

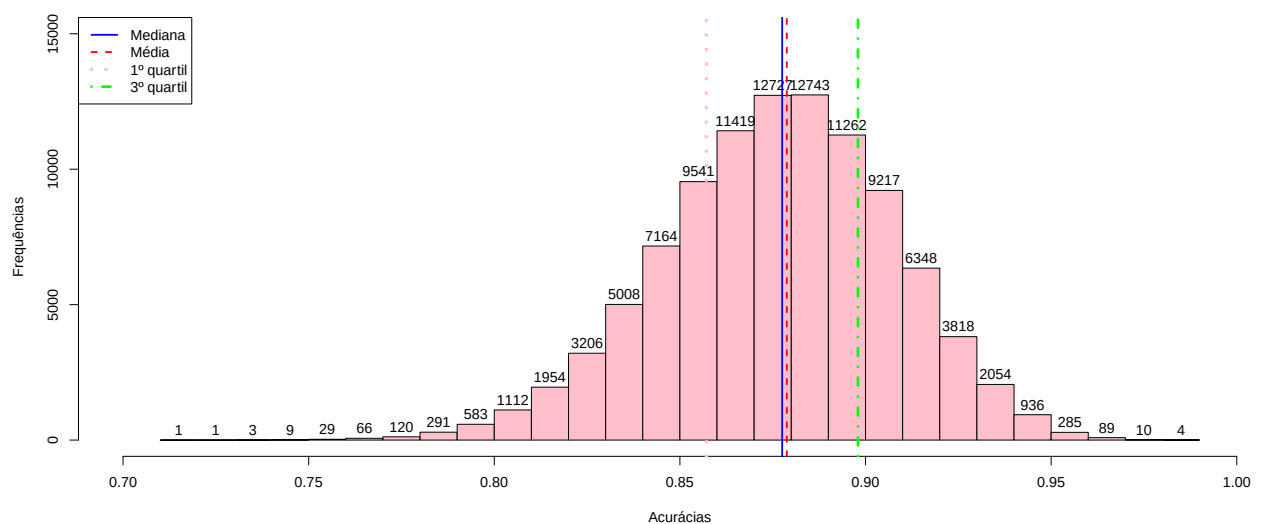


Figura 57 – Histograma das 100 mil acurácias para árvore de classificação dos DEGs -LUAD-LUSC.

A árvore de classificação com 99% de acurácia foi gerada, contudo, como pode-se observar na Tabela 57, formou-se com 22 nós e ficou inviável de ser utilizada. Dessa forma, foi

necessário realizar a pós-poda. Para podar a árvore foi selecionado o valor $cp=0,016$ que tem o menor erro de validação cruzada, isso irá gerar uma árvore com 4 divisões.

Tabela 57 – Resumo do modelo da árvore de sobrevivência para os DEGs - LUAD-LUSC.

CP	Nº de divisões	Erro relativo	Erro de validação cruzada	Erro padrão
0,732	0	1,000	1,000	0,054
0,066	1	0,268	0,301	0,038
0,033	2	0,202	0,273	0,036
0,022	3	0,169	0,262	0,035
0,016	4	0,148	0,240	0,034
0,011	5	0,131	0,268	0,036
0,008	10	0,077	0,306	0,038
0,007	12	0,060	0,322	0,039
0,005	15	0,038	0,322	0,039
0,000	22	0,000	0,306	0,038

A árvore de classificação pós-poda pode ser visualizada na Figura 58, o primeiro nó contém o *JAG1*, seguido dos *ERBB3*, *HOXA1* e *TDG*, observa-se maior quantidade de observações no nó 3 para LUAD e nó 9 para LUSC.

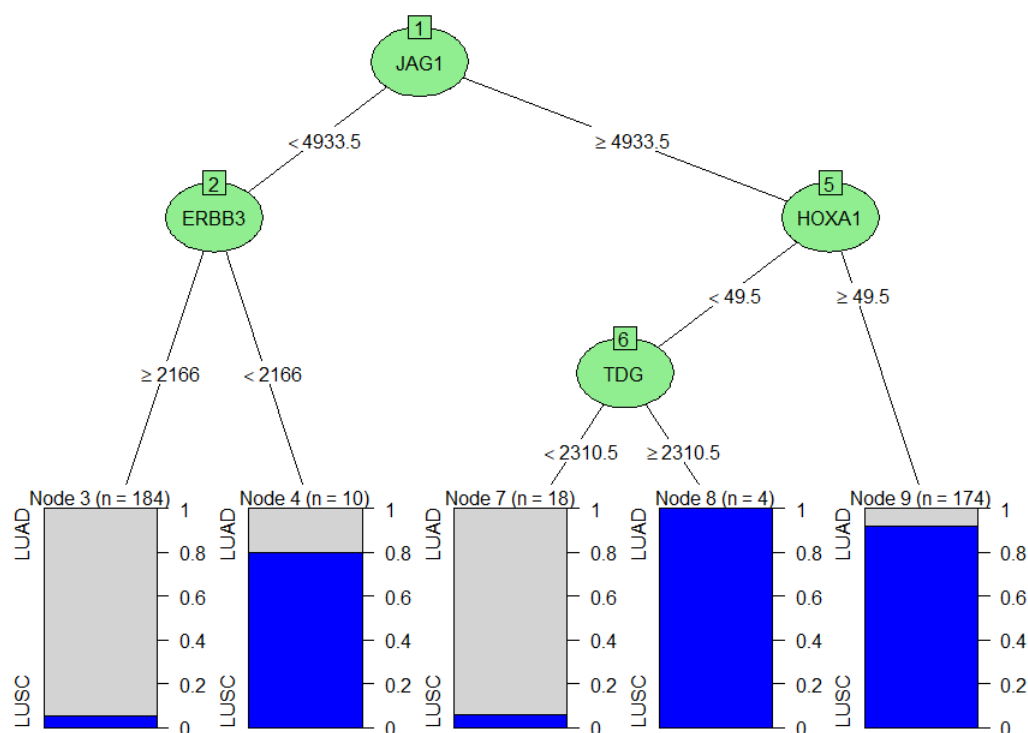


Figura 58 – Árvore de classificação pós-poda para os DEGs - LUAD-LUSC.

A Figura 59 apresenta os DEGs com maiores importâncias para a diferenciação da doença no modelo de árvore de classificação pós-poda, com destaque para o gene *JAG1*, acima de 90 de importância, e depois, *SLC16A1*, *E2F7* e *AP2M1* em sequência.

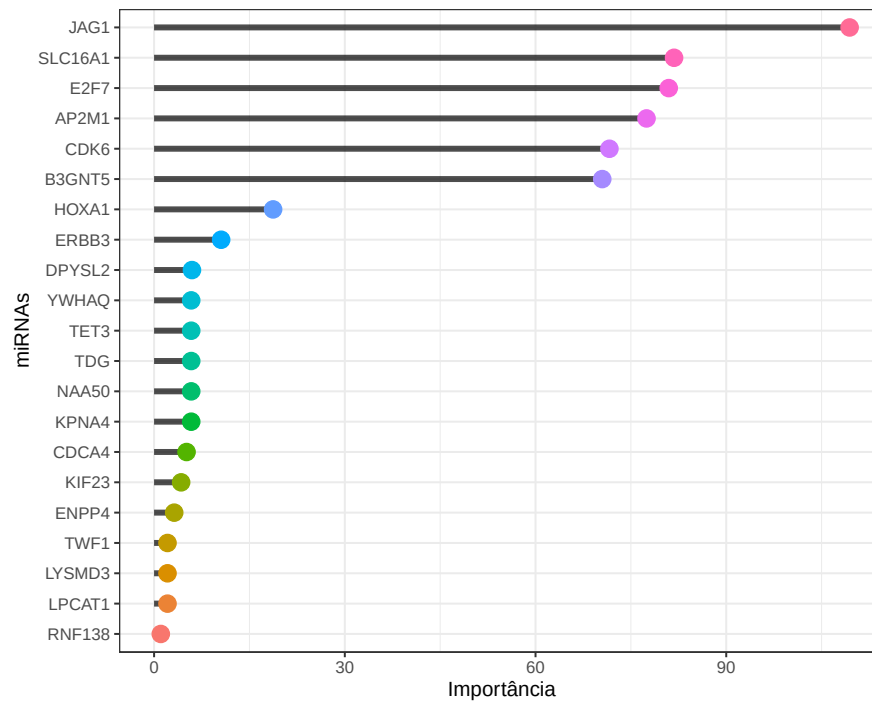


Figura 59 – Variáveis importantes para os DEGs - LUAD-LUSC.

A baixo na matriz de confusão ao testar o banco de teste no modelo da árvore pós-poda, destaca-se quatro classificações erradas. Posteriormente tem-se as medidas de desempenho do modelo, com especificidade de 100% e sensibilidade 93%. A curva ROC, Figura 60, obteve AUC= 95%.

Predição	Referência	
	LUAD	LUSC
LUAD	56	4
LUSC	0	38

Medida	Valor
Acurácia	96%
Sensibilidade	93%
Especificidade	100%

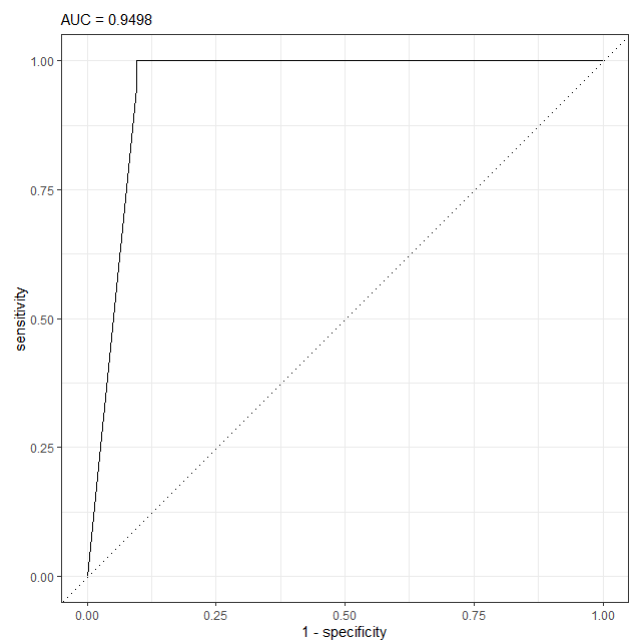


Figura 60 – Curva ROC para DEGs - LUAD-LUSC.

Florestas Aleatórias

A análise pelo modelo de Florestas Aleatórias foi realizada utilizando os DE-miRNAs e DEGs validados experimentalmente. O modelo foi construído com 5000 árvores de classificação, identificando os 10 DE-miRNAs e 10 DEGs mais significativos para a diferenciação das doenças.

A Figura 61a apresenta os 10 DE-miRNAs mais importantes identificados pelo modelo para a separação dos subtipos, com destaque maior para o miR-944. Ao lado, a Figura 61b mostra os erros de modelo, houve estabilidade próximo de 1000 árvores.

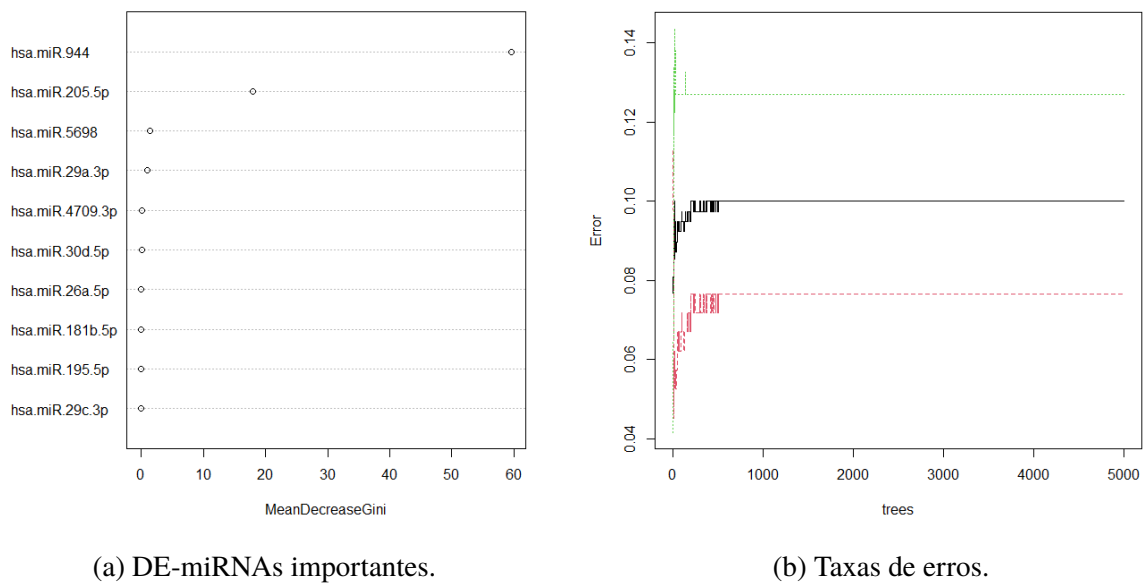


Figura 61 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DE-miRNA - LUAD-LUSC.

Na Figura 62a são apresentados os 10 DEGs mais importantes identificados pelo modelo para a separação dos subtipos e notavelmente, o gene *JAG1* se destaca. Adicionalmente, a Figura 62b ilustra os erros do modelo, revelando uma estabilidade próxima de 2000 árvores.

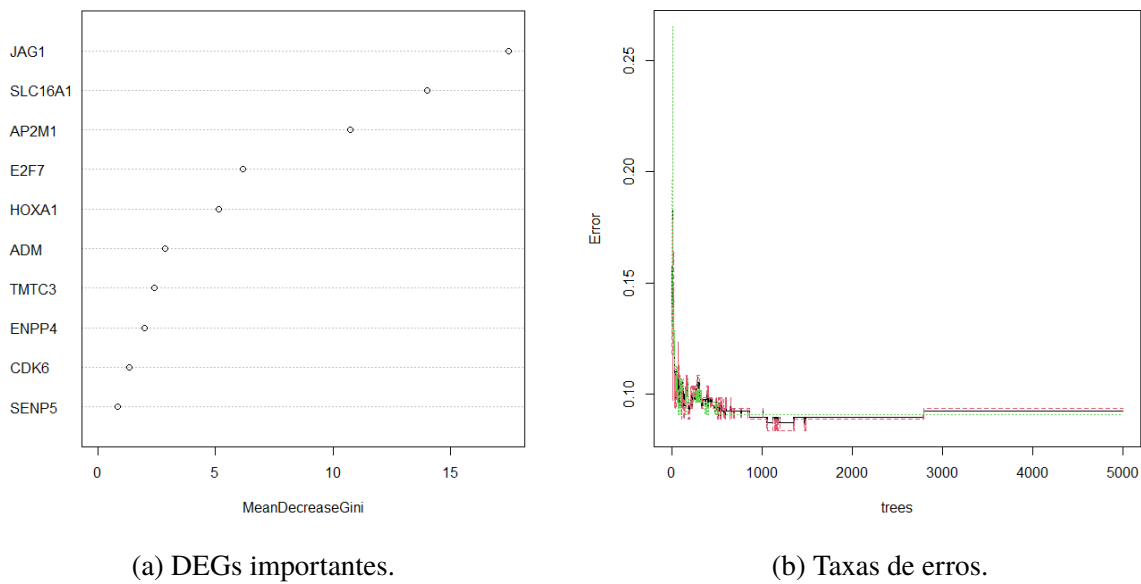


Figura 62 – Variáveis importantes e taxas de erros para o modelo de Florestas Aleatórias para DEGs - LUAD-LUSC.

Discussões

Após realizar análises descritivas, biológicas e estatísticas nos bancos de dados de miRNA-seq e RNA-seq, foi possível examinar e identificar padrões e comportamentos relevantes. Todas as análises foram conduzidas para identificar DE-miRNAs e DEGs, que desempenham um papel crucial na diferenciação entre tecido LUSC vs. LUAD em estágios precoces.

A análise de expressão diferencial foi conduzida utilizando o teste de Wilcoxon-Mann-Whitney com correção de Bonferroni. Esta abordagem foi escolhida para preservar os dados originais e empregar estatísticas não-paramétricas, diferenciando-se das estratégias convencionais. Durante a análise, foram identificados 301 DE-miRNAs e 12.816 DEGs.

Após a identificação dos DE-miRNAs e DEGs, as análises biológicas conduziram à identificação de 86 pares potenciais (DE-miRNA - DEG candidato) que exibem expressões opostas, apresentam correlações negativas significativas e foram validados experimentalmente.

Ao analisar o conjunto dos 79 pares de DE-miRNAs *down* e DEGs *up*, a rede regulatória (Figura 50) aponta vários miRNAs importantes relacionados a DEGs. Em relação ao miR-26a-5p, Liang *et al.* (2018) revelam que o SNHG6 pode atuar como um lncRNA oncogênico na carcinogênese LUAD, regulando o eixo miR-26a-5p/E2F7, esse eixo foi identificado na rede.

Li *et al.* (2020) concluem que o miR-195-5p pode atuar como um biomarcador que contribui para o diagnóstico do câncer de pulmão e a detecção de sua população de alto risco.

A família do miR-29 apareceu, com o miR-29a-3p, miR-29b-3p, miR-29c-3p e também a família do miR-30, como miR-30b-5p e miR-30d-5p. Além disso, foram identificadas importantes

vias enriquecidas relacionadas aos DE-miRNAs e DEGs, como Vias no câncer, ciclo celular, via de sinalização p53 e resistência ao inibidor de tirosina quinase EGFR.

Para os 7 DE-miRNAs *upregulated* e DEGs *downregulated*, a rede regulatória ficou pequena, apenas com os miR-205-5p e miR-944. Mesmo com poucos pares foram identificadas importantes vias enriquecidas, como Via de sinalização FoxO, MAPK, mTOR e Wnt.

A rede PPI gerada com os DEGs ficou mais simples, mas também mostrou relações entre fortes entre CHEK1 e CDC25A, por exemplo.

Em relação às análises estatísticas para os DE-miRNAs a árvore de classificação destacou o miR-944 duas vezes e para o Florestas Aleatórias esse miRNA teve a maior importância na separação de LUSC vs. LUAD. O trabalho de [Shen et al. \(2022\)](#) faz um vasto estudo a respeito desse miR-944, destacando o potencial como biomarcador de LUSC. Além disso, o miR-944 foi significativamente subexpresso em tecidos LUAD e linhas celulares (como A549 e H1299) e significativamente superexpresso em tecidos LUSC (incluindo TCGA-LUSC), resultando em conclusões semelhantes ao ilustrado na Figura 63.

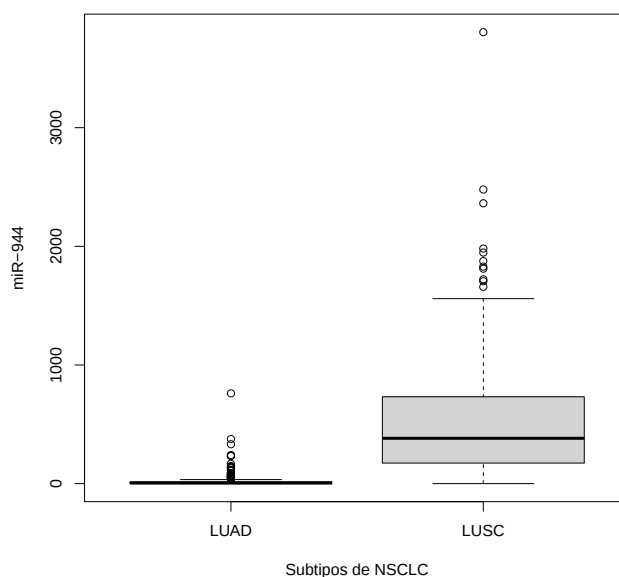


Figura 63 – MiR-944 em LUSC vs. LUAD.

[Liu, Zhou e Cao \(2016\)](#) relataram que o miR-944 tem como alvo o *EPHA7* (*EPH Receptor A7*) no NSCLC e regula a proliferação celular do NSCLC, o que pode oferecer um novo mecanismo subjacente ao desenvolvimento e progressão do NSCLC.

O miR-205-5p foi o segundo miRNA mais importante na diferenciação das doenças, segundo Florestas Aleatórias. Este miRNA também apareceu na árvore de classificação. Em relação aos miR-205-3p e 205-5p, [Lebanony et al. \(2009\)](#) conclue que o miR-205 é um marcador altamente preciso para câncer de pulmão de histologia escamosa e que o ensaio diagnóstico

padronizado no estudo pode fornecer uma subclassificação altamente precisa de pacientes com NSCLC. Boeri *et al.* (2011) destacam que o mir-205 e mir-21 discriminaram significativamente adenocarcinoma de carcinoma de células escamosas ($p \leq 0,001$).

Lu *et al.* (2012) identificaram um grande conjunto de miRNAs, incluindo membros da família Let-7 e miR-205, cujos perfis de expressão diferiram fortemente entre LUAD e LUSC.

Em relação à árvore de classificação com os DEGs, o destaque foi o JAG1 que também foi o mais importante detectado em Florestas Aleatórias. Chang *et al.* (2016) destacam que JAG1 serve como um biomarcador prognóstico e o eixo JAG1/HSPA2 contribui para a malignidade do câncer de pulmão.

Neste estudo comparativo entre LUSC vs. LUAD, optou-se por não gerar árvores de sobrevivência, uma vez que o foco da análise era a identificação de DE-miRNAs e DEGs distintos entre essas doenças nos estágios iniciais.

Este estudo destaca marcadores genéticos cruciais revelando vias essenciais. Estas descobertas têm o potencial de impactar significativamente o diagnóstico precoce e o tratamento personalizado do câncer de pulmão, melhorando as perspectivas para pacientes com LUSC e LUAD em estágios iniciais. A integração desses resultados na prática clínica pode representar um avanço crucial para estratégias terapêuticas mais eficazes e direcionadas.

5 Conclusão

Este trabalho utilizou técnicas de bioinformática, incluindo metodologias estatísticas e ferramentas biológicas, para analisar grandes conjuntos de dados de miRNA-Seq e RNA-Seq, visando identificar potenciais biomarcadores associados ao LUAD e LUSC nos estadiamentos iniciais da doença que diferenciem T vs. N, que estejam relacionadas a sobrevivência dos pacientes e que diferenciam as doenças LUSC vs. LUAD. Todos esses procedimentos foram realizados com o intuito de conseguir compreender os mecanismos subjacentes à formação do câncer e colaborar para o desenvolvimento de estratégias diagnósticas mais precisas.

Vale destacar, que com o avanço da medicina moderna e a integração de NGS como ferramentas rotineiras de diagnóstico, prognóstico ou tratamento, há a necessidade de identificação precisa de moléculas clinicamente úteis, incluindo os RNAs e miRNAs, para LUAD e LUSC em estadiamento inicial.

Abaixo, estão destacados os resultados mais importantes deste trabalho e as conclusões pontuais em relação às análises I, II e III:

- **Análise I:** O teste de Wilcoxon-Mann-Whitney com correção de Bonferroni mostrou ser uma ferramenta mais simples e conservadora para a identificação de DE-miRNAs no banco de dados LUAD miRNA-seq quando comparado ao EdgeR e Deseq2, sendo assim, pode ser sugerido para detecção eficiente de DE-miRNAs, ao demonstrar altas taxas de acurácia na diferenciação de T vs. N em estadiamentos iniciais. Contudo, vale ressaltar que neste estudo é evidenciado que a escolha da metodologia pode influenciar nos resultados e consequentemente nas interpretações biológicas, sendo assim, sempre é importante estudar e analisar metodologias e ferramentas antes de utilizá-las.
- **Análises II e III:** Os Testes de Wilcoxon-Mann-Whitney, com correção de Bonferroni, realizados nas análises identificaram de 18% a 33% DE-miRNAs ou DEGs. As análises biológicas foram fundamentais para identificar pares de DE-miRNAs e DEGs com expressões opostas, correlações negativas e validados experimentalmente, foram identificados nas análises de 80 a 160 pares. Identificaram-se importantes vias biológicas relacionadas ao câncer, redes mostrando as ligações entre os pares e, por fim, a PPI confirmou a relação funcional e os fenômenos biológicos associados. As análises estatísticas em conjunto possibilitaram entender a importância desses DE-miRNAs e DEGs identificados para a discriminação

do tipo de tecido ou da doença e, também, para a sobrevivência dos pacientes. Assim, seguindo o fluxograma proposto, nas análises II e III, foram identificados importantes e potenciais RNAs e miRNAs para os pacientes em estadiamento inicial.

Por fim, foram identificadas valiosas informações a respeito de DE-miRNAs e DEGs que obtiveram comportamentos diferentes dos demais, podendo ser candidatos a potenciais biomarcadores de LUAD e LUSC em estadiamento inicial, e, além disso, apresentou árvores tanto para a classificação, quanto para a sobrevivência, que possuem potencial a serem mais investigadas e posteriormente utilizadas para diagnóstico e entendimento a respeito do tipo de tecido e subtipo da doença.

Referências

- AGRESTI, A. A survey of exact inference for contingency tables. *Statistical science*, Institute of Mathematical Statistics, v. 7, n. 1, p. 131–153, 1992. [36](#)
- AGRESTI, A. *An introduction to categorical data analysis*. [S.l.]: John Wiley & Sons, 2018. [36](#)
- ALBERG, A. J.; BROCK, M. V.; SAMET, J. M. Epidemiology of lung cancer: looking to the future. *Journal of clinical oncology*, American Society of Clinical Oncology, v. 23, n. 14, p. 3175–3185, 2005. [7](#)
- ALBERTS, B. *et al. Molecular biology of the cell*. [S.l.]: Garland science, 2014. [12](#)
- AMBROS, V. The functions of animal micrnas. *Nature*, Nature Publishing Group UK London, v. 431, n. 7006, p. 350–355, 2004. [16](#)
- AMIT, Y.; GEMAN, D. Shape quantization and recognition with randomized trees. *Neural computation*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., v. 9, n. 7, p. 1545–1588, 1997. [30](#)
- ANUSEWICZ, D.; ORZECOWSKA, M.; BEDNAREK, A. K. Lung squamous cell carcinoma and lung adenocarcinoma differential gene expression regulation through pathways of notch, hedgehog, wnt, and erbb signalling. *Scientific reports*, Springer, v. 10, n. 1, p. 1–15, 2020. [8](#), [9](#)
- ARDEKANI, A. M.; NAEINI, M. M. The role of micrnas in human diseases. *Avicenna journal of medical biotechnology*, Avicenna Research Institute, v. 2, n. 4, p. 161, 2010. [18](#)
- AZIZI, M. I. H. N.; OTHMAN, I.; NAIDU, R. The role of micrnas in lung cancer metabolism. *Cancers*, MDPI, v. 13, n. 7, p. 1716, 2021. [20](#)
- BARANAUSKAS, J. A.; MONARD, M. C. Reviewing some machine learning concepts and methods. 2000. [43](#)
- BARBA-ALIAGA, M.; ALEPUZ, P.; PÉREZ-ORTÍN, J. E. Eukaryotic rna polymerases: the many ways to transcribe a gene. *Frontiers in Molecular Biosciences*, Frontiers Media SA, v. 8, p. 663209, 2021. [17](#)
- BARBACID, M. Ras genes. *Annual review of biochemistry*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 56, n. 1, p. 779–827, 1987. [21](#)
- BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, Wiley Online Library, v. 57, n. 1, p. 289–300, 1995. [24](#), [38](#), [40](#)
- BHAYANI, M. K.; CALIN, G. A.; LAI, S. Y. Functional relevance of mirna* sequences in human disease. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, Elsevier, v. 731, n. 1-2, p. 14–19, 2012. [20](#)
- BIAU, G.; SCORNET, E. A random forest guided tour. *Test*, Springer, v. 25, p. 197–227, 2016. [30](#), [59](#), [62](#)
- BIORENDER. *BioRender*. 2022. [ix](#), [13](#), [17](#)

- BISCHL, B. *et al.* Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, p. e1484, 2021. [61](#)
- BLAGUS, R.; LUSA, L. Smote for high-dimensional class-imbalanced data. *BMC bioinformatics*, Springer, v. 14, p. 1–16, 2013. [46](#)
- BOERI, M. *et al.* Microrna signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 108, n. 9, p. 3713–3718, 2011. [150](#)
- BOFFETTA, P.; JÄRVHOLM, B.; BRENNAN, P.; NYRÉN, O. Incidence of lung cancer in a large cohort of non-smoking men from sweden. *International journal of cancer*, Wiley Online Library, v. 94, n. 4, p. 591–593, 2001. [6](#)
- BOGEDALE, K.; JAGANNATHAN, V.; GERBER, V.; UNGER, L. Differentially expressed micrnas, including a large microrna cluster on chromosome 24, are associated with equine sarcoid and squamous cell carcinoma. *Veterinary and comparative oncology*, Wiley Online Library, v. 17, n. 2, p. 155–164, 2019. [80](#)
- BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, p. 123–140, 1996. [30](#)
- BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001. [30](#), [58](#), [59](#)
- BREIMAN, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, Institute of Mathematical Statistics, v. 16, n. 3, p. 199–231, 2001. [41](#)
- BREIMAN, L.; CUTLER, A. *Random forests: Classification/clustering*. [S.l.]: Retrieved May, 2004. [59](#)
- BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Classification and regression trees. *CA: Wadsworth International Group*, n. 368, 1984. [29](#), [47](#), [55](#), [56](#), [60](#)
- CALIN, G. A. *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*, National Acad Sciences, v. 99, n. 24, p. 15524–15529, 2002. [21](#)
- CAMARGO, B. da R. *et al.* A decision tree-based classifier compares three data analysis methods for the identification of mirnas associated with early-stage lung cancer. *REVISTA FOCO*, v. 16, n. 5, p. e2031–e2031, 2023. [ix](#), [68](#), [69](#), [70](#), [72](#), [75](#), [76](#), [77](#), [78](#)
- CHANG, J. T.-H.; LEE, Y. M.; HUANG, R. S. The impact of the cancer genome atlas on lung cancer. *Translational Research*, Elsevier, v. 166, n. 6, p. 568–585, 2015. [22](#)
- CHANG, W.-H. *et al.* Jag1 is associated with poor survival through inducing metastasis in lung cancer. *PloS one*, Public Library of Science San Francisco, CA USA, v. 11, n. 3, p. e0150355, 2016. [150](#)
- CHARKIEWICZ, R. *et al.* Gene expression signature differentiates histology but not progression status of early-stage nslc. *Translational oncology*, Elsevier, v. 10, n. 3, p. 450–458, 2017. [9](#)
- CHAUDHURI, P.; LOH, W.-Y. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, JSTOR, p. 561–576, 2002. [29](#)

- CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. 46
- CHEN, B. *et al.* A regulatory circuitry comprising tp53, mir-29 family, and setdb1 in non-small cell lung cancer. *Bioscience reports*, Portland Press Ltd., v. 38, n. 5, p. BSR20180678, 2018. 129
- CHEN, H. *VennDiagram: Generate High-Resolution Venn and Euler Plots*. [S.l.], 2022. R package version 1.7.3. Disponível em: <<https://CRAN.R-project.org/package=VennDiagram>>. 40
- CHEN, J. W.; DHAHBI, J. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Scientific reports*, Nature Publishing Group UK London, v. 11, n. 1, p. 13323, 2021. 30
- CHEN, L. *et al.* Machine learning and network methods for biology and medicine. *Computational and Mathematical Methods in Medicine*, Hindawi, v. 2015, 2015. 104
- CHEN, M. *et al.* Differentiated regulation of immune-response related genes between luad and lusc subtypes of lung cancers. *Oncotarget*, Impact Journals, LLC, v. 8, n. 1, p. 133, 2017. 1, 8
- CHEN, Q. *et al.* Microrna-301a promotes growth and migration by repressing tgfr2 in non-small cell lung cancer. *Int J Clin Exp Pathol*, v. 10, n. 2, p. 957–971, 2017. 103
- CHEN, Q. *et al.* mir-210-3p promotes lung cancer development and progression by modulating usf1 and pcg3. *Oncotargets and therapy*, Dove Press, v. 14, p. 3687, 2021. 129
- CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. *Genomics*, Elsevier, v. 99, n. 6, p. 323–329, 2012. 27, 28, 30
- CHI, X.-J. *et al.* Identification of high expression profiles of mir-31-5p and its vital role in lung squamous cell carcinoma: a survey based on qrt-pcr and bioinformatics analysis. *Translational Cancer Research*, AME Publications, v. 8, n. 3, p. 788, 2019. 130
- CIMMINO, A. *et al.* mir-15 and mir-16 induce apoptosis by targeting bcl2. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 102, n. 39, p. 13944–13949, 2005. 22
- CINEGAGLIA, N. C. *et al.* Integrative transcriptome analysis identifies deregulated microrna-transcription factor networks in lung adenocarcinoma. *Oncotarget*, Impact Journals, LLC, v. 7, n. 20, p. 28920, 2016. 21
- COLOSIMO, E. A.; GIOLO, S. R. Análise de sobrevivência. *São Paulo: Abe-Projeto Fisher*, 2006. 51, 52
- CORCHETE, L. A. *et al.* Systematic comparison and assessment of rna-seq procedures for gene expression quantitative analysis. *Scientific reports*, Nature Publishing Group UK London, v. 10, n. 1, p. 19737, 2020. 80
- CORRALES, L. *et al.* Lung cancer in never smokers: The role of different risk factors other than tobacco smoking. *Critical reviews in oncology/hematology*, Elsevier, v. 148, p. 102895, 2020. 6
- COURAUD, S. *et al.* Lung cancer in never smokers—a review. *European journal of cancer*, Elsevier, v. 48, n. 9, p. 1299–1311, 2012. 6

- COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. [53](#), [54](#)
- COX, D. R. Partial likelihood. *Biometrika*, Oxford University Press, v. 62, n. 2, p. 269–276, 1975. [54](#)
- CRUZ, C. S. D.; TANOUE, L. T.; MATTHAY, R. A. Lung cancer: epidemiology, etiology, and prevention. *Clinics in chest medicine*, Elsevier, v. 32, n. 4, p. 605–644, 2011. [6](#)
- DAVID, C. J.; MASSAGUÉ, J. Contextual determinants of $\text{tgf}\beta$ action in development, immunity and cancer. *Nature reviews Molecular cell biology*, Nature Publishing Group UK London, v. 19, n. 7, p. 419–435, 2018. [103](#)
- DENISON, D. G.; MALLICK, B. K.; SMITH, A. F. A bayesian cart algorithm. *Biometrika*, Oxford University Press, v. 85, n. 2, p. 363–377, 1998. [29](#)
- DETTERBECK, F. C.; BOFFA, D. J.; KIM, A. W.; TANOUE, L. T. The eighth edition lung cancer stage classification. *Chest*, Elsevier, v. 151, n. 1, p. 193–203, 2017. [10](#)
- DEVROYE, L.; GYÖRFI, L.; LUGOSI, G. *A probabilistic theory of pattern recognition*. [S.l.]: Springer Science & Business Media, 2013. v. 31. [59](#)
- DIAS, M. *et al.* Lung cancer in never-smokers—what are the differences? *Acta oncologica*, Taylor & Francis, v. 56, n. 7, p. 931–935, 2017. [7](#)
- DÍAZ-URIARTE, R.; ANDRÉS, S. Alvarez de. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, Springer, v. 7, p. 1–13, 2006. [58](#)
- DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., v. 10, n. 7, p. 1895–1923, 1998. [41](#)
- DIETTERICH, T. G. Ensemble methods in machine learning. In: SPRINGER. *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings I*. [S.l.], 2000. p. 1–15. [30](#)
- DONSON, A. M. *et al.* Mgmt promoter methylation correlates with survival benefit and sensitivity to temozolomide in pediatric glioblastoma. *Pediatric blood & cancer*, Wiley Online Library, v. 48, n. 4, p. 403–407, 2007. [14](#)
- DUBIN, S.; GRIFFIN, D. Lung cancer in non-smokers. *Missouri medicine*, Missouri State Medical Association, v. 117, n. 4, p. 375, 2020. [1](#), [6](#), [7](#)
- EFRON, B. Bootstrap methods: another look at the jackknife. In: *Breakthroughs in statistics: Methodology and distribution*. [S.l.]: Springer, 1992. p. 569–593. [58](#)
- ELGELDAWI, E.; SAYED, A.; GALAL, A. R.; ZAKI, A. M. Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In: MULTIDISCIPLINARY DIGITAL PUBLISHING INSTITUTE. *Informatics*. [S.l.], 2021. v. 8, n. 4, p. 79. [61](#)
- ESTELLER, M. *et al.* Inactivation of the dna repair gene o 6-methylguanine-dna methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. *Cancer research*, AACR, v. 59, n. 4, p. 793–797, 1999. [14](#)

- ETTINGER, D. S. Overview and state of the art in the management of lung cancer. *Oncology (Williston Park, NY)*, v. 18, n. 7 Suppl 4, p. 3–9, 2004. [7](#)
- FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. Inteligência artificial: uma abordagem de aprendizado de máquina. 2011. [ix](#), [27](#), [28](#), [41](#), [42](#), [43](#), [44](#), [47](#), [50](#), [62](#)
- FARUKI, H. *et al.* Lung adenocarcinoma and squamous cell carcinoma gene expression subtypes demonstrate significant differences in tumor immune landscape. *Journal of Thoracic Oncology*, Elsevier, v. 12, n. 6, p. 943–953, 2017. [9](#)
- FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. [64](#)
- FENG, Y.-H.; TSAO, C.-J. Emerging role of microRNA-21 in cancer. *Biomedical reports*, Spandidos Publications, v. 5, n. 4, p. 395–402, 2016. [22](#)
- FISHER, R. A. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the royal statistical society*, JSTOR, v. 85, n. 1, p. 87–94, 1922. [36](#)
- FISHER, R. A. *Statistical methods for research workers 12 th ed.* [S.l.]: Oliver & Body, 1954. [36](#)
- FONSECA, J. da. *Indução de Árvores de decisão*. Tese (Doutorado) — Dissertação de Mestrado, Universidade Nova de Lisboa, Lisboa, 1994. [48](#)
- FOSS, K. M. *et al.* mir-1254 and mir-574-5p: serum-based microRNA biomarkers for early-stage non-small cell lung cancer. *Journal of thoracic oncology*, Elsevier, v. 6, n. 3, p. 482–488, 2011. [22](#)
- FOTOUHI, S.; ASADI, S.; KATTAN, M. W. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of biomedical informatics*, Elsevier, v. 90, p. 103089, 2019. [45](#)
- FRANKISH, A. *et al.* Gencode 2021. *Nucleic acids research*, Oxford University Press, v. 49, n. D1, p. D916–D923, 2021. [34](#)
- FRIEDMAN, J. H.; KOHAVI, R.; YUN, Y. Lazy decision trees. In: *AAAI/IAAI, Vol. 1*. [S.l.: s.n.], 1996. p. 717–724. [48](#)
- GAO, M.; KONG, W.; HUANG, Z.; XIE, Z. Identification of key genes related to lung squamous cell carcinoma using bioinformatics analysis. *International journal of molecular sciences*, MDPI, v. 21, n. 8, p. 2994, 2020. [8](#)
- GAUTHIER, M.; AGNIEL, D.; THIÉBAUT, R.; HEJBLUM, B. P. dearseq: a variance component score test for rna-seq differential analysis that effectively controls the false discovery rate. *NAR genomics and bioinformatics*, Oxford University Press, v. 2, n. 4, p. lqaa093, 2020. [25](#), [80](#)
- GAZDAR, A. Activating and resistance mutations of egfr in non-small-cell lung cancer: role in clinical response to egfr tyrosine kinase inhibitors. *Oncogene*, Nature Publishing Group, v. 28, n. 1, p. S24–S31, 2009. [12](#)
- GENG, Q. *et al.* Five microRNAs in plasma as novel biomarkers for screening of early-stage non-small cell lung cancer. *Respiratory research*, BioMed Central, v. 15, n. 1, p. 1–9, 2014. [79](#)

- GENTLES, A. J. *et al.* Integrating tumor and stromal gene expression signatures with clinical indices for survival stratification of early-stage non-small cell lung cancer. *JNCI: Journal of the National Cancer Institute*, Oxford University Press, v. 107, n. 10, 2015. 9
- GIOLO, S. R. *Introdução à análise de dados categóricos com aplicações*. [S.l.]: Editora Blucher, 2017. 35, 36, 37, 54
- GIRBIG, M.; MISIASZEK, A. D.; MÜLLER, C. W. Structural insights into nuclear transcription by eukaryotic dna-dependent rna polymerases. *Nature Reviews Molecular Cell Biology*, Nature Publishing Group UK London, v. 23, n. 9, p. 603–622, 2022. 17
- GOLDSTRAW, P. *et al.* The iaslc lung cancer staging project: proposals for revision of the tnm stage groupings in the forthcoming (eighth) edition of the tnm classification for lung cancer. *Journal of Thoracic Oncology*, Elsevier, v. 11, n. 1, p. 39–51, 2016. 11
- GORLOVA, O. Y. *et al.* Aggregation of cancer among relatives of never-smoking lung cancer patients. *International journal of cancer*, Wiley Online Library, v. 121, n. 1, p. 111–118, 2007. 7
- GRIDELLI, C. *et al.* Non-small-cell lung cancer. *Nature reviews Disease primers*, Nature Publishing Group, v. 1, n. 1, p. 1–16, 2015. 8
- HAMFJORD, J. *et al.* Differential expression of mirnas in colorectal cancer: comparison of paired tumor tissue and adjacent normal mucosa using high-throughput sequencing. *PloS one*, Public Library of Science San Francisco, USA, v. 7, n. 4, p. e34150, 2012. 80
- HAMNER, B.; FRASCO, M. *Metrics: Evaluation Metrics for Machine Learning*. [S.l.], 2018. R package version 0.1.4. Disponível em: <<https://CRAN.R-project.org/package=Metrics>>. 65
- Harrell Jr, F. E. *Hmisc: Harrell Miscellaneous*. [S.l.], 2023. R package version 5.0-1. Disponível em: <<https://CRAN.R-project.org/package=Hmisc>>. 40
- HART, A. Mann-whitney test is not just a test of medians: differences in spread can be important. *Bmj*, British Medical Journal Publishing Group, v. 323, n. 7309, p. 391–393, 2001. 37
- HE, J.-H. *et al.* Analyzing the lncrna, mirna, and mrna regulatory network in prostate cancer with bioinformatics software. *Journal of Computational Biology*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 25, n. 2, p. 146–157, 2018. 80
- HENNESSEY, P. T. *et al.* Serum microrna biomarkers for detection of non-small cell lung cancer. *PloS one*, Public Library of Science San Francisco, USA, v. 7, n. 2, p. e32307, 2012. 22
- HERBST, R. S.; JR, P. A. B. Targeting the epidermal growth factor receptor in non-small cell lung cancer. *Clinical Cancer Research*, AACR, v. 9, n. 16, p. 5813–5824, 2003. 9
- HIRONO, T. *et al.* Microrna-130b functions as an oncomirna in non-small cell lung cancer by targeting tissue inhibitor of metalloproteinase-2. *Scientific Reports*, Nature Publishing Group UK London, v. 9, n. 1, p. 6956, 2019. 103
- HO, T. K. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, Ieee, v. 20, n. 8, p. 832–844, 1998. 30
- HOGEWEG, P. The roots of bioinformatics in theoretical biology. *PLoS computational biology*, Public Library of Science San Francisco, USA, v. 7, n. 3, p. e1002021, 2011. 23

- HOLM, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, JSTOR, p. 65–70, 1979. 38
- HOOD, L.; ROWEN, L. The human genome project: big science transforms biology and medicine. *Genome medicine*, Springer, v. 5, p. 1–8, 2013. 23
- HOTHORN, T.; HORNIK, K.; ZEILEIS, A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, Taylor & Francis, v. 15, n. 3, p. 651–674, 2006. 29
- HOTHORN, T.; ZEILEIS, A. partykit: A modular toolkit for recursive partytioning in r. *The Journal of Machine Learning Research*, JMLR. org, v. 16, n. 1, p. 3905–3909, 2015. 48, 50
- HOU, J. *et al.* Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PloS one*, Public Library of Science San Francisco, USA, v. 5, n. 4, p. e10312, 2010. 9
- HU, J. *et al.* Mir-497-5p down-regulates cdca4 to restrains lung squamous cell carcinoma progression. *Journal of Cardiothoracic Surgery*, Springer, v. 16, p. 1–8, 2021. 130
- HU, Y. *et al.* Identification of key differentially expressed micrnas in cancer patients through pan-cancer analysis. *Computers in biology and medicine*, Elsevier, v. 103, p. 183–197, 2018. 80
- HUANG, C.-Y. *et al.* A review on the effects of current chemotherapy drugs and natural agents in treating non–small cell lung cancer. *Biomedicine*, China Medical University, v. 7, n. 4, 2017. 11
- HUANG, L. *et al.* Let-7c-5p represses cisplatin resistance of lung adenocarcinoma cells by targeting cdc25a. *Applied Biochemistry and Biotechnology*, Springer, v. 195, n. 3, p. 1644–1655, 2023. 104
- HUYNH, P.-H.; NGUYEN, V. H.; DO, T.-N. Improvements in the large p, small n classification issue. *SN Computer Science*, Springer, v. 1, p. 1–19, 2020. 30
- INAMURA, K.; ISHIKAWA, Y. Microrna in lung cancer: Novel biomarkers and potential tools for treatment. *Journal of Clinical Medicine*, v. 5, n. 3, 2016. ISSN 2077-0383. Disponível em: <<https://www.mdpi.com/2077-0383/5/3/36>>. ix, 19, 20
- INAMURA, K. *et al.* let-7 microrna expression is reduced in bronchioloalveolar carcinoma, a non-invasive carcinoma, and is not correlated with prognosis. *Lung cancer*, Elsevier, v. 58, n. 3, p. 392–396, 2007. 22
- INCA. Instituto Nacional de Câncer José Alencar Gomes da Silva. *Estimativa 2025: Incidência de Câncer no Brasil*. Rio de Janeiro. 2022. <https://www.gov.br/inca/pt-br/assuntos/noticias/2022/inca-estima-704-mil-casos-de-cancer-por-ano-no-brasil-ate-2025>. Acesso em 25 de janeiro de 2023. 5
- IORIO, M. V.; CROCE, C. M. Microrna dysregulation in cancer: diagnostics, monitoring and therapeutics. a comprehensive review. *EMBO molecular medicine*, WILEY-VCH Verlag Weinheim, v. 4, n. 3, p. 143–159, 2012. 17
- ISHWARAN, H. *et al.* High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, Taylor & Francis, v. 105, n. 489, p. 205–217, 2010. 30

- ISLAM, K. *et al.* Comorbidity and survival in lung cancer patients. *Cancer Epidemiology, Biomarkers & Prevention*, AACR, v. 24, n. 7, p. 1079–1085, 2015. [6](#)
- IWAMOTO, T. *et al.* Distinct gene expression profiles between primary breast cancers and brain metastases from pair-matched samples. *Scientific reports*, Nature Publishing Group UK London, v. 9, n. 1, p. 13343, 2019. [25](#)
- IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.]: Rafael Izbicki, 2020. [43](#)
- JANSSEN-HEIJNEN, M. L. *et al.* Prevalence of co-morbidity in lung cancer patients and its relationship with treatment: a population-based study. *Lung cancer*, Elsevier, v. 21, n. 2, p. 105–113, 1998. [6](#)
- JOHNSON, S. M. *et al.* Ras is regulated by the let-7 microRNA family. *Cell*, Elsevier, v. 120, n. 5, p. 635–647, 2005. [22](#)
- KADARA, H. *et al.* Early diagnosis and screening for lung cancer. *Cold Spring Harbor perspectives in medicine*, Cold Spring Harbor Laboratory Press, v. 11, n. 9, p. a037994, 2021. [1](#)
- KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, Oxford University Press, v. 28, n. 1, p. 27–30, 2000. [39](#), [40](#)
- KANTHAJE, S.; BAIKUNJE, N.; KANDAL, I.; RATNACARAM, C. K. Repertoires of microRNA-30 family as gate-keepers in lung cancer. *Frontiers in Bioscience-Scholar*, IMR Press, v. 13, n. 2, p. 141–156, 2021. [104](#), [129](#)
- KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. [52](#)
- KASS, G. V. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 29, n. 2, p. 119–127, 1980. [29](#)
- KAUR, B.; KUMAR, S.; KAUSHIK, B. K. Recent advancements in optical biosensors for cancer detection. *Biosensors and Bioelectronics*, Elsevier, v. 197, p. 113805, 2022. [7](#)
- KAWAGUCHI, T. *et al.* Japanese ethnicity compared with caucasian ethnicity and never-smoking status are independent favorable prognostic factors for overall survival in non-small cell lung cancer: a collaborative epidemiologic study of the national hospital organization study group for lung cancer (nhsglc) in japan and a southern california regional cancer registry databases. *Journal of Thoracic Oncology*, Elsevier, v. 5, n. 7, p. 1001–1010, 2010. [6](#)
- KENT, O.; MENDELL, J. A small piece in the cancer puzzle: microRNAs as tumor suppressors and oncogenes. *Oncogene*, Nature Publishing Group, v. 25, n. 46, p. 6188–6196, 2006. [20](#)
- KERN, F. *et al.* mieaa 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic acids research*, Oxford University Press, v. 48, n. W1, p. W521–W528, 2020. [39](#)
- KIM, H.; LOH, W.-Y. Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, Taylor & Francis, v. 96, n. 454, p. 589–604, 2001. [29](#)

- KIM, V. N.; NAM, J.-W. Genomics of microRNA. *TRENDS in Genetics*, Elsevier, v. 22, n. 3, p. 165–173, 2006. 17, 18, 20
- KINGSFORD, C.; SALZBERG, S. L. What are decision trees? *Nature biotechnology*, Nature Publishing Group US New York, v. 26, n. 9, p. 1011–1013, 2008. 30
- KLECZKO, E. K.; KWAK, J. W.; SCHENK, E. L.; NEMENOFF, R. A. Targeting the complement pathway as a therapeutic strategy in lung cancer. *Frontiers in immunology*, Frontiers Media SA, v. 10, p. 954, 2019. 9
- KOBAYASHI, S. *et al.* Egfr mutation and resistance of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, Mass Medical Soc, v. 352, n. 8, p. 786–792, 2005. 12
- KOTHANDAN, R. Handling class imbalance problem in mirna dataset associated with cancer. *Bioinformatics*, Biomedical Informatics Publishing Group, v. 11, n. 1, p. 6, 2015. 46
- KOZOMARA, A.; BIRGAOANU, M.; GRIFFITHS-JONES, S. mirbase: from microRNA sequences to function. *Nucleic acids research*, Oxford University Press, v. 47, n. D1, p. D155–D162, 2019. 20
- KRIS, M. G. *et al.* Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. *Jama*, American Medical Association, v. 311, n. 19, p. 1998–2006, 2014. 21
- KUBAT, M.; BRATKO, I.; MICHALSKI, R. S. A review of machine learning methods. *Machine learning and data mining: methods and applications*, Citeseer, p. 3–69, 1998. 47
- KUBAT, M.; KUBAT, J. *An introduction to machine learning*. [S.l.]: Springer, 2017. v. 2. 45
- KUHN, M. *caret: Classification and Regression Training*. [S.l.], 2022. R package version 6.0-93. Disponível em: <<https://CRAN.R-project.org/package=caret>>. 64
- KUHN, M.; VAUGHAN, D.; HVITFELDT, E. *yardstick: Tidy Characterizations of Model Performance*. [S.l.], 2022. R package version 1.1.0. Disponível em: <<https://CRAN.R-project.org/package=yardstick>>. 65
- KUO, W.-T. *et al.* Bioinformatic interrogation of 5p-arm and 3p-arm specific microRNA expression using tcga datasets. *Journal of clinical medicine*, MDPI, v. 4, n. 9, p. 1798–1814, 2015. 34
- LANDER, E.; LINTON, L.; BIRREN, B. Initial sequencing and analysis of the human genome [published correction appears in nature. 2001; 411 (6838): 720]. *Nature*, v. 409, n. 6822, p. 860–921, 2001. 12
- LANE, D. P.; CRAWFORD, L. V. T antigen is bound to a host protein in sy40-transformed cells. *Nature*, Nature Publishing Group UK London, v. 278, n. 5701, p. 261–263, 1979. 21
- LEBANONY, D. *et al.* Diagnostic assay based on hsa-mir-205 expression distinguishes squamous from nonsquamous non–small-cell lung carcinoma. *Journal of clinical oncology*, American Society of Clinical Oncology, v. 27, n. 12, p. 2030–2037, 2009. 128, 149
- 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. ISSN 0092-8674. Disponível em: <<https://www.sciencedirect.com/science/article/pii/009286749390529Y>>. 18

- LEE, Y.-h.; BANG, H.; KIM, D. J. How to establish clinical prediction models. *Endocrinology and Metabolism*, Korean Endocrine Society, v. 31, n. 1, p. 38–44, 2016. 41
- LI, L. *et al.* Microrna biomarker hsa-mir-195-5p for detecting the risk of lung cancer. *International Journal of Genomics*, Hindawi, v. 2020, 2020. 104, 148
- LI, Y. *et al.* Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome biology*, Springer, v. 23, n. 1, p. 79, 2022. 24, 25, 79
- LIANG, R. *et al.* Snhg6 functions as a competing endogenous rna to regulate e2f7 expression by sponging mir-26a-5p in lung adenocarcinoma. *Biomedicine & Pharmacotherapy*, Elsevier, v. 107, p. 1434–1446, 2018. 148
- LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002. Disponível em: <<https://CRAN.R-project.org/doc/Rnews/>>. 61, 62
- LIM, E. L. *et al.* Microrna expression-based model indicates event-free survival in pediatric acute myeloid leukemia. *Journal of Clinical Oncology*, American Society of Clinical Oncology, v. 35, n. 35, p. 3964, 2017. 25
- LIM, H. J.; CROWE, P.; YANG, J.-L. Current clinical regulation of pi3k/pten/akt/mtor signalling in treatment of human cancer. *Journal of cancer research and clinical oncology*, Springer, v. 141, p. 671–689, 2015. 129
- LIN, T.-C. *et al.* Microrna-184 deregulated by the microrna-21 promotes tumor malignancy and poor outcomes in non-small cell lung cancer via targeting cdc25a and c-myc. *Annals of surgical oncology*, Springer, v. 22, p. 1532–1539, 2015. 105
- LIU, M.; ZHOU, K.; CAO, Y. Microrna-944 affects cell growth by targeting epha7 in non-small cell lung cancer. *International journal of molecular sciences*, MDPI, v. 17, n. 10, p. 1493, 2016. 149
- LOH, W.-Y. Fifty years of classification and regression trees. *International Statistical Review*, Wiley Online Library, v. 82, n. 3, p. 329–348, 2014. 28
- LOH, W.-Y.; SHIH, Y.-S. Split selection methods for classification trees. *Statistica sinica*, JSTOR, p. 815–840, 1997. 29
- LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, BioMed Central, v. 15, n. 12, p. 1–21, 2014. 24
- LU, T. X.; ROTHENBERG, M. E. Microrna. *Journal of allergy and clinical immunology*, Elsevier, v. 141, n. 4, p. 1202–1207, 2018. 18
- LU, Y. *et al.* Microrna profiling and prediction of recurrence/relapse-free survival in stage i lung cancer. *Carcinogenesis*, Oxford University Press, v. 33, n. 5, p. 1046–1054, 2012. 150
- LÜCHTENBORG, M. *et al.* The effect of comorbidity on stage-specific survival in resected non-small cell lung cancer patients. *European journal of cancer*, Elsevier, v. 48, n. 18, p. 3386–3395, 2012. 6
- LUNARDON, N.; MENARDI, G.; TORELLI, N. ROSE: a Package for Binary Imbalanced Learning. *R Journal*, v. 6, n. 1, p. 82–92, 2014. 64

- LYNCH, T. J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *New England Journal of Medicine*, Mass Medical Soc, v. 350, n. 21, p. 2129–2139, 2004. [12](#)
- MA, J. *et al.* Bioinformatic analysis reveals an exosomal mirna-mrna network in colorectal cancer. *BMC medical genomics*, BioMed Central, v. 14, n. 1, p. 1–18, 2021. [80](#)
- MAK, D. W.; LI, S.; MINCHOM, A. Challenging the recalcitrant disease—developing molecularly driven treatments for small cell lung cancer. *European Journal of Cancer*, Elsevier, v. 119, p. 132–150, 2019. [11](#)
- 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*, JSTOR, p. 50–60, 1947. [25](#), [37](#)
- MANTEL, N. *et al.* Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, v. 50, n. 3, p. 163–170, 1966. [52](#)
- MARIAN, A. Sequencing your genome: what does it mean? *Methodist DeBakey cardiovascular journal*, Methodist DeBakey Heart & Vascular Center, v. 10, n. 1, p. 3, 2014. [14](#)
- MEISTER, G.; TUSCHL, T. Mechanisms of gene silencing by double-stranded rna. *Nature*, Nature Publishing Group, v. 431, n. 7006, p. 343–349, 2004. [17](#)
- MENCK, C. F.; SLUYS, M. *Genética molecular básica: dos genes ao genoma*. [S.l.]: Editora Guanabara Koogan: Rio de Janeiro, Brazil, 2017. [15](#)
- MESSENGER, R.; MANDELL, L. A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American statistical association*, Taylor & Francis, v. 67, n. 340, p. 768–772, 1972. [29](#)
- MICHIE, D.; SPIEGELHALTER, D. J.; TAYLOR, C. C.; CAMPBELL, J. *Machine learning, neural and statistical classification*. [S.l.]: Ellis Horwood, 1995. [43](#)
- MILLER, K. D. *et al.* Cancer statistics for the us hispanic/latino population, 2021. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 71, n. 6, p. 466–487, 2021. [4](#)
- MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-hill New York, 1997. [27](#)
- MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003. [62](#)
- MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, Taylor & Francis, v. 58, n. 302, p. 415–434, 1963. [29](#)
- MUELLER, J. P.; MASSARON, L. *Machine learning for dummies*. [S.l.]: John Wiley & Sons, 2021. [43](#)
- MYERS, D. J.; WALLEN, J. M. Lung adenocarcinoma. In: *StatPearls [Internet]*. [S.l.]: StatPearls Publishing, 2022. [8](#)
- NESBITT, J. C. *et al.* Survival in early-stage non-small cell lung cancer. *The Annals of thoracic surgery*, Elsevier, v. 60, n. 2, p. 466–472, 1995. [11](#)

- NEVE, J. D.; THAS, O.; OTTOY, J.-P.; CLEMENT, L. An extension of the wilcoxon-mann-whitney test for analyzing rt-qpcr data. *Statistical Applications in Genetics and Molecular Biology*, De Gruyter, v. 12, n. 3, p. 333–346, 2013. [24](#)
- NYBERG, F. *et al.* A european validation study of smoking and environmental tobacco smoke exposure in nonsmoking lung cancer cases and controls. *Cancer Causes & Control*, Springer, v. 9, p. 173–182, 1998. [7](#)
- ÖBERG, M. *et al.* Worldwide burden of disease from exposure to second-hand smoke: a retrospective analysis of data from 192 countries. *The lancet*, Elsevier, v. 377, n. 9760, p. 139–146, 2011. [1](#), [6](#)
- O'BRIEN, J.; HAYDER, H.; ZAYED, Y.; PENG, C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in endocrinology*, Frontiers Media SA, v. 9, p. 402, 2018. [17](#)
- OLIVEROS, J. C. Venny. an interactive tool for comparing lists with venn diagrams. <http://bioinfogp.cnb.csic.es/tools/venny/index.html>, 2007. [39](#)
- OMS, O. M. da S. *Estimativas Globais de Saúde 2020: Mortes por Causa, Idade, Sexo, por País e por Região, 2000-2019*. 2020. <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death>. Acesso em 12 de dezembro de 2022. [3](#)
- PAEZ, J. G. *et al.* Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, American Association for the Advancement of Science, v. 304, n. 5676, p. 1497–1500, 2004. [14](#)
- PARKIN, D. M.; BRAY, F.; FERLAY, J.; PISANI, P. Global cancer statistics, 2002. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 55, n. 2, p. 74–108, 2005. [7](#)
- 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*, Taylor & Francis, v. 50, n. 302, p. 157–175, 1900. [35](#)
- PENG, Y.; CROCE, C. M. The role of microRNAs in human cancer. *Signal transduction and targeted therapy*, Nature Publishing Group, v. 1, n. 1, p. 1–9, 2016. [24](#)
- PERNEGER, T. V. What's wrong with bonferroni adjustments. *Bmj*, British Medical Journal Publishing Group, v. 316, n. 7139, p. 1236–1238, 1998. [38](#)
- PETO, R. *et al.* Smoking, smoking cessation, and lung cancer in the uk since 1950: combination of national statistics with two case-control studies. *Bmj*, British Medical Journal Publishing Group, v. 321, n. 7257, p. 323–329, 2000. [6](#)
- PFEFFER, S. R.; YANG, C. H.; PFEFFER, L. M. The role of mir-21 in cancer. *Drug development research*, Wiley Online Library, v. 76, n. 6, p. 270–277, 2015. [22](#)
- PIKOR, L. A.; RAMNARINE, V. R.; LAM, S.; LAM, W. L. Genetic alterations defining nslc subtypes and their therapeutic implications. *Lung cancer*, Elsevier, v. 82, n. 2, p. 179–189, 2013. [1](#), [8](#)

- PRATI, R. C. *Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos*. Tese (Doutorado) — Universidade de São Paulo, 2006. 65
- PRATI, R. C.; FLACH, P. A. Roccer: An algorithm for rule learning based on roc analysis. In: *Ijcai*. [S.l.: s.n.], 2005. p. 823–828. 64
- PROBST, P.; BOULESTEIX, A.-L.; BISCHL, B. Tunability: Importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, JMLR. org, v. 20, n. 1, p. 1934–1965, 2019. 62
- QUINLAN, J. C4. 5: Programs for machine learning, san mateo 1993. *Google Scholar Google Scholar Digital Library Digital Library*, p. 1–302, 1993. 29
- QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, v. 1, p. 81–106, 1986. 29
- QUINLAN, J. R. Decision trees and multivalued attributes. *Machine intelligence 11*, Oxford University Press, p. 305–318, 1988. 55
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <<https://www.R-project.org/>>. 31, 32
- REIS, P. P. *et al.* Circulating mir-16-5p, mir-92a-3p, and mir-451a in plasma from lung cancer patients: potential application in early detection and a regulatory role in tumorigenesis pathways. *Cancers*, Multidisciplinary Digital Publishing Institute, v. 12, n. 8, p. 2071, 2020. 22
- RELLI, V.; TREROTOLA, M.; GUERRA, E.; ALBERTI, S. Abandoning the notion of non-small cell lung cancer. *Trends in Molecular Medicine*, Elsevier, v. 25, n. 7, p. 585–594, 2019. 8
- RESS, A. L. *et al.* Mir-96-5p influences cellular growth and is associated with poor survival in colorectal cancer patients. *Molecular carcinogenesis*, Wiley Online Library, v. 54, n. 11, p. 1442–1450, 2015. 104
- RHODES, D. R.; CHINNAIYAN, A. M. Integrative analysis of the cancer transcriptome. *Nature genetics*, Nature Publishing Group US New York, v. 37, n. Suppl 6, p. S31–S37, 2005. 1
- RITCHIE, M. E. *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, Oxford Academic, v. 43, n. 7, p. e47–e47, 2015. 24
- ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, Oxford University Press, v. 26, n. 1, p. 139–140, 2010. 24
- ROSSUM, G. V.; JR, D.; L, F. The python language reference manual (version 2.5). *Network Theory Ltd*, 2006. 31
- RUANO-RAVIÑA, A. *et al.* Lung cancer symptoms at diagnosis: results of a nationwide registry study. *ESMO open*, Elsevier, v. 5, n. 6, p. e001021, 2020. 11
- RUDIN, C. M.; BRAMBILLA, E.; FAIVRE-FINN, C.; SAGE, J. Small-cell lung cancer. *Nature Reviews Disease Primers*, Nature Publishing Group UK London, v. 7, n. 1, p. 3, 2021. 8
- SACHIDANANDAM, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, v. 409, n. 6822, p. 928–933, 2001. 13

- SANDLER, D. P. *et al.* Indoor radon and lung cancer risk in connecticut and utah. *Journal of Toxicology and Environmental Health, Part A*, Taylor & Francis, v. 69, n. 7-8, p. 633–654, 2006. [7](#)
- SANTOS, G. da C.; SHEPHERD, F. A.; TSAO, M. S. Egfr mutations and lung cancer. *Annual Review of Pathology: Mechanisms of Disease*, Annual Reviews, v. 6, p. 49–69, 2011. [14](#)
- SARDO, F. L. *et al.* Yap/taz and ezh2 synergize to impair tumor suppressor activity of tgfr2 in non-small cell lung cancer. *Cancer Letters*, Elsevier, v. 500, p. 51–63, 2021. [103](#)
- SARUMI, O. A.; LEUNG, C. K. Adaptive machine learning algorithm and analytics of big genomic data for gene prediction. *Tracking and preventing diseases with artificial intelligence*, Springer, p. 103–123, 2022. [23](#)
- SCHABATH, M. B.; COTE, M. L. Cancer progress and priorities: lung cancer. *Cancer epidemiology, biomarkers & prevention*, AACR, v. 28, n. 10, p. 1563–1579, 2019. [1](#)
- SCHMID, K.; KUWERT, T.; DREXLER, H. Radon in indoor spaces: an underestimated risk factor for lung cancer in environmental medicine. *Deutsches Arzteblatt international*, Deutscher Arzte-Verlag GmbH, v. 107, n. 11, p. 181, 2010. [7](#)
- SCHURCH, N. J. *et al.* How many biological replicates are needed in an rna-seq experiment and which differential expression tool should you use? *Rna*, Cold Spring Harbor Lab, v. 22, n. 6, p. 839–851, 2016. [80](#)
- SEIJO, L. M. *et al.* Biomarkers in lung cancer screening: achievements, promises, and challenges. *Journal of Thoracic Oncology*, Elsevier, v. 14, n. 3, p. 343–357, 2019. [20](#)
- SEYEDNASROLLAH, F.; LAIHO, A.; ELO, L. L. Comparison of software packages for detecting differential expression in rna-seq studies. *Briefings in bioinformatics*, Oxford University Press, v. 16, n. 1, p. 59–70, 2015. [80](#)
- SHANNON, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, Cold Spring Harbor Lab, v. 13, n. 11, p. 2498–2504, 2003. [40](#)
- SHAW, A. T.; SOLOMON, B. Targeting anaplastic lymphoma kinase in lung canceralk in lung cancer. *Clinical Cancer Research*, AACR, v. 17, n. 8, p. 2081–2086, 2011. [14](#)
- SHEN, J. *et al.* Novel insights into mir-944 in cancer. *Cancers*, MDPI, v. 14, n. 17, p. 4232, 2022. [149](#)
- SHENDURE, J. *et al.* Dna sequencing at 40: past, present and future. *Nature*, Nature Publishing Group UK London, v. 550, n. 7676, p. 345–353, 2017. [13](#)
- SHERAFATIAN, M.; ARJMAND, F. Decision tree-based classifiers for lung cancer diagnosis and subtyping using tcga mirna expression data. *Oncology letters*, Spandidos Publications, v. 18, n. 2, p. 2125–2131, 2019. [30](#)
- SIEGEL, R.; NAISHADHAM, D.; JEMAL, A. Cancer statistics, 2013. *CA: a cancer journal for clinicians*, v. 63, n. 1, p. 11–30, 2013. [8](#)
- SOUZA, V. G. *et al.* Identifying new contributors to brain metastasis in lung adenocarcinoma: A transcriptomic meta-analysis. *Cancers*, MDPI, v. 15, n. 18, p. 4526, 2023. [27](#)

- SPACKMAN, K. A. Signal detection theory: Valuable tools for evaluating inductive learning. In: ELSEVIER. *Proceedings of the sixth international workshop on Machine learning*. [S.l.], 1989. p. 160–163. 64
- STRIMBU, K.; TAVEL, J. A. What are biomarkers? *Current Opinion in HIV and AIDS*, NIH Public Access, v. 5, n. 6, p. 463, 2010. 20
- SUN, M. *et al.* Integrated analysis identifies microRNA-195 as a suppressor of hippo-yap pathway in colorectal cancer. *Journal of hematology & oncology*, Springer, v. 10, p. 1–16, 2017. 104
- SUN, S.; SCHILLER, J. H.; GAZDAR, A. F. Lung cancer in never smokers—a different disease. *Nature reviews cancer*, Nature Publishing Group UK London, v. 7, n. 10, p. 778–790, 2007. 1, 6
- SUNG, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 71, n. 3, p. 209–249, 2021. 1, 4
- ŚWITLIK, W. *et al.* mir-30a-5p together with mir-210-3p as a promising biomarker for non-small cell lung cancer: A preliminary study. *Cancer Biomarkers*, IOS Press, v. 21, n. 2, p. 479–488, 2018. 104
- SZYMANSKI, M.; ERDMANN, V. A.; BARCISZEWSKI, J. Noncoding regulatory rnas database. *Nucleic acids research*, Oxford University Press, v. 31, n. 1, p. 429–431, 2003. 17
- TAFT, R. J. *et al.* Nuclear-localized tiny rnas are associated with transcription initiation and splice sites in metazoans. *Nature structural & molecular biology*, Nature Publishing Group, v. 17, n. 8, p. 1030–1034, 2010. 17
- TAKAMIZAWA, J. *et al.* Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer research*, AACR, v. 64, n. 11, p. 3753–3756, 2004. 22
- TAN, M.; WU, J.; CAI, Y. Suppression of wnt signaling by the mir-29 family is mediated by demethylation of wif-1 in non-small-cell lung cancer. *Biochemical and biophysical research communications*, Elsevier, v. 438, n. 4, p. 673–679, 2013. 129
- TARAZONA, S. *et al.* Noiseq: a rna-seq differential expression method robust for sequencing depth biases. *EMBnet. journal*, v. 17, n. B, p. 18–19, 2011. 25
- THANDRA, K. C. *et al.* Epidemiology of lung cancer. *Contemporary Oncology/Współczesna Onkologia*, Termedia, v. 25, n. 1, p. 45–52, 2021. 6
- THERNEAU, T. *A package for survival analysis in R (R package version 3.5-0)*. [S.l.]: Springer: New York, NY, USA, 2023. 50
- THERNEAU, T.; ATKINSON, B. *rpart: Recursive Partitioning and Regression Trees*. [S.l.], 2022. R package version 4.1.19. Disponível em: <<https://CRAN.R-project.org/package=rpart>>. 48, 50, 56, 62
- THERNEAU, T. M.; ATKINSON, E. J. *et al.* *An introduction to recursive partitioning using the RPART routines*. [S.l.], 1997. 56
- THUN, M. J. *et al.* Lung cancer death rates in lifelong nonsmokers. *Journal of the National Cancer Institute*, Oxford University Press, v. 98, n. 10, p. 691–699, 2006. 7

- TIAN, S. Classification and survival prediction for early-stage lung adenocarcinoma and squamous cell carcinoma patients. *Oncology letters*, Spandidos Publications, v. 14, n. 5, p. 5464–5470, 2017. 9
- TIMOFEEV, R. Classification and regression trees (cart) theory and applications. *Humboldt University, Berlin*, v. 54, 2004. 48, 49, 50
- TOUSSAINT, G. Bibliography on estimation of misclassification. *IEEE Transactions on information Theory*, IEEE, v. 20, n. 4, p. 472–479, 1974. 43
- TRAVIS, W. D. Lung cancer pathology: current concepts. *Clinics in chest medicine*, Elsevier, v. 41, n. 1, p. 67–85, 2020. 9
- TRAVIS, W. D.; BRAMBILLA, E.; RIELY, G. J. New pathologic classification of lung cancer: relevance for clinical practice and clinical trials. *J Clin Oncol*, v. 31, n. 8, p. 992–1001, 2013. 7
- VLACHOS, I. S. *et al.* Diana-mirpath v3. 0: deciphering microrna function with experimental support. *Nucleic acids research*, Oxford University Press, v. 43, n. W1, p. W460–W466, 2015. 39
- VYKOUKAL, J. *et al.* Contributions of circulating micrornas for early detection of lung cancer. *Cancers*, MDPI, v. 14, n. 17, p. 4221, 2022. 22
- WANG, W. *et al.* Early detection of non-small cell lung cancer by using a 12-microrna panel and a nomogram for assistant diagnosis. *Frontiers in Oncology*, Frontiers Media SA, v. 10, p. 855, 2020. 22
- WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, Nature Publishing Group UK London, v. 10, n. 1, p. 57–63, 2009. 15
- WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, Nature Publishing Group UK London, v. 171, n. 4356, p. 737–738, 1953. 12
- WEISSTEIN, E. W. Bonferroni correction. <https://mathworld.wolfram.com/>, Wolfram Research, Inc., 2004. 38
- WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: <<https://ggplot2.tidyverse.org/>>. 48
- WICKHAM, H.; FRANÇOIS, R.; HENRY, L.; MÜLLER, K. *dplyr: A Grammar of Data Manipulation*. [S.l.], 2022. R package version 1.0.9. Disponível em: <<https://CRAN.R-project.org/package=dplyr>>. 64
- WILCOXON, F. *Individual comparisons by ranking methods*. *Biom. Bull.*, 1, 80–83. 1945. 37
- WILKERSON, M. D. *et al.* Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. *PloS one*, Public Library of Science San Francisco, USA, v. 7, n. 5, p. e36530, 2012. 9
- WILLIAMS, G. J. *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer, 2011. (Use R!). Disponível em: <<https://rd.springer.com/book/10.1007/978-1-4419-9890-3>>. 48
- WITTEN, I. H.; FRANK, E.; MARK, A. *Hall, and Christopher J Pal. Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. 43

- WOOSTER, R. *et al.* Identification of the breast cancer susceptibility gene brca2. *Nature*, Nature Publishing Group UK London, v. 378, n. 6559, p. 789–792, 1995. [21](#)
- WU, K.-L. *et al.* The roles of microrna in lung cancer. *International journal of molecular sciences*, MDPI, v. 20, n. 7, p. 1611, 2019. [20](#)
- WU, W. *et al.* Lncrna dleu2 accelerates the tumorigenesis and invasion of non–small cell lung cancer by sponging mir-30a-5p. *Journal of cellular and molecular medicine*, Wiley Online Library, v. 24, n. 1, p. 441–450, 2020. [105](#)
- XING, P.-Y. *et al.* What are the clinical symptoms and physical signs for non-small cell lung cancer before diagnosis is made? a nation-wide multicenter 10-year retrospective study in china. *Cancer medicine*, Wiley Online Library, v. 8, n. 8, p. 4055–4069, 2019. [8](#)
- YANG, X. *et al.* Temporal trends of the lung cancer mortality attributable to smoking from 1990 to 2017: a global, regional and national analysis. *Lung Cancer*, Elsevier, v. 152, p. 49–57, 2021. [1](#)
- YANG, Y. *et al.* Machine learning application in personalised lung cancer recurrence and survivability prediction. *Computational and Structural Biotechnology Journal*, Elsevier, v. 20, p. 1811–1820, 2022. [30](#)
- YU, I. T. *et al.* Dose-response relationship between cooking fumes exposures and lung cancer among chinese nonsmoking women. *Cancer research*, AACR, v. 66, n. 9, p. 4961–4967, 2006. [7](#)
- ZAMAY, T. N. *et al.* Current and prospective protein biomarkers of lung cancer. *Cancers*, MDPI, v. 9, n. 11, p. 155, 2017. [4](#), [12](#)
- ZAPPA, C.; MOUSA, S. A. Non-small cell lung cancer: current treatment and future advances. *Translational lung cancer research*, AME Publications, v. 5, n. 3, p. 288, 2016. [8](#)
- ZEILEIS, A.; HOTHORN, T.; HORNIK, K. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 17, n. 2, p. 492–514, 2008. [29](#)
- ZHANG, H. *et al.* Plasma mir-145, mir-20a, mir-21 and mir-223 as novel biomarkers for screening early-stage non-small cell lung cancer. *Oncology letters*, Spandidos Publications, v. 13, n. 2, p. 669–676, 2017. [79](#)
- ZHANG, Y.; YANG, Q.; WANG, S. Micrnas: a new key in lung cancer. *Cancer chemotherapy and pharmacology*, Springer, v. 74, p. 1105–1111, 2014. [11](#)
- ZONG, F.-Y. *et al.* The rna-binding protein qki suppresses cancer-associated aberrant splicing. *PLoS genetics*, Public Library of Science San Francisco, USA, v. 10, n. 4, p. e1004289, 2014. [104](#)

6 Material Suplementar

Tabela S1 – Maiores correlações negativas entre os pares - LUAD.

miRNA	Gene	corr	valor p	miRNA	Gene	corr	valor p	miRNA	Gene	corr	valor p
miR-130b-3p	TGFBR2	-0,46	p<0,05	miR-6720-3p	JPT1	-0,35	p<0,05	miR-9-5p	PALMD	-0,32	p<0,05
miR-139-3p	LAGE3	-0,46	p<0,05	miR-708-5p	ASPA	-0,35	p<0,05	miR-6720-3p	ABT1	-0,32	p<0,05
miR-147b-3p	NFIX	-0,43	p<0,05	miR-590-5p	ROBO2	-0,35	p<0,05	miR-196a-5p	AQP4	-0,32	p<0,05
miR-130b-3p	ROBO2	-0,43	p<0,05	miR-182-5p	STARD13	-0,35	p<0,05	miR-141-5p	QKI	-0,32	p<0,05
miR-30a-5p	R3HDM1	-0,43	p<0,05	miR-30a-3p	FAM104A	-0,35	p<0,05	miR-1247-3p	LUC7L	-0,32	p<0,05
miR-30c-2-3p	MCM6	-0,42	p<0,05	miR-139-3p	ANKRD16	-0,35	p<0,05	let-7a-5p	RRM2	-0,32	p<0,05
miR-30c-2-3p	E2F2	-0,42	p<0,05	miR-590-5p	TGFBR2	-0,35	p<0,05	miR-200b-3p	QKI	-0,32	p<0,05
miR-30a-3p	INTS8	-0,42	p<0,05	miR-139-3p	METTTL22	-0,35	p<0,05	miR-184	ABCF2	-0,32	p<0,05
miR-30c-2-3p	MTCH2	-0,41	p<0,05	miR-30a-3p	GNPNAT1	-0,35	p<0,05	miR-30c-2-3p	NAT8L	-0,32	p<0,05
miR-9-5p	FLRT3	-0,41	p<0,05	miR-9-5p	STARD13	-0,35	p<0,05	miR-592	ERBB4	-0,32	p<0,05
miR-378a-3p	UBE2W	-0,41	p<0,05	miR-556-3p	ACSL4	-0,35	p<0,05	let-7c-5p	ATP2A2	-0,32	p<0,05
miR-184	SLC7A5	-0,41	p<0,05	miR-30a-3p	PDK1	-0,35	p<0,05	miR-96-5p	ZEB1	-0,32	p<0,05
miR-195-5p	CDCA4	-0,40	p<0,05	miR-30a-3p	CNPY2	-0,34	p<0,05	miR-139-5p	DENR	-0,32	p<0,05
miR-30a-3p	NAA25	-0,40	p<0,05	miR-30a-3p	UBE2K	-0,34	p<0,05	miR-184	TARBP2	-0,32	p<0,05
miR-1247-3p	B4GALT3	-0,40	p<0,05	miR-1-3p	SRSF9	-0,34	p<0,05	miR-96-5p	ARHGAP6	-0,32	p<0,05
miR-200b-3p	CFL2	-0,40	p<0,05	let-7c-5p	BZW1	-0,34	p<0,05	miR-147b-3p	ADCY9	-0,32	p<0,05
miR-1247-3p	PTGES2	-0,39	p<0,05	miR-139-3p	HGH1	-0,34	p<0,05	miR-93-5p	ETV1	-0,32	p<0,05
miR-139-3p	ALYREF	-0,39	p<0,05	let-7c-5p	E2F5	-0,34	p<0,05	miR-423-5p	KLC2	-0,32	p<0,05
miR-195-5p	BZW1	-0,39	p<0,05	miR-708-5p	THRB	-0,34	p<0,05	miR-139-5p	TPD52	-0,32	p<0,05
miR-195-5p	TLK1	-0,39	p<0,05	let-7f-5p	CDC34	-0,34	p<0,05	miR-96-5p	GRIA1	-0,31	p<0,05
let-7c-5p	ERO1A	-0,39	p<0,05	miR-200a-3p	QKI	-0,34	p<0,05	miR-301a-3p	TGFBR2	-0,31	p<0,05
miR-30a-3p	ARMC1	-0,39	p<0,05	miR-195-5p	CAPRIN1	-0,34	p<0,05	miR-143-3p	CRELD1	-0,31	p<0,05
miR-195-5p	NAA25	-0,39	p<0,05	miR-378a-3p	TMCO1	-0,34	p<0,05	miR-423-5p	CELSR3	-0,31	p<0,05
miR-195-5p	USP14	-0,39	p<0,05	miR-139-3p	CABLES2	-0,34	p<0,05	miR-30a-5p	TWF1	-0,31	p<0,05
let-7c-5p	MARS2	-0,38	p<0,05	miR-130b-3p	ZEB2	-0,34	p<0,05	miR-30c-2-3p	RNF220	-0,31	p<0,05
let-7a-5p	TARBP2	-0,38	p<0,05	miR-1301-3p	PAQR5	-0,34	p<0,05	miR-195-5p	TBL1XR1	-0,31	p<0,05
miR-30a-3p	LMNB1	-0,38	p<0,05	miR-195-5p	UBE2Q1	-0,34	p<0,05	miR-429	ZEB2	-0,31	p<0,05
let-7f-5p	TARBP2	-0,38	p<0,05	miR-429	ETS1	-0,34	p<0,05	miR-221-5p	MEX3B	-0,31	p<0,05
miR-221-5p	CREBZF	-0,38	p<0,05	miR-429	CFL2	-0,34	p<0,05	miR-374a-5p	STYX	-0,31	p<0,05
let-7c-5p	CDC25A	-0,37	p<0,05	miR-6892-5p	CELSR3	-0,34	p<0,05	miR-20a-5p	CYBRD1	-0,31	p<0,05
miR-130b-3p	SECISBP2L	-0,37	p<0,05	miR-556-3p	EPM2A	-0,34	p<0,05	miR-30a-3p	TRIM37	-0,31	p<0,05
miR-17-5p	TGFBR2	-0,37	p<0,05	miR-30a-5p	SYNGR3	-0,34	p<0,05	miR-30a-5p	TFDP1	-0,31	p<0,05
miR-139-3p	PYM1	-0,37	p<0,05	let-7c-5p	SLC20A1	-0,34	p<0,05	miR-423-5p	DBN1	-0,31	p<0,05
miR-130b-5p	GRIA1	-0,37	p<0,05	miR-301b-3p	TGFBR2	-0,34	p<0,05	miR-148b-5p	LRRK2	-0,31	p<0,05
miR-671-5p	DLC1	-0,37	p<0,05	miR-210-3p	CELF2	-0,34	p<0,05	miR-378a-3p	DCAF12	-0,31	p<0,05
miR-30c-2-3p	MRPL19	-0,37	p<0,05	miR-423-5p	CREBZF	-0,34	p<0,05	miR-671-5p	FAT4	-0,31	p<0,05
miR-30a-5p	YWHAZ	-0,37	p<0,05	let-7a-5p	MARS2	-0,34	p<0,05	miR-29b-3p	FERMT2	-0,31	p<0,05
miR-378a-3p	ZDHHC9	-0,37	p<0,05	miR-616-5p	NFIX	-0,33	p<0,05	miR-423-5p	ATXN7L3	-0,31	p<0,05
miR-30a-5p	CALU	-0,37	p<0,05	miR-556-3p	PFIFBP1	-0,33	p<0,05	let-7a-5p	E2F5	-0,31	p<0,05
miR-1-3p	HSPD1	-0,37	p<0,05	miR-3617-3p	GIPC2	-0,33	p<0,05	miR-577	CYBRD1	-0,31	p<0,05
let-7c-5p	CDC34	-0,37	p<0,05	miR-17-5p	ETV1	-0,33	p<0,05	miR-423-5p	DOLPP1	-0,31	p<0,05
miR-141-3p	QKI	-0,37	p<0,05	miR-30a-5p	CELSR3	-0,33	p<0,05	miR-374a-5p	GNPNAT1	-0,31	p<0,05
miR-96-5p	TACC1	-0,37	p<0,05	miR-195-5p	CNE1	-0,33	p<0,05	miR-195-5p	PEX13	-0,31	p<0,05
miR-183-5p	ZEB2	-0,37	p<0,05	miR-6892-5p	SLC12A8	-0,33	p<0,05	miR-6720-3p	STRIP2	-0,31	p<0,05
miR-1247-3p	AP5Z1	-0,37	p<0,05	miR-30c-2-3p	E2F3	-0,33	p<0,05	miR-30a-3p	LIN9	-0,31	p<0,05
miR-141-3p	ZEB2	-0,37	p<0,05	miR-20a-5p	TGFBR2	-0,33	p<0,05	miR-454-3p	TGFBR2	-0,31	p<0,05
miR-96-5p	FHL1	-0,37	p<0,05	miR-30a-3p	NEMPI	-0,33	p<0,05	miR-301b-3p	MTMR10	-0,31	p<0,05
miR-195-5p	KIF23	-0,37	p<0,05	miR-671-5p	BMPR2	-0,33	p<0,05	miR-148b-5p	DPYSL2	-0,30	p<0,05
miR-96-5p	MITF	-0,37	p<0,05	miR-708-5p	GPM6A	-0,33	p<0,05	miR-210-3p	CORO2B	-0,30	p<0,05
miR-143-3p	MOGS	-0,37	p<0,05	miR-30a-3p	KRAS	-0,33	p<0,05	miR-195-5p	CCNT2	-0,30	p<0,05
miR-31-3p	CADM1	-0,37	p<0,05	miR-423-5p	EFNA3	-0,33	p<0,05	miR-184	PSMA4	-0,30	p<0,05
miR-195-5p	CHEK1	-0,37	p<0,05	miR-130b-3p	ARAP2	-0,33	p<0,05	miR-96-5p	CACNB4	-0,30	p<0,05
miR-671-5p	CELF2	-0,37	p<0,05	miR-195-5p	E2F7	-0,33	p<0,05	miR-708-5p	MAP3K3	-0,30	p<0,05
miR-93-5p	TGFBR2	-0,37	p<0,05	miR-147b-3p	MACROD2	-0,33	p<0,05	miR-182-5p	TACC1	-0,30	p<0,05
miR-195-5p	E2F3	-0,37	p<0,05	miR-195-5p	SLC25A22	-0,33	p<0,05	miR-374a-5p	CPC350	-0,30	p<0,05
miR-30c-2-3p	HNRNPL	-0,36	p<0,05	miR-6720-3p	TMEM241	-0,33	p<0,05	miR-378a-3p	RIT1	-0,30	p<0,05
let-7c-5p	HMGA1	-0,36	p<0,05	miR-184	SND1	-0,33	p<0,05	miR-3617-3p	NEGR1	-0,30	p<0,05
miR-139-3p	METTTL2A	-0,36	p<0,05	miR-31-5p	TNS1	-0,33	p<0,05	miR-345-5p	ARHGAP31	-0,30	p<0,05
miR-423-3p	PABPN1	-0,36	p<0,05	miR-9-5p	ID4	-0,33	p<0,05	miR-937-3p	CBX7	-0,30	p<0,05
miR-139-5p	TMPO	-0,36	p<0,05	miR-423-5p	THBS3	-0,33	p<0,05	let-7a-5p	E2F6	-0,30	p<0,05
miR-550a-5p	ATP1A2	-0,36	p<0,05	miR-592	HLF	-0,33	p<0,05	miR-6892-5p	CRELD2	-0,30	p<0,05
miR-30a-5p	UBE2V2	-0,36	p<0,05	miR-30a-3p	PITX2	-0,33	p<0,05	miR-556-3p	SIRPA	-0,30	p<0,05
miR-6720-3p	LSM4	-0,36	p<0,05	miR-6720-3p	HGH1	-0,33	p<0,05	miR-196a-5p	CELF2	-0,30	p<0,05
miR-182-5p	QKI	-0,36	p<0,05	miR-184	CDC25A	-0,33	p<0,05	miR-627-5p	SPTBN1	-0,30	p<0,05
miR-210-3p	NFIX	-0,36	p<0,05	miR-30c-2-3p	PAK1	-0,33	p<0,05	miR-6720-3p	RPL27A	-0,30	p<0,05
miR-130b-3p	ZBTB4	-0,36	p<0,05	miR-184	BUB3	-0,32	p<0,05	miR-1247-3p	BRSK1	-0,30	p<0,05
miR-550a-5p	SCN2B	-0,35	p<0,05	miR-139-3p	RAB1A	-0,32	p<0,05	miR-301b-3p	DPYSL2	-0,30	p<0,05
miR-30a-5p	CCNE2	-0,35	p<0,05	miR-9-5p	EIF4E3	-0,32	p<0,05	miR-616-5p	EMP2	-0,30	p<0,05
miR-423-5p	COPS7B	-0,35	p<0,05	miR-195-5p	SSR1	-0,32	p<0,05	miR-200b-3p	ETS1	-0,30	p<0,05
miR-30a-5p	PHTF2	-0,35	p<0,05	miR-30a-5p	EED	-0,32	p<0,05	miR-20a-5p	ETV1	-0,30	p<0,05
miR-182-5p	MITF	-0,35	p<0,05	miR-374a-5p	TFDP1	-0,32	p<0,05	miR-423-5p	NACC1	-0,30	p<0,05
miR-6720-3p	SYT7	-0,35	p<0,05	miR-137-3p	WIF1	-0,32	p<0,05	miR-200a-3p	ZEB2	-0,30	p<0,05
miR-6720-3p	MTPAP	-0,35	p<0,05	miR-423-5p	SRM	-0,32	p<0,05	miR-30a-5p	LCLAT1	-0,30	p<0,05
miR-130b-3p	DPYSL2	-0,35	p<0,05	miR-139-5p	CCT5	-0,32	p<0,05	miR-148b-5p	ETS1	-0,30	p<0,05
let-7c-5p	TARBP2	-0,35	p<0,05	miR-423-5p	IDH2	-0,32	p<0,05	miR-423-5p	PA2G4	-0,30	p<0,05
miR-9-5p	SHROOM4	-0,35	p<0,05	miR-30c-2-3p	MARCKSL1	-0,32	p<0,05	miR-182-5p	TNS3	-0,30	p<0,05
let-7a-5p	CDC34	-0,35	p<0,05	miR-139-5p	ATP5MC3	-0,32	p<0,05	miR-183-5p	NTN4	-0,30	p<0,05

Tabela S2 – Vias enriquecidas para DE-miRNAs *up* - LUAD.

<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Adherens junction	79	54	15	p<0,01
Autophagy - animal	146	83	15	p<0,01
Ubiquitin mediated proteolysis	142	81	15	p<0,01
Bacterial invasion of epithelial cells	80	54	15	p<0,01
Shigellosis	268	129	15	p<0,01
Cell cycle	129	75	15	p<0,01
Pathways in cancer	555	221	15	p<0,01
Neurotrophin signaling pathway	124	69	15	p<0,01
Protein processing in endoplasmic reticulum	194	96	15	p<0,01
FoxO signaling pathway	139	74	15	p<0,01
Proteoglycans in cancer	220	103	15	p<0,01
Focal adhesion	213	100	15	p<0,01
Salmonella infection	277	122	15	p<0,01
Chronic myeloid leukemia	79	47	15	p<0,01
Colorectal cancer	88	50	15	p<0,01
Mitophagy - animal	76	45	15	p<0,01
Amyotrophic lateral sclerosis	408	163	15	p<0,01
Renal cell carcinoma	70	42	15	p<0,01
Regulation of actin cytoskeleton	224	100	15	p<0,01
Pancreatic cancer	78	45	15	p<0,01
Prostate cancer	101	54	15	p<0,01
HIF-1 signaling pathway	112	58	15	p<0,01
Yersinia infection	147	71	15	p<0,01
Endocytosis	311	128	15	p<0,01
p53 signaling pathway	75	43	15	p<0,01
Huntington disease	339	137	15	p<0,01
Spinocerebellar ataxia	145	69	15	p<0,01
EGFR tyrosine kinase inhibitor resistance	82	45	15	p<0,01
Small cell lung cancer	100	52	15	p<0,01
MAPK signaling pathway	329	132	15	p<0,01
Axon guidance	186	83	15	p<0,01
Tight junction	182	81	15	p<0,01
Longevity regulating pathway	105	53	15	p<0,01
mRNA surveillance pathway	108	54	15	p<0,01
Pathways of neurodegeneration - multiple diseases	539	196	15	p<0,01
AMPK signaling pathway	130	61	15	p<0,01
TGF-beta signaling pathway	103	51	15	p<0,01
mTOR signaling pathway	177	77	15	p<0,01
ErbB signaling pathway	86	44	15	p<0,01
Thermogenesis	252	102	15	p<0,01
Glioma	79	41	15	p<0,01
Circadian rhythm	31	21	14	p<0,01
Oocyte meiosis	134	61	15	p<0,01
PI3K-Akt signaling pathway	372	140	15	p<0,01
Thyroid hormone signaling pathway	137	62	15	p<0,01
Fc gamma R-mediated phagocytosis	101	49	15	p<0,01
Rap1 signaling pathway	214	88	15	p<0,01
Hepatitis C	173	74	15	p<0,01
Alzheimer disease	426	156	15	p<0,01
Non-small cell lung cancer	81	41	15	p<0,01
Hepatitis B	177	75	15	p<0,01
Endocrine resistance	118	54	15	p<0,01
Hepatocellular carcinoma	177	74	15	p<0,01
Cellular senescence	219	88	15	p<0,01
RNA transport	199	81	15	p<0,01
Parkinson disease	282	108	15	p<0,01
Measles	161	68	15	p<0,01
Hippo signaling pathway	164	69	15	p<0,01
Endometrial cancer	61	32	15	p<0,01
Acute myeloid leukemia	69	35	15	p<0,01
Apoptosis	151	64	15	p<0,01
Platinum drug resistance	75	37	14	p<0,01
Growth hormone synthesis, secretion and action	129	56	15	p<0,01
Prolactin signaling pathway	73	36	15	p<0,01

continuação				
<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Pathogenic Escherichia coli infection	222	87	15	p<0,01
Progesterone-mediated oocyte maturation	104	47	15	p<0,01
Various types of N-glycan biosynthesis	40	23	14	p<0,01
Bladder cancer	43	24	15	p<0,01
N-Glycan biosynthesis	51	27	14	p<0,01
Choline metabolism in cancer	106	47	15	p<0,01
Kaposi sarcoma-associated herpesvirus infection	245	93	15	p<0,01
Phosphatidylinositol signaling system	101	45	15	p<0,01
Sphingolipid signaling pathway	133	56	15	p<0,01
Vasopressin-regulated water reabsorption	47	25	15	0,001
TNF signaling pathway	131	55	15	0,001
Central carbon metabolism in cancer	74	35	15	0,001
T cell receptor signaling pathway	115	49	15	0,001
Insulin resistance	124	52	15	0,001
Autophagy - other	33	19	12	0,001
Fluid shear stress and atherosclerosis	149	60	15	0,001
Signaling pathways regulating pluripotency of stem cells	156	62	15	0,001
Insulin signaling pathway	153	61	15	0,001
Transcriptional misregulation in cancer	206	78	15	0,001
Breast cancer	163	64	15	0,001
AGE-RAGE signaling pathway in diabetic complications	115	48	15	0,001
Wnt signaling pathway	173	67	15	0,002
Viral carcinogenesis	265	96	15	0,002
Apelin signaling pathway	140	56	15	0,002
Hedgehog signaling pathway	59	28	15	0,002
Cushing syndrome	176	67	15	0,003
Non-alcoholic fatty liver disease	177	67	15	0,003
Human papillomavirus infection	406	137	15	0,004
Melanoma	76	33	15	0,005
Platelet activation	136	53	14	0,005
Epithelial cell signaling in Helicobacter pylori infection	79	34	15	0,005
Longevity regulating pathway - multiple species	79	34	15	0,005
Apoptosis - multiple species	32	17	13	0,005
Thyroid cancer	43	21	15	0,006
cGMP-PKG signaling pathway	175	65	15	0,006
PD-L1 expression and PD-1 checkpoint pathway in cancer	101	41	15	0,006
RNA degradation	101	41	15	0,006
Relaxin signaling pathway	138	53	15	0,006
Glucagon signaling pathway	114	45	15	0,007
Lysosome	145	55	15	0,007
C-type lectin receptor signaling pathway	116	45	15	0,010
Ras signaling pathway	241	84	15	0,011
Prion disease	331	111	15	0,012
Endocrine and other factor-regulated calcium reabsorption	57	25	15	0,013
Gap junction	99	39	14	0,013
Gastric cancer	162	59	15	0,014
GnRH signaling pathway	97	38	15	0,016
Vibrio cholerae infection	60	25	14	0,027
Lysine degradation	69	28	15	0,027
Oxytocin signaling pathway	161	57	15	0,030
Phospholipase D signaling pathway	158	56	15	0,031
Leukocyte transendothelial migration	120	44	15	0,033
VEGF signaling pathway	61	25	14	0,033
Regulation of lipolysis in adipocytes	61	25	14	0,033
Long-term potentiation	71	28	14	0,040
Necroptosis	174	60	15	0,044
Inositol phosphate metabolism	78	30	15	0,046
Chagas disease	116	42	15	0,046
Terpenoid backbone biosynthesis	22	11	13	0,048

Tabela S3 – Vias enriquecidas para DE-miRNAs *down* - LUAD.

<i>Term Name</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Ubiquitin mediated proteolysis	142	85	11	p<0,01
Cell cycle	129	78	12	p<0,01
Pathways in cancer	555	243	12	p<0,01
Proteoglycans in cancer	220	116	12	p<0,01
Autophagy - animal	146	84	11	p<0,01
Neurotrophin signaling pathway	124	74	12	p<0,01
Hippo signaling pathway	164	91	12	p<0,01
Protein processing in endoplasmic reticulum	194	103	12	p<0,01
Shigellosis	268	131	11	p<0,01
Focal adhesion	213	109	12	p<0,01
FoxO signaling pathway	139	78	11	p<0,01
p53 signaling pathway	75	49	11	p<0,01
Salmonella infection	277	132	11	p<0,01
Renal cell carcinoma	70	46	11	p<0,01
AMPK signaling pathway	130	71	11	p<0,01
Adherens junction	79	49	11	p<0,01
TGF-beta signaling pathway	103	58	12	p<0,01
Prostate cancer	101	57	11	p<0,01
Colorectal cancer	88	51	11	p<0,01
Human papillomavirus infection	406	172	12	p<0,01
Endocytosis	311	137	11	p<0,01
Insulin signaling pathway	153	77	11	p<0,01
Tight junction	182	88	11	p<0,01
mTOR signaling pathway	177	86	11	p<0,01
Regulation of actin cytoskeleton	224	104	11	p<0,01
Spinocerebellar ataxia	145	73	11	p<0,01
MAPK signaling pathway	329	142	11	p<0,01
EGFR tyrosine kinase inhibitor resistance	82	47	11	p<0,01
Phosphatidylinositol signaling system	101	55	11	p<0,01
Bacterial invasion of epithelial cells	80	46	11	p<0,01
Hepatocellular carcinoma	177	85	11	p<0,01
Huntington disease	339	145	11	p<0,01
Oocyte meiosis	134	68	12	p<0,01
Insulin resistance	124	64	11	p<0,01
Thyroid hormone signaling pathway	137	69	12	p<0,01
Small cell lung cancer	100	54	12	p<0,01
Alzheimer disease	426	174	11	p<0,01
HIF-1 signaling pathway	112	58	11	p<0,01
Longevity regulating pathway	105	55	11	p<0,01
AGE-RAGE signaling pathway in diabetic complications	115	59	11	p<0,01
Pathways of neurodegeneration - multiple diseases	539	212	11	p<0,01
Hepatitis B	177	83	11	p<0,01
ErbB signaling pathway	86	47	11	p<0,01
Apelin signaling pathway	140	68	11	p<0,01
Signaling pathways regulating pluripotency of stem cells	156	74	12	p<0,01
Long-term potentiation	71	40	11	p<0,01
mRNA surveillance pathway	108	55	12	p<0,01
Cushing syndrome	176	81	11	p<0,01
Chronic myeloid leukemia	79	43	11	p<0,01
PI3K-Akt signaling pathway	372	151	12	p<0,01
Parkinson disease	282	119	11	p<0,01
Relaxin signaling pathway	138	66	11	p<0,01
Amyotrophic lateral sclerosis	408	163	11	p<0,01
Pancreatic cancer	78	42	11	p<0,01
Rap1 signaling pathway	214	94	11	p<0,01
Growth hormone synthesis, secretion and action	129	62	11	p<0,01
Glucagon signaling pathway	114	56	11	p<0,01
Axon guidance	186	83	11	p<0,01
Fluid shear stress and atherosclerosis	149	69	11	p<0,01
Inositol phosphate metabolism	78	41	11	p<0,01
Hedgehog signaling pathway	59	33	11	p<0,01
Mitophagy - animal	76	40	11	p<0,01
Apoptosis	151	69	12	p<0,01
Apoptosis - multiple species	32	21	11	p<0,01
Acute myeloid leukemia	69	37	12	p<0,01
Transcriptional misregulation in cancer	206	89	12	p<0,01

<i>Term Name</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	continuação	
				FDR	
Glioma	79	41	11	p<0,01	
Yersinia infection	147	67	11	p<0,01	
Viral carcinogenesis	265	109	12	p<0,01	
Wnt signaling pathway	173	76	11	p<0,01	
Circadian rhythm	31	20	11	p<0,01	
Thermogenesis	252	104	12	p<0,01	
Gap junction	99	48	11	p<0,01	
Melanogenesis	107	51	11	p<0,01	
Central carbon metabolism in cancer	74	38	11	p<0,01	
GnRH signaling pathway	97	47	11	p<0,01	
Bladder cancer	43	25	11	0,001	
Hepatitis C	173	75	12	0,001	
Gastric cancer	162	71	11	0,001	
Ferroptosis	41	24	11	0,001	
Fc gamma R-mediated phagocytosis	101	48	11	0,001	
RNA degradation	101	48	12	0,001	
RNA transport	199	84	12	0,001	
Sphingolipid signaling pathway	133	60	11	0,001	
Pathogenic Escherichia coli infection	222	92	11	0,001	
N-Glycan biosynthesis	51	28	10	0,001	
Non-small cell lung cancer	81	40	11	0,001	
Dopaminergic synapse	143	63	11	0,001	
Various types of N-glycan biosynthesis	40	23	9	0,001	
Progesterone-mediated oocyte maturation	104	48	11	0,001	
Prolactin signaling pathway	73	36	11	0,002	
Endocrine resistance	118	53	12	0,002	
Parathyroid hormone synthesis, secretion and action	118	53	11	0,002	
Metabolic pathways	1634	548	11	0,002	
Cellular senescence	219	89	11	0,002	
Endometrial cancer	61	31	11	0,002	
Long-term depression	64	32	11	0,002	
Vasopressin-regulated water reabsorption	47	25	12	0,003	
Platinum drug resistance	75	36	11	0,003	
Breast cancer	163	68	12	0,003	
Chagas disease	116	51	12	0,003	
Thyroid cancer	43	23	11	0,004	
Oxytocin signaling pathway	161	67	11	0,004	
Cysteine and methionine metabolism	53	27	9	0,004	
Epithelial cell signaling in Helicobacter pylori infection	79	37	11	0,004	
Longevity regulating pathway - multiple species	79	37	11	0,004	
Kaposi sarcoma-associated herpesvirus infection	245	96	11	0,004	
Lysine degradation	69	33	11	0,004	
Platelet activation	136	57	11	0,007	
Melanoma	76	35	11	0,007	
TNF signaling pathway	131	55	12	0,007	
Autophagy - other	33	18	10	0,009	
VEGF signaling pathway	61	29	11	0,009	
cGMP-PKG signaling pathway	175	70	11	0,009	
Renin secretion	72	33	10	0,010	
Choline metabolism in cancer	106	45	11	0,013	
Vibrio cholerae infection	60	28	11	0,014	
Measles	161	64	11	0,015	
Ras signaling pathway	241	91	11	0,017	
C-type lectin receptor signaling pathway	116	48	11	0,017	
Gastric acid secretion	80	35	10	0,018	
Carbon metabolism	122	50	11	0,018	
Non-alcoholic fatty liver disease	177	69	11	0,018	
Spliceosome	193	74	11	0,022	
Adrenergic signaling in cardiomyocytes	161	63	11	0,023	
Endocrine and other factor-regulated calcium reabsorption	57	26	11	0,025	
Basal cell carcinoma	66	29	11	0,031	
Phospholipase D signaling pathway	158	61	11	0,034	
Biosynthesis of amino acids	78	33	11	0,037	
Terpenoid backbone biosynthesis	22	12	9	0,037	
Type II diabetes mellitus	57	25	11	0,049	

Tabela S4 – Maiores correlações negativas entre os pares - LUSC.

miRNA	Gene	cor	p	miRNA	Gene	cor	p	miRNA	Gene	cor	p
miR-205-5p	ENPP4	-0,55	p<0,05	miR-944	DPYSL2	-0,42	p<0,05	miR-4449	MYADM	-0,39	p<0,05
miR-205-5p	MGAT4A	-0,55	p<0,05	miR-320a-3p	TWF1	-0,42	p<0,05	miR-205-3p	ST8SIA4	-0,39	p<0,05
miR-141-3p	ZEB2	-0,54	p<0,05	miR-30a-5p	CCNE2	-0,42	p<0,05	miR-205-3p	EBF1	-0,39	p<0,05
miR-205-5p	CFL2	-0,53	p<0,05	miR-29b-3p	LYSMD1	-0,42	p<0,05	miR-497-5p	CDC25A	-0,39	p<0,05
miR-140-3p	OLA1	-0,53	p<0,05	miR-101-5p	GINS1	-0,42	p<0,05	miR-96-5p	TNS3	-0,39	p<0,05
miR-100-5p	BIRC5	-0,51	p<0,05	miR-22-3p	MTNFD2	-0,42	p<0,05	miR-103a-3p	SINHCAF	-0,39	p<0,05
miR-944	FGL2	-0,50	p<0,05	miR-141-3p	APBB2	-0,42	p<0,05	miR-362-3p	CDCA4	-0,39	p<0,05
miR-944	FLI1	-0,50	p<0,05	miR-181a-5p	METAP1	-0,42	p<0,05	miR-708-5p	MAP3K3	-0,39	p<0,05
miR-205-3p	CELF2	-0,50	p<0,05	miR-150-5p	NDC1	-0,42	p<0,05	miR-671-5p	CFL2	-0,39	p<0,05
miR-146b-5p	PGK1	-0,49	p<0,05	miR-205-5p	GOLM2	-0,42	p<0,05	miR-145-5p	ABRACL	-0,39	p<0,05
miR-205-3p	CPED1	-0,49	p<0,05	miR-139-5p	CCT5	-0,42	p<0,05	miR-455-5p	TOMM22	-0,39	p<0,05
miR-205-3p	RAB8B	-0,49	p<0,05	miR-301b-3p	ZEB2	-0,42	p<0,05	miR-30a-5p	TFDP1	-0,39	p<0,05
miR-30c-2-3p	PAK1	-0,49	p<0,05	miR-223-3p	PFN2	-0,42	p<0,05	miR-183-5p	RAB8B	-0,39	p<0,05
miR-205-3p	CYRIA	-0,49	p<0,05	miR-205-5p	CALCRL	-0,42	p<0,05	miR-301b-5p	ARRDC2	-0,39	p<0,05
miR-195-3p	PDCD10	-0,48	p<0,05	miR-342-3p	RRM2	-0,41	p<0,05	miR-30d-3p	LIN9	-0,39	p<0,05
miR-708-5p	RAB8B	-0,48	p<0,05	miR-486-5p	TWF1	-0,41	p<0,05	miR-29b-3p	PXYLP1	-0,39	p<0,05
miR-29b-3p	TDG	-0,48	p<0,05	miR-141-3p	MIER1	-0,41	p<0,05	let-7g-5p	BZW2	-0,39	p<0,05
miR-141-3p	ZCCHC24	-0,48	p<0,05	miR-141-3p	NR3C1	-0,41	p<0,05	miR-199b-5p	UNG	-0,39	p<0,05
miR-150-5p	PFN2	-0,48	p<0,05	miR-149-5p	RAP1A	-0,41	p<0,05	miR-218-5p	YEATS4	-0,39	p<0,05
miR-96-5p	RAB8B	-0,48	p<0,05	let-7b-5p	CDC25A	-0,41	p<0,05	miR-34a-5p	NRIP3	-0,39	p<0,05
miR-30c-5p	TWF1	-0,47	p<0,05	miR-29c-3p	TDG	-0,41	p<0,05	miR-1293	ZNF662	-0,39	p<0,05
miR-497-5p	CDCA4	-0,47	p<0,05	miR-30b-5p	TWF1	-0,41	p<0,05	miR-140-3p	DCUN1D1	-0,39	p<0,05
miR-210-3p	CNRIP1	-0,47	p<0,05	miR-4449	NCF4	-0,41	p<0,05	miR-708-5p	STK4	-0,39	p<0,05
miR-877-5p	HSD11B1	-0,47	p<0,05	miR-130b-3p	CD69	-0,41	p<0,05	miR-181a-3p	PERP	-0,39	p<0,05
miR-205-5p	MFNG	-0,46	p<0,05	miR-26a-5p	TWF1	-0,41	p<0,05	miR-139-5p	SSX2IP	-0,39	p<0,05
miR-205-5p	HSD17B11	-0,46	p<0,05	let-7f-5p	BZW2	-0,41	p<0,05	miR-146b-3p	GOT2	-0,39	p<0,05
miR-361-5p	CKS1B	-0,46	p<0,05	miR-146b-5p	PRMT3	-0,41	p<0,05	miR-96-5p	MITF	-0,39	p<0,05
miR-944	UBL3	-0,46	p<0,05	miR-944	VGLL3	-0,41	p<0,05	miR-944	SLC41A2	-0,39	p<0,05
miR-141-3p	ATP6V1B2	-0,46	p<0,05	miR-451a	SPC25	-0,41	p<0,05	miR-345-5p	LAT2	-0,39	p<0,05
miR-660-5p	SINHCAF	-0,46	p<0,05	miR-193a-5p	OLA1	-0,41	p<0,05	miR-140-5p	SLC2A1	-0,39	p<0,05
miR-141-3p	ZEB1	-0,46	p<0,05	miR-126-3p	E2F7	-0,41	p<0,05	miR-125a-5p	EIF4EBP1	-0,38	p<0,05
miR-141-3p	SDC2	-0,45	p<0,05	miR-210-3p	CD22	-0,41	p<0,05	miR-205-5p	JADE1	-0,38	p<0,05
let-7e-5p	BZW2	-0,45	p<0,05	miR-944	STK17B	-0,41	p<0,05	miR-671-5p	SASH3	-0,38	p<0,05
miR-30d-5p	TWF1	-0,45	p<0,05	miR-141-3p	HGF	-0,41	p<0,05	miR-362-3p	OLA1	-0,38	p<0,05
miR-195-5p	CDCA4	-0,45	p<0,05	miR-141-3p	DOCK4	-0,41	p<0,05	miR-181a-2-3p	RAN	-0,38	p<0,05
miR-210-3p	ISCU	-0,45	p<0,05	miR-15a-5p	CDCA4	-0,41	p<0,05	miR-30d-5p	YWHAZ	-0,38	p<0,05
miR-30d-5p	DBF4	-0,45	p<0,05	miR-708-3p	RGS18	-0,41	p<0,05	miR-29b-3p	SLC16A1	-0,38	p<0,05
let-7e-3p	NUP37	-0,45	p<0,05	miR-30a-5p	LCLAT1	-0,40	p<0,05	miR-1910-5p	IKZF1	-0,38	p<0,05
miR-210-3p	KCNA3	-0,45	p<0,05	miR-195-3p	DCUN1D5	-0,40	p<0,05	miR-181b-5p	LRRRC8D	-0,38	p<0,05
miR-181a-3p	NUP37	-0,45	p<0,05	miR-146b-5p	ZNF367	-0,40	p<0,05	miR-30a-3p	PDK1	-0,38	p<0,05
miR-30a-5p	YWHAZ	-0,45	p<0,05	miR-140-3p	RNF7	-0,40	p<0,05	miR-205-5p	LRRK2	-0,38	p<0,05
miR-140-5p	TTK	-0,45	p<0,05	miR-29a-3p	LYSMD1	-0,40	p<0,05	miR-130b-3p	KBTBD8	-0,38	p<0,05
miR-31-5p	STK40	-0,45	p<0,05	miR-29b-3p	TMT3	-0,40	p<0,05	miR-552-5p	DEPDC1B	-0,38	p<0,05
miR-708-5p	DRAM1	-0,45	p<0,05	miR-26b-5p	KPNA2	-0,40	p<0,05	miR-4777-3p	CKAP2	-0,38	p<0,05
miR-205-5p	LYSMD3	-0,45	p<0,05	miR-6499-3p	PDE4B	-0,40	p<0,05	miR-34a-3p	OLA1	-0,38	p<0,05
miR-30c-5p	YWHAZ	-0,44	p<0,05	miR-34a-5p	CCNE2	-0,40	p<0,05	miR-141-3p	ERG	-0,38	p<0,05
miR-141-3p	QKI	-0,44	p<0,05	miR-221-3p	KPNA2	-0,40	p<0,05	miR-30a-5p	R3HDM1	-0,38	p<0,05
miR-29a-3p	TDG	-0,44	p<0,05	miR-6499-3p	LPIN2	-0,40	p<0,05	miR-223-3p	PAX6	-0,38	p<0,05
miR-30d-3p	CACYBP	-0,44	p<0,05	miR-30d-3p	ORC5	-0,40	p<0,05	miR-3913-5p	LDHA	-0,38	p<0,05
miR-100-5p	CDC25A	-0,44	p<0,05	miR-141-3p	PTPRG	-0,40	p<0,05	miR-181a-5p	SPP1	-0,38	p<0,05
miR-205-3p	ENTPD1	-0,44	p<0,05	miR-34a-5p	LDHA	-0,40	p<0,05	miR-34a-5p	CERS6	-0,38	p<0,05
miR-708-5p	S1PR1	-0,44	p<0,05	miR-30c-5p	CCNE2	-0,40	p<0,05	miR-195-5p	E2F7	-0,38	p<0,05
miR-708-5p	TNS3	-0,44	p<0,05	miR-181a-5p	LRRRC8D	-0,40	p<0,05	miR-130b-3p	ESR1	-0,38	p<0,05
miR-205-5p	TRAK2	-0,44	p<0,05	miR-183-3p	FGL2	-0,40	p<0,05	miR-205-5p	CHN1	-0,38	p<0,05
miR-338-3p	JPT1	-0,44	p<0,05	miR-345-5p	CTRAM	-0,40	p<0,05	miR-500b-5p	TMEM182	-0,38	p<0,05
miR-146b-5p	SIAH2	-0,44	p<0,05	miR-708-5p	PSAP	-0,40	p<0,05	miR-6499-3p	SVIP	-0,38	p<0,05
miR-6499-5p	CELF2	-0,43	p<0,05	miR-30d-5p	PFN2	-0,40	p<0,05	miR-30c-5p	R3HDM1	-0,38	p<0,05
miR-30c-5p	LCLAT1	-0,43	p<0,05	miR-205-3p	CAPZA2	-0,40	p<0,05	miR-150-5p	BORA	-0,38	p<0,05
miR-139-5p	DENR	-0,43	p<0,05	miR-4778-3p	PIK3AP1	-0,40	p<0,05	miR-29b-3p	NKRF	-0,38	p<0,05
miR-181a-5p	DCUN1D1	-0,43	p<0,05	miR-877-5p	LMBRD1	-0,40	p<0,05	miR-96-5p	ADGRA2	-0,38	p<0,05
let-7i-5p	BZW2	-0,43	p<0,05	miR-101-5p	PBK	-0,40	p<0,05	miR-224-5p	ACSL4	-0,38	p<0,05
miR-141-3p	UNC5C	-0,43	p<0,05	miR-141-3p	DLC1	-0,40	p<0,05	miR-29b-3p	DNMT3B	-0,38	p<0,05
miR-145-5p	MCM2	-0,43	p<0,05	let-7i-5p	RRM2	-0,40	p<0,05	miR-195-5p	KIF23	-0,38	p<0,05
miR-944	TFEC	-0,43	p<0,05	miR-944	GATA6	-0,40	p<0,05	miR-206	PDCD10	-0,38	p<0,05
miR-140-5p	RAN	-0,43	p<0,05	miR-671-5p	RNF41	-0,40	p<0,05	miR-26b-5p	TWF1	-0,38	p<0,05
miR-9-5p	LYVE1	-0,43	p<0,05	miR-181b-5p	METAP1	-0,40	p<0,05	miR-552-5p	DNA2	-0,38	p<0,05
miR-210-3p	SH3BGRL	-0,43	p<0,05	miR-218-1-3p	OLA1	-0,40	p<0,05	miR-125a-5p	SUV39H1	-0,38	p<0,05
miR-223-3p	ECT2	-0,42	p<0,05	miR-205-5p	CADM1	-0,40	p<0,05	miR-133a-3p	PFN2	-0,37	p<0,05
miR-205-5p	ZEB2	-0,42	p<0,05	miR-708-5p	TAGAP	-0,39	p<0,05	miR-9-5p	EIF4E3	-0,37	p<0,05
miR-100-5p	GMPS	-0,42	p<0,05	miR-210-3p	LEPROTL1	-0,39	p<0,05	miR-4777-3p	WDHD1	-0,37	p<0,05
miR-301b-3p	ESR1	-0,42	p<0,05	miR-944	KCNT2	-0,39	p<0,05	miR-183-3p	LAPTM5	-0,37	p<0,05
let-7a-5p	RRM2	-0,42	p<0,05	miR-210-3p	CELF2	-0,39	p<0,05	miR-181d-5p	HMB5	-0,37	p<0,05
miR-132-3p	H2AZ1	-0,42	p<0,05	miR-9-5p	FLRT3	-0,39	p<0,05	miR-30a-5p	XPO1	-0,37	p<0,05
miR-103a-3p	TWF1	-0,42	p<0,05	miR-6499-3p	PTPRO	-0,39	p<0,05	miR-301b-5p	TXNIP	-0,37	p<0,05
miR-6499-3p	ADGRG2	-0,42	p<0,05	miR-944	TMEM47	-0,39	p<0,05	miR-1248	SAMD4A	-0,37	p<0,05
miR-96-5p	MSN	-0,42	p<0,05	miR-130b-3p	ZEB2	-0,39	p<0,05	miR-6499-5p	TNFSF13	-0,37	p<0,05
miR-125a-5p	PPAT	-0,42	p<0,05	miR-342-5p	WDR43	-0,39	p<0,05	miR-374b-5p	CCNE2	-0,37	p<0,05

												Continuação
miRNA	Gene	corr	p	miRNA	Gene	corr	p	miRNA	Gene	corr	p	
miR-4449	SORCS2	-0,37	p<0,05	miR-6499-3p	EPB41L5	-0,36	p<0,05	miR-29a-3p	SLC16A1	-0,34	p<0,05	
miR-30a-5p	GCLC	-0,37	p<0,05	miR-206	PFN2	-0,36	p<0,05	miR-30d-3p	PKD1	-0,34	p<0,05	
miR-141-3p	LDLRAD4	-0,37	p<0,05	miR-660-5p	CEP55	-0,36	p<0,05	miR-4449	PREX1	-0,34	p<0,05	
miR-30c-2-3p	MCM6	-0,37	p<0,05	miR-205-5p	SLC4A4	-0,36	p<0,05	miR-149-5p	CD47	-0,34	p<0,05	
miR-146b-5p	SIX4	-0,37	p<0,05	miR-181c-5p	SPP1	-0,36	p<0,05	miR-17-3p	TWF1	-0,34	p<0,05	
miR-185-5p	PARBPB	-0,37	p<0,05	miR-210-3p	PSAP	-0,36	p<0,05	miR-133a-3p	SRPK1	-0,34	p<0,05	
miR-4781-3p	PFN2	-0,37	p<0,05	miR-146b-5p	TIMELESS	-0,36	p<0,05	miR-99b-3p	SPRR2A	-0,34	p<0,05	
miR-30a-5p	LRR8D	-0,37	p<0,05	miR-306b-3p	DBF4	-0,36	p<0,05	miR-183-5p	ZEB2	-0,34	p<0,05	
miR-205-5p	SEPTIN4	-0,37	p<0,05	miR-10a-5p	ZNF367	-0,36	p<0,05	miR-4449	KCNK5	-0,34	p<0,05	
miR-185-5p	SPATS2	-0,37	p<0,05	miR-6510-3p	STX7	-0,35	p<0,05	miR-1976	GINS4	-0,34	p<0,05	
miR-130b-3p	TGFBR2	-0,37	p<0,05	miR-34a-3p	RMI1	-0,35	p<0,05	miR-194-5p	GOT2	-0,34	p<0,05	
miR-181c-5p	METAP1	-0,37	p<0,05	miR-221-3p	PDCD10	-0,35	p<0,05	miR-190a-5p	AK4	-0,34	p<0,05	
miR-181d-5p	SPP1	-0,37	p<0,05	miR-34a-5p	TBL1XR1	-0,35	p<0,05	miR-140-3p	TBL1XR1	-0,34	p<0,05	
miR-141-3p	MPEG1	-0,37	p<0,05	let-7a-5p	CDC25A	-0,35	p<0,05	miR-138-5p	CCNE1	-0,34	p<0,05	
miR-195-5p	CHEK1	-0,37	p<0,05	miR-146b-3p	PTDSS1	-0,35	p<0,05	miR-1-3p	HSPD1	-0,34	p<0,05	
miR-205-5p	TSC22D1	-0,37	p<0,05	miR-130b-5p	ETS1	-0,35	p<0,05	miR-577	CAV1	-0,34	p<0,05	
miR-96-5p	SH3KBP1	-0,37	p<0,05	miR-205-5p	PTPRJ	-0,35	p<0,05	miR-552-5p	BRIP1	-0,34	p<0,05	
miR-141-3p	STAT4	-0,37	p<0,05	miR-183-5p	MEF2C	-0,35	p<0,05	miR-210-3p	RUNX1T1	-0,34	p<0,05	
miR-26b-5p	CKS2	-0,37	p<0,05	miR-29a-3p	DNMT3B	-0,35	p<0,05	miR-708-5p	SYT11	-0,34	p<0,05	
miR-29c-3p	SLC16A1	-0,37	p<0,05	miR-301b-3p	ACSL4	-0,35	p<0,05	miR-574-3p	CSEIL	-0,34	p<0,05	
miR-205-3p	CNOT8	-0,37	p<0,05	miR-497-5p	KIF23	-0,35	p<0,05	miR-143-3p	SINHCAP	-0,34	p<0,05	
miR-6499-3p	ZEB2	-0,37	p<0,05	miR-6499-3p	SCN1A	-0,35	p<0,05	miR-4683	OLA1	-0,34	p<0,05	
miR-223-3p	RCN2	-0,37	p<0,05	let-7i-5p	HMGA1	-0,35	p<0,05	miR-29b-3p	SMS	-0,34	p<0,05	
miR-141-3p	EDEM1	-0,37	p<0,05	miR-500b-5p	HDAC2	-0,35	p<0,05	miR-141-3p	PTEN	-0,34	p<0,05	
miR-205-5p	KAT2B	-0,37	p<0,05	miR-301b-3p	MEOX2	-0,35	p<0,05	miR-130b-5p	CPEB4	-0,34	p<0,05	
miR-671-5p	PJA2	-0,37	p<0,05	miR-1248	NRP1	-0,35	p<0,05	miR-30a-5p	TWF1	-0,34	p<0,05	
miR-1976	PDRG1	-0,37	p<0,05	miR-205-5p	ZFYVE16	-0,35	p<0,05	miR-6499-3p	ROBO2	-0,34	p<0,05	
miR-185-5p	RAE1	-0,37	p<0,05	miR-205-3p	JAM3	-0,35	p<0,05	miR-4772-3p	FBXO45	-0,34	p<0,05	
miR-29a-3p	PXYLP1	-0,37	p<0,05	miR-141-3p	DUSP3	-0,35	p<0,05	miR-210-3p	SORBS2	-0,34	p<0,05	
miR-30b-5p	CCNE2	-0,37	p<0,05	miR-4652-5p	CD247	-0,35	p<0,05	miR-130b-3p	ACSL4	-0,34	p<0,05	
miR-141-3p	FOXP1	-0,37	p<0,05	let-7b-3p	CCNA2	-0,35	p<0,05	miR-708-3p	CAPZA2	-0,34	p<0,05	
miR-577	ETS2	-0,37	p<0,05	miR-210-3p	GPD1L	-0,35	p<0,05	miR-223-5p	WASF1	-0,34	p<0,05	
miR-374a-5p	CCNE2	-0,37	p<0,05	miR-29a-5p	PSMD14	-0,35	p<0,05	miR-374a-5p	MAPK6	-0,34	p<0,05	
miR-205-5p	SLC35A1	-0,36	p<0,05	miR-708-5p	GIT2	-0,35	p<0,05	miR-6499-3p	SECTM1	-0,34	p<0,05	
miR-1910-5p	MEF2C	-0,36	p<0,05	miR-125a-5p	MFN1	-0,35	p<0,05	miR-6499-3p	KCNK3	-0,34	p<0,05	
miR-767-5p	SH3GLB1	-0,36	p<0,05	miR-5586-5p	DSCC1	-0,35	p<0,05	miR-205-5p	PJA2	-0,34	p<0,05	
miR-1293	S1PR1	-0,36	p<0,05	miR-1248	ETS1	-0,35	p<0,05	miR-516a-5p	STIL	-0,34	p<0,05	
miR-93-3p	RAP1A	-0,36	p<0,05	miR-361-3p	PTDSS1	-0,35	p<0,05	miR-877-5p	MAN1A1	-0,34	p<0,05	
miR-149-5p	WIPF1	-0,36	p<0,05	miR-3913-5p	YWHAZ	-0,35	p<0,05	miR-205-3p	MBNL2	-0,34	p<0,05	
miR-200b-5p	ZEB2	-0,36	p<0,05	miR-1293	STARD8	-0,35	p<0,05	miR-132-3p	TTK	-0,34	p<0,05	
miR-1293	CEL2F	-0,36	p<0,05	miR-1293	DPYSL2	-0,35	p<0,05	miR-125b-5p	SUV39H1	-0,34	p<0,05	
let-7i-5p	MAPK6	-0,36	p<0,05	let-7e-5p	CDC25A	-0,35	p<0,05	miR-130b-3p	IRF1	-0,34	p<0,05	
miR-4778-3p	MEF2C	-0,36	p<0,05	miR-206	TWF1	-0,35	p<0,05	miR-6499-5p	ZSWIM7	-0,34	p<0,05	
miR-138-5p	MAPK6	-0,36	p<0,05	miR-181a-5p	NKAIN2	-0,35	p<0,05	miR-301b-3p	KBTBD8	-0,34	p<0,05	
miR-4652-5p	DNAJC15	-0,36	p<0,05	miR-4449	PTPRE	-0,35	p<0,05	miR-516a-5p	OLA1	-0,34	p<0,05	
miR-497-5p	CHEK1	-0,36	p<0,05	miR-101-5p	TRIM59	-0,35	p<0,05	miR-30d-5p	LCLAT1	-0,34	p<0,05	
miR-139-5p	GOT2	-0,36	p<0,05	miR-141-3p	PCDH9	-0,35	p<0,05	miR-1910-5p	CKNIP1	-0,34	p<0,05	
miR-103a-3p	PDCD10	-0,36	p<0,05	miR-29a-3p	TMT3C	-0,35	p<0,05	miR-378c	SINHCAP	-0,34	p<0,05	
miR-4777-3p	CENPK	-0,36	p<0,05	miR-200b-5p	KCTD12	-0,35	p<0,05	miR-26a-5p	MAPK6	-0,34	p<0,05	
miR-181d-5p	METAP1	-0,36	p<0,05	miR-205-5p	FOXF1	-0,35	p<0,05	miR-374b-5p	LIN9	-0,33	p<0,05	
miR-141-3p	STX2	-0,36	p<0,05	miR-107	TWF1	-0,35	p<0,05	miR-141-3p	DCUN1D3	-0,33	p<0,05	
miR-6499-3p	SGTB	-0,36	p<0,05	miR-374b-5p	MAPK6	-0,35	p<0,05	miR-342-5p	PPME1	-0,33	p<0,05	
miR-149-5p	PXK	-0,36	p<0,05	miR-877-5p	AOX1	-0,35	p<0,05	miR-217-5p	LIN9	-0,33	p<0,05	
miR-338-3p	TBL1XR1	-0,36	p<0,05	miR-205-5p	ZEB1	-0,35	p<0,05	miR-150-5p	PERP	-0,33	p<0,05	
miR-205-3p	TSPAN2	-0,36	p<0,05	miR-500b-5p	TMEM237	-0,35	p<0,05	miR-138-5p	CLNS1A	-0,33	p<0,05	
miR-9-5p	KCTD12	-0,36	p<0,05	miR-320a-3p	AK4	-0,35	p<0,05	miR-4709-3p	NLN	-0,33	p<0,05	
miR-96-5p	VAMP3	-0,36	p<0,05	miR-30c-5p	TFDP1	-0,35	p<0,05	miR-141-3p	ANKFY1	-0,33	p<0,05	
miR-455-5p	INTS13	-0,36	p<0,05	miR-301b-3p	PTEN	-0,35	p<0,05	miR-3926	CKAP2	-0,33	p<0,05	
miR-205-5p	QKI	-0,36	p<0,05	let-7b-5p	IGF2BP3	-0,35	p<0,05	let-7f-5p	RRM2	-0,33	p<0,05	
miR-196b-5p	CPM	-0,36	p<0,05	miR-301b-3p	RTN1	-0,35	p<0,05	miR-183-5p	FCHO2	-0,33	p<0,05	
miR-1976	AP1S1	-0,36	p<0,05	miR-301b-5p	CD59	-0,34	p<0,05	miR-320a-3p	YWHAZ	-0,33	p<0,05	
miR-6499-3p	ZNF763	-0,36	p<0,05	miR-210-3p	PALM2AKAP2	-0,34	p<0,05	miR-362-3p	AK4	-0,33	p<0,05	
miR-205-5p	LHFPL2	-0,36	p<0,05	miR-708-5p	FMNL3	-0,34	p<0,05	miR-150-5p	METAP1	-0,33	p<0,05	
miR-96-5p	NAV3	-0,36	p<0,05	miR-196a-5p	CALM1	-0,34	p<0,05	miR-101-5p	ATAD2	-0,33	p<0,05	
miR-342-5p	GINS3	-0,36	p<0,05	miR-196b-5p	AQP4	-0,34	p<0,05	miR-26a-5p	E2F7	-0,33	p<0,05	
miR-5579-3p	FAM107B	-0,36	p<0,05	miR-301b-3p	IRF1	-0,34	p<0,05	miR-199b-5p	MCM2	-0,33	p<0,05	
miR-140-5p	PFN2	-0,36	p<0,05	miR-29a-3p	NKRF	-0,34	p<0,05	miR-3926	NLM	-0,33	p<0,05	
miR-126-5p	E2F7	-0,36	p<0,05	miR-210-3p	RCAN2	-0,34	p<0,05	miR-9-5p	NR5A2	-0,33	p<0,05	
miR-301b-3p	TGFBR2	-0,36	p<0,05	miR-342-3p	SLC16A1	-0,34	p<0,05	miR-141-3p	CREBRF	-0,33	p<0,05	
miR-26a-5p	HMGA1	-0,36	p<0,05	miR-205-3p	TMEM170B	-0,34	p<0,05	miR-26b-5p	RCN2	-0,33	p<0,05	
miR-500b-5p	TBCCD1	-0,36	p<0,05	miR-9-5p	VAMP3	-0,34	p<0,05	miR-96-5p	CAMK2N1	-0,33	p<0,05	
miR-125b-5p	TDG	-0,36	p<0,05	miR-944	CPEB4	-0,34	p<0,05	miR-205-3p	TMEM245	-0,33	p<0,05	
miR-944	KAT2B	-0,36	p<0,05	miR-99b-5p	CDC25A	-0,34	p<0,05	miR-30c-5p	XPO1	-0,33	p<0,05	
miR-361-3p	PSMD2	-0,36	p<0,05	miR-29a-5p	APOO	-0,34	p<0,05	miR-516a-5p	XXYLT1	-0,33	p<0,05	
miR-206	YWHAZ	-0,36	p<0,05	miR-877-5p	EPS15	-0,34	p<0,05	miR-146b-3p	SLC9A3R1	-0,33	p<0,05	
miR-30c-5p	LRR8D	-0,36	p<0,05	miR-126-3p	TP63	-0,34	p<0,05	miR-133b	PFN2	-0,33	p<0,05	
miR-205-3p	PCGF5	-0,36	p<0,05	miR-26a-5p	RCN2	-0,34	p<0,05	miR-30d-5p	CCNE2	-0,33	p<0,05	

											Continuação
miRNA	Gene	corr	p	miRNA	Gene	corr	p	miRNA	Gene	corr	p
miR-31-5p	AKAP7	-0,33	p<0,05	miR-3662	PRRG1	-0,32	p<0,05	miR-33a-5p	S1PR1	-0,31	p<0,05
miR-944	SGPP1	-0,33	p<0,05	miR-301b-3p	QKI	-0,32	p<0,05	miR-1247-3p	MELTF	-0,31	p<0,05
miR-107	PDCD10	-0,33	p<0,05	let-7g-5p	MAPK6	-0,32	p<0,05	miR-7-5p	PRKCB	-0,31	p<0,05
miR-130b-3p	EPS15	-0,33	p<0,05	miR-96-5p	BNC2	-0,32	p<0,05	let-7e-5p	MAPK6	-0,31	p<0,05
miR-34a-5p	WASF1	-0,33	p<0,05	miR-9-5p	SGMS2	-0,32	p<0,05	miR-183-5p	QKI	-0,31	p<0,05
miR-30c-5p	PFN2	-0,33	p<0,05	miR-598-3p	FKBP4	-0,32	p<0,05	miR-139-5p	TPD52	-0,31	p<0,05
miR-6510-3p	SYNE3	-0,33	p<0,05	miR-573	PRRG1	-0,32	p<0,05	miR-30d-5p	R3HDM1	-0,31	p<0,05
miR-96-5p	PTGER3	-0,33	p<0,05	miR-375-3p	SHOX2	-0,32	p<0,05	miR-29a-5p	TWF1	-0,31	p<0,05
miR-573	ST6GALNAC3	-0,33	p<0,05	miR-708-3p	ATG7	-0,32	p<0,05	miR-224-5p	SH3KBP1	-0,31	p<0,05
miR-320a-3p	RCN2	-0,33	p<0,05	miR-9-5p	ITM2B	-0,32	p<0,05	miR-205-3p	RSBN1	-0,31	p<0,05
miR-708-5p	VAMP2	-0,33	p<0,05	miR-96-5p	GPM6A	-0,32	p<0,05	miR-30d-5p	RAB38	-0,31	p<0,05
miR-96-5p	PROK2	-0,33	p<0,05	miR-93-3p	ITGA4	-0,32	p<0,05	miR-210-3p	CYGB	-0,31	p<0,05
miR-1248	JCAD	-0,33	p<0,05	let-7b-3p	LIN9	-0,32	p<0,05	miR-221-3p	STMN1	-0,31	p<0,05
miR-29b-3p	KDM5B	-0,33	p<0,05	miR-1976	PRR11	-0,32	p<0,05	miR-205-5p	PTPRM	-0,31	p<0,05
miR-210-3p	SACM1L	-0,33	p<0,05	miR-151b	TOMM40	-0,32	p<0,05	miR-141-3p	PDS5B	-0,31	p<0,05
miR-96-5p	EIF4EBP2	-0,33	p<0,05	miR-1248	MS4A2	-0,32	p<0,05	miR-6499-3p	DIXDC1	-0,31	p<0,05
miR-101-5p	HDAC2	-0,33	p<0,05	miR-205-3p	LDLRAD4	-0,32	p<0,05	miR-1293	MEIS1	-0,31	p<0,05
miR-26a-5p	SSX2IP	-0,33	p<0,05	miR-1293	FMNL3	-0,32	p<0,05	miR-877-5p	ITM2B	-0,31	p<0,05
miR-195-3p	RCN2	-0,33	p<0,05	miR-181a-5p	HOXA1	-0,32	p<0,05	miR-133a-3p	XPO1	-0,31	p<0,05
miR-96-5p	ZEB1	-0,33	p<0,05	miR-1293	GAS7	-0,32	p<0,05	miR-30b-5p	R3HDM1	-0,31	p<0,05
miR-4709-3p	ALG10	-0,33	p<0,05	miR-210-3p	KLF9	-0,32	p<0,05	miR-1-3p	PDCD10	-0,31	p<0,05
miR-210-3p	FCHSD2	-0,33	p<0,05	miR-4778-3p	PRKCB	-0,32	p<0,05	miR-26b-5p	MAPK6	-0,31	p<0,05
miR-29c-3p	DNMT3B	-0,33	p<0,05	miR-219a-2-3p	TMEM182	-0,32	p<0,05	miR-185-5p	YWHAZ	-0,31	p<0,05
miR-149-5p	GAB2	-0,33	p<0,05	miR-573	MBNL2	-0,32	p<0,05	miR-183-3p	RHOA	-0,31	p<0,05
miR-33a-5p	ZEB2	-0,33	p<0,05	miR-107	SINHCAF	-0,32	p<0,05	miR-181c-5p	HOXA1	-0,31	p<0,05
miR-138-5p	EIF4EBP1	-0,33	p<0,05	miR-146b-5p	FAM169A	-0,32	p<0,05	miR-26a-5p	FAM136A	-0,31	p<0,05
miR-200b-5p	NEK7	-0,33	p<0,05	miR-6499-5p	TMEM130	-0,32	p<0,05	miR-141-3p	CLASP2	-0,31	p<0,05
miR-205-3p	PARVA	-0,33	p<0,05	miR-29c-5p	IGF2BP1	-0,32	p<0,05	miR-15a-5p	TBL1XR1	-0,31	p<0,05
miR-511-5p	SINHCAF	-0,33	p<0,05	miR-708-5p	SESNI	-0,32	p<0,05	let-7e-5p	AP1S1	-0,31	p<0,05
miR-29b-3p	NUDT11	-0,33	p<0,05	miR-9-5p	UBASH3B	-0,32	p<0,05	miR-141-3p	CD47	-0,31	p<0,05
miR-125b-5p	PPAT	-0,33	p<0,05	miR-664a-3p	MAPK6	-0,32	p<0,05	miR-3662	CPEB4	-0,31	p<0,05
miR-944	BMPR2	-0,33	p<0,05	miR-34a-3p	PFN2	-0,32	p<0,05	miR-26a-5p	PLOD2	-0,31	p<0,05
miR-378c	PDK1	-0,33	p<0,05	miR-210-3p	KLF6	-0,32	p<0,05	let-7c-5p	HMGA1	-0,31	p<0,05
miR-96-5p	PLPPR4	-0,33	p<0,05	miR-217-5p	TDG	-0,32	p<0,05	miR-22-3p	YWHAZ	-0,31	p<0,05
miR-30b-3p	HDAC2	-0,33	p<0,05	miR-139-3p	ALYREF	-0,32	p<0,05	miR-221-3p	DCUN1D1	-0,31	p<0,05
miR-181c-5p	LRRC8D	-0,33	p<0,05	miR-338-5p	CBX3	-0,32	p<0,05	miR-224-5p	DPYSL2	-0,31	p<0,05
miR-145-5p	SELENOI	-0,33	p<0,05	miR-205-5p	PDE3B	-0,32	p<0,05	miR-877-5p	YPEL2	-0,31	p<0,05
miR-574-3p	JPT1	-0,33	p<0,05	miR-4772-3p	FRMD6	-0,32	p<0,05	miR-146b-5p	SESNI	-0,31	p<0,05
miR-1248	BCL6B	-0,33	p<0,05	miR-4709-3p	SNAI2	-0,32	p<0,05	miR-29c-5p	ATAD2	-0,31	p<0,05
miR-708-5p	QKI	-0,33	p<0,05	miR-183-3p	MAN1A1	-0,32	p<0,05	miR-183-5p	TRAM1	-0,31	p<0,05
miR-497-5p	TBL1XR1	-0,33	p<0,05	miR-96-5p	FN1	-0,32	p<0,05	miR-195-5p	TBL1XR1	-0,31	p<0,05
miR-1248	NR2F2	-0,33	p<0,05	miR-126-3p	SLC7A5	-0,32	p<0,05	miR-944	RAPGEF2	-0,31	p<0,05
miR-96-5p	LRCH2	-0,33	p<0,05	miR-205-5p	CREBRF	-0,32	p<0,05	miR-205-5p	SECISBP2L	-0,31	p<0,05
miR-30a-3p	ORC5	-0,33	p<0,05	miR-1976	HMGA2	-0,32	p<0,05	miR-708-5p	GPM6A	-0,31	p<0,05
miR-30b-3p	SKA1	-0,33	p<0,05	miR-96-5p	SDC2	-0,32	p<0,05	miR-4652-5p	GNG7	-0,31	p<0,05
miR-181a-5p	SSX2IP	-0,33	p<0,05	miR-301b-3p	CSF1	-0,32	p<0,05	miR-519a-3p	ACTL6A	-0,31	p<0,05
miR-519a-3p	E2F7	-0,33	p<0,05	miR-3912-3p	DENR	-0,32	p<0,05	miR-146b-5p	JAG1	-0,31	p<0,05
miR-511-3p	DENR	-0,33	p<0,05	miR-31-5p	TACC1	-0,31	p<0,05	miR-125a-5p	MLF2	-0,31	p<0,05
let-7b-5p	HMGA2	-0,33	p<0,05	miR-342-5p	TBL1XR1	-0,31	p<0,05	miR-497-5p	HMGA2	-0,31	p<0,05
miR-181a-2-3p	PDCD10	-0,33	p<0,05	miR-146b-5p	ZDHHC13	-0,31	p<0,05	miR-1293	ARL6IP5	-0,31	p<0,05
miR-140-5p	FBXO45	-0,32	p<0,05	miR-490-3p	HMGA2	-0,31	p<0,05	miR-205-3p	PCDH17	-0,31	p<0,05
miR-4778-3p	ENTPD1	-0,32	p<0,05	miR-1293	KIAA1328	-0,31	p<0,05	let-7d-3p	TIMMDC1	-0,31	p<0,05
miR-125a-5p	BDH1	-0,32	p<0,05	miR-100-5p	E2F2	-0,31	p<0,05	miR-877-5p	CAMLG	-0,31	p<0,05
miR-9-5p	CPEB2	-0,32	p<0,05	miR-93-3p	PAG1	-0,31	p<0,05	miR-144-5p	KCTD1	-0,31	p<0,05
miR-1248	DLC1	-0,32	p<0,05	miR-4449	MMP25	-0,31	p<0,05	miR-501-5p	FRMD6	-0,31	p<0,05
miR-146b-5p	NEMP1	-0,32	p<0,05	miR-3065-3p	COL11A1	-0,31	p<0,05	miR-342-3p	AP2M1	-0,31	p<0,05
miR-6499-5p	SERPINB9	-0,32	p<0,05	miR-183-3p	MBNL2	-0,31	p<0,05	miR-519a-3p	TWF1	-0,31	p<0,05
miR-708-5p	NR4A3	-0,32	p<0,05	miR-5579-3p	OLR1	-0,31	p<0,05	miR-146b-5p	IRAK1	-0,31	p<0,05
miR-6499-3p	SLC1A1	-0,32	p<0,05	miR-100-5p	OPA1	-0,31	p<0,05	miR-181d-5p	LRRC8D	-0,31	p<0,05
miR-301a-5p	ITM2A	-0,32	p<0,05	miR-140-3p	HS6ST2	-0,31	p<0,05	miR-29a-3p	NUDT11	-0,31	p<0,05
miR-375-3p	CLCA2	-0,32	p<0,05	miR-205-5p	SORBS1	-0,31	p<0,05	miR-552-5p	DNAJC9	-0,31	p<0,05
miR-4772-5p	MSI2	-0,32	p<0,05	miR-34a-3p	WDR72	-0,31	p<0,05	miR-6499-5p	LRRC27	-0,31	p<0,05
miR-210-3p	CHN1	-0,32	p<0,05	miR-96-5p	CELF2	-0,31	p<0,05	miR-1258	AK4	-0,31	p<0,05
miR-1287-5p	CDCA4	-0,32	p<0,05	miR-126-3p	CCNE2	-0,31	p<0,05	miR-205-5p	WWC2	-0,31	p<0,05
miR-33a-5p	TENT5C	-0,32	p<0,05	miR-26b-5p	E2F7	-0,31	p<0,05	miR-183-5p	MTMR6	-0,31	p<0,05
miR-210-3p	FOXP1	-0,32	p<0,05	miR-99a-5p	SELENOI	-0,31	p<0,05	miR-205-5p	SGMS1	-0,31	p<0,05
miR-4683	CACYBP	-0,32	p<0,05	miR-6499-5p	SLC4A8	-0,31	p<0,05	miR-183-3p	MEF2C	-0,31	p<0,05
miR-1910-5p	GLCCI1	-0,32	p<0,05	miR-1293	PDZRN3	-0,31	p<0,05	miR-671-5p	VGLL3	-0,31	p<0,05
miR-205-3p	HDX	-0,32	p<0,05	miR-130b-3p	MEOX2	-0,31	p<0,05	miR-616-5p	EMP2	-0,31	p<0,05
miR-196b-5p	CELF2	-0,32	p<0,05	miR-126-5p	LIN9	-0,31	p<0,05	miR-183-5p	PTEN	-0,31	p<0,05
miR-223-5p	CHAC2	-0,32	p<0,05	miR-30c-5p	GCLC	-0,31	p<0,05	miR-183-3p	CSGALNACT2	-0,31	p<0,05
miR-301b-3p	PSAP	-0,32	p<0,05	miR-6892-5p	AK4	-0,31	p<0,05	miR-887-5p	ANAPC7	-0,31	p<0,05
miR-374a-5p	LIN9	-0,32	p<0,05	miR-6510-3p	NDFIP1	-0,31	p<0,05	miR-6499-5p	TBXA2R	-0,31	p<0,05
miR-30c-2-3p	OPA1	-0,32	p<0,05	miR-29c-3p	TMTC3	-0,31	p<0,05	miR-6499-3p	MTMR7	-0,31	p<0,05
miR-30a-3p	LIN9	-0,32	p<0,05	miR-4449	LEPROTL1	-0,31	p<0,05	miR-3662	RAP1A	-0,31	p<0,05
miR-29a-3p	MLLT11	-0,32	p<0,05	miR-4652-5p	IKZF1	-0,31	p<0,05	miR-30e-3p	LIN9	-0,31	p<0,05
miR-301b-3p	MAT2B	-0,32	p<0,05	miR-708-5p	ASPA	-0,31	p<0,05	miR-1258	NCAPG	-0,31	p<0,05

										Continuação
miRNA	Gene	corr	p	miRNA	Gene	corr	p	miRNA	Gene	corr p
miR-301a-5p	ZEB1	-0,31	p<0,05	miR-141-3p	ATXN1	-0,30	p<0,05	miR-345-5p	IRF1	-0,30 p<0,05
miR-145-3p	R3HDM1	-0,31	p<0,05	miR-301b-5p	UBTD1	-0,30	p<0,05	miR-130b-5p	SP100	-0,30 p<0,05
miR-205-3p	CSGALNACT2	-0,31	p<0,05	miR-141-3p	ATP8A1	-0,30	p<0,05	miR-6510-3p	SMIM10	-0,30 p<0,05
miR-30d-5p	TFDP1	-0,31	p<0,05	miR-30b-5p	LCLAT1	-0,30	p<0,05	miR-4449	KIAA0040	-0,30 p<0,05
miR-141-3p	SIRT1	-0,31	p<0,05	miR-20a-3p	CPEB4	-0,30	p<0,05	miR-20a-3p	SIAE	-0,30 p<0,05
miR-944	ZFYVE16	-0,31	p<0,05	miR-205-3p	RRM2B	-0,30	p<0,05	miR-573	FGL2	-0,30 p<0,05
miR-144-5p	CTSV	-0,31	p<0,05	miR-1293	FLT4	-0,30	p<0,05	miR-1287-5p	DSG1	-0,30 p<0,05
miR-4449	VENTX	-0,31	p<0,05	miR-140-5p	WASF1	-0,30	p<0,05	miR-205-5p	RTN3	-0,30 p<0,05
miR-345-5p	PDE4B	-0,31	p<0,05	miR-6718-5p	RRM2	-0,30	p<0,05	miR-5586-5p	SSX2IP	-0,30 p<0,05
miR-29c-3p	NKRF	-0,31	p<0,05	miR-105-5p	FGL2	-0,30	p<0,05	miR-598-3p	DCAF13	-0,30 p<0,05
miR-573	FGF7	-0,31	p<0,05	miR-877-5p	RFTN2	-0,30	p<0,05	miR-141-3p	MBNL1	-0,30 p<0,05
miR-139-3p	NRIP3	-0,31	p<0,05	miR-374b-5p	ATAD2	-0,30	p<0,05	miR-183-3p	PPP3CB	-0,30 p<0,05
miR-301a-5p	ARRDC2	-0,31	p<0,05	miR-210-3p	IQSEC1	-0,30	p<0,05	miR-497-5p	CCNE1	-0,30 p<0,05
miR-210-3p	NR3C2	-0,31	p<0,05	miR-4772-3p	PRR11	-0,30	p<0,05	miR-29c-5p	ONECUT2	-0,30 p<0,05
miR-194-5p	RAB38	-0,31	p<0,05	miR-194-5p	PFN2	-0,30	p<0,05	miR-944	SECISBP2L	-0,30 p<0,05
miR-877-5p	CREBRF	-0,31	p<0,05	miR-144-3p	CEP55	-0,30	p<0,05	miR-100-5p	CEP85	-0,30 p<0,05
miR-30a-3p	INTS8	-0,30	p<0,05	miR-23a-3p	MARCKSL1	-0,30	p<0,05	miR-1293	ACACB	-0,30 p<0,05
miR-1293	KCNAB2	-0,30	p<0,05	miR-30d-5p	LRRC8D	-0,30	p<0,05	let-7g-5p	CDC25A	-0,30 p<0,05
miR-4777-3p	SRSF12	-0,30	p<0,05	miR-200b-5p	VGLL3	-0,30	p<0,05	miR-130b-3p	PTEN	-0,30 p<0,05
miR-6499-3p	SMAD9	-0,30	p<0,05	miR-181a-5p	CPOX	-0,30	p<0,05	miR-29a-3p	SMS	-0,30 p<0,05
miR-4778-3p	HBEGF	-0,30	p<0,05	miR-205-5p	MICU3	-0,30	p<0,05	miR-1976	MAPK6	-0,30 p<0,05
miR-301b-3p	EPS15	-0,30	p<0,05	miR-150-5p	PDIA6	-0,30	p<0,05	miR-1248	MEF2C	-0,30 p<0,05
miR-224-5p	MBNL2	-0,30	p<0,05	miR-219a-2-3p	SINHCAF	-0,30	p<0,05	miR-301b-5p	BTN3A2	-0,30 p<0,05
miR-301b-3p	DPYSL2	-0,30	p<0,05	miR-150-5p	PLP2	-0,30	p<0,05			
miR-96-5p	CNN3	-0,30	p<0,05	miR-15a-5p	CHEK1	-0,30	p<0,05			

Tabela S5 – Vias enriquecidas para DE-miRNAs *up* - LUSC.

<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Bacterial invasion of epithelial cells	80	47	12	p<0,01
Prostate cancer	101	53	13	p<0,01
Adherens junction	79	45	12	p<0,01
FoxO signaling pathway	139	66	13	p<0,01
Renal cell carcinoma	70	41	12	p<0,01
Focal adhesion	213	87	13	p<0,01
Shigellosis	268	102	11	p<0,01
Cell cycle	129	59	13	p<0,01
Chronic myeloid leukemia	79	42	13	p<0,01
Regulation of actin cytoskeleton	224	87	13	p<0,01
Ubiquitin mediated proteolysis	142	62	11	p<0,01
Autophagy - animal	146	63	13	p<0,01
Pathways in cancer	555	176	13	p<0,01
Proteoglycans in cancer	220	85	13	p<0,01
p53 signaling pathway	75	39	12	p<0,01
Bladder cancer	43	27	13	p<0,01
Neurotrophin signaling pathway	124	55	12	p<0,01
Cellular senescence	219	83	13	p<0,01
Colorectal cancer	88	42	13	p<0,01
PI3K-Akt signaling pathway	372	123	13	p<0,01
Protein processing in endoplasmic reticulum	194	74	12	p<0,01
Hippo signaling pathway	164	65	12	p<0,01
EGFR tyrosine kinase inhibitor resistance	82	39	12	p<0,01
Endocytosis	311	105	13	p<0,01
HIF-1 signaling pathway	112	48	13	p<0,01
Endometrial cancer	61	31	13	p<0,01
Rap1 signaling pathway	214	77	13	p<0,01
Central carbon metabolism in cancer	74	35	12	p<0,01
Pancreatic cancer	78	36	12	p<0,01
AMPK signaling pathway	130	52	12	p<0,01
Salmonella infection	277	93	13	p<0,01
Axon guidance	186	68	12	p<0,01
Endocrine resistance	118	48	13	p<0,01
Hepatitis B	177	65	13	p<0,01
Tight junction	182	66	12	p<0,01
Small cell lung cancer	100	42	13	p<0,01
Non-small cell lung cancer	81	36	12	p<0,01
MAPK signaling pathway	329	105	13	p<0,01
Glioma	79	35	13	p<0,01
Melanoma	76	34	13	p<0,01
Hepatocellular carcinoma	177	63	13	p<0,01
RNA transport	199	69	12	p<0,01
Pathogenic Escherichia coli infection	222	75	12	p<0,01
Amyotrophic lateral sclerosis	408	123	11	p<0,01
Longevity regulating pathway	105	42	12	p<0,01
Viral carcinogenesis	265	86	13	p<0,01
Yersinia infection	147	54	12	p<0,01
Human papillomavirus infection	406	122	13	p<0,01
mTOR signaling pathway	177	62	12	p<0,01
Spinocerebellar ataxia	145	53	11	p<0,01
Relaxin signaling pathway	138	51	13	p<0,01
Insulin signaling pathway	153	55	13	p<0,01
Fc gamma R-mediated phagocytosis	101	40	11	p<0,01
Thyroid hormone signaling pathway	137	50	13	p<0,01
Oocyte meiosis	134	49	13	p<0,01
ErbB signaling pathway	86	35	12	p<0,01
Prolactin signaling pathway	73	31	13	p<0,01
N-Glycan biosynthesis	51	24	10	p<0,01
Fluid shear stress and atherosclerosis	149	52	11	p<0,01
Alzheimer disease	426	123	13	p<0,01
TGF-beta signaling pathway	103	39	12	p<0,01
mRNA surveillance pathway	108	40	11	0,001
AGE-RAGE signaling pathway in diabetic complications	115	42	12	0,001
Sphingolipid signaling pathway	133	47	13	0,001

KEGG pathway	continuação			
	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Breast cancer	163	55	13	0,001
Thyroid cancer	43	20	13	0,001
Epithelial cell signaling in Helicobacter pylori infection	79	31	10	0,001
Hepatitis C	173	57	13	0,001
Growth hormone synthesis, secretion and action	129	45	12	0,001
Pathways of neurodegeneration - multiple diseases	539	148	13	0,001
Huntington disease	339	99	12	0,001
Vasopressin-regulated water reabsorption	47	21	10	0,001
C-type lectin receptor signaling pathway	116	41	12	0,001
TNF signaling pathway	131	45	13	0,001
Biosynthesis of amino acids	78	30	10	0,002
cGMP-PKG signaling pathway	175	56	12	0,002
Signaling pathways regulating pluripotency of stem cells	156	51	13	0,002
Progesterone-mediated oocyte maturation	104	37	11	0,002
T cell receptor signaling pathway	115	40	13	0,002
Mitophagy - animal	76	29	12	0,002
Cushing syndrome	176	56	12	0,002
Various types of N-glycan biosynthesis	40	18	10	0,003
Gastric cancer	162	52	13	0,003
Ras signaling pathway	241	72	13	0,003
Human cytomegalovirus infection	306	88	13	0,004
Parathyroid hormone synthesis, secretion and action	118	40	12	0,004
Thermogenesis	252	74	13	0,005
Kaposi sarcoma-associated herpesvirus infection	245	72	13	0,005
Apoptosis	151	48	13	0,005
Selenocompound metabolism	18	10	8	0,005
Platelet activation	136	44	11	0,006
Choline metabolism in cancer	106	36	12	0,006
Apelin signaling pathway	140	45	13	0,006
Hippo signaling pathway - multiple species	30	14	7	0,006
Glucagon signaling pathway	114	38	13	0,006
Phosphatidylinositol signaling system	101	34	10	0,008
Circadian rhythm	31	14	10	0,008
Terpenoid backbone biosynthesis	22	11	9	0,009
Lysine degradation	69	25	9	0,010
Hedgehog signaling pathway	59	22	11	0,011
Notch signaling pathway	63	23	11	0,012
Vibrio cholerae infection	60	22	9	0,014
Insulin resistance	124	39	13	0,015
SNARE interactions in vesicular transport	33	14	9	0,016
Gap junction	99	32	13	0,020
Transcriptional misregulation in cancer	206	59	13	0,021
Alanine, aspartate and glutamate metabolism	38	15	9	0,025
GnRH signaling pathway	97	31	12	0,027
PD-L1 expression and PD-1 checkpoint pathway in cancer	101	32	12	0,027
Wnt signaling pathway	173	50	11	0,030
Cysteine and methionine metabolism	53	19	10	0,030
Longevity regulating pathway - multiple species	79	26	12	0,031
VEGF signaling pathway	61	21	12	0,034
Alcoholism	195	55	13	0,034
MicroRNAs in cancer	334	88	13	0,037
Non-alcoholic fatty liver disease	177	50	12	0,044
Oxidative phosphorylation	157	45	11	0,045
Parkinson disease	282	75	11	0,047
Phospholipase D signaling pathway	158	45	13	0,050

Tabela S6 – Vias enriquecidas para DE-miRNAs *down* - LUSC.

<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Shigellosis	268	184	46	p>0,01
Proteoglycans in cancer	220	156	45	p>0,01
Renal cell carcinoma	70	63	43	p>0,01
Cell cycle	129	101	45	p>0,01
Pathways in cancer	555	333	46	p>0,01
Ubiquitin mediated proteolysis	142	107	44	p>0,01
Focal adhesion	213	147	45	p>0,01
Autophagy - animal	146	108	44	p>0,01
Prostate cancer	101	80	44	p>0,01
FoxO signaling pathway	139	102	45	p>0,01
Spinocerebellar ataxia	145	105	43	p>0,01
Colorectal cancer	88	70	43	p>0,01
Regulation of actin cytoskeleton	224	148	44	p>0,01
Salmonella infection	277	175	45	p>0,01
Protein processing in endoplasmic reticulum	194	130	45	p>0,01
EGFR tyrosine kinase inhibitor resistance	82	65	45	p>0,01
Pathways of neurodegeneration - multiple diseases	539	309	45	p>0,01
Amyotrophic lateral sclerosis	408	242	45	p>0,01
Huntington disease	339	206	45	p>0,01
Alzheimer disease	426	250	45	p>0,01
Rap1 signaling pathway	214	139	46	p>0,01
Adherens junction	79	61	45	p>0,01
Pancreatic cancer	78	60	44	p>0,01
Parkinson disease	282	172	45	p>0,01
Chronic myeloid leukemia	79	60	44	p>0,01
Apelin signaling pathway	140	95	43	p>0,01
ErbB signaling pathway	86	64	43	p>0,01
Axon guidance	186	120	44	p>0,01
HIF-1 signaling pathway	112	79	44	p>0,01
Endocytosis	311	184	44	p>0,01
PI3K-Akt signaling pathway	372	214	45	p>0,01
Fluid shear stress and atherosclerosis	149	98	44	p>0,01
Bacterial invasion of epithelial cells	80	59	45	p>0,01
p53 signaling pathway	75	56	44	p>0,01
Neurotrophin signaling pathway	124	84	45	p>0,01
AMPK signaling pathway	130	87	45	p>0,01
Alcoholism	195	122	43	p>0,01
Hepatocellular carcinoma	177	112	45	p>0,01
Thyroid hormone signaling pathway	137	90	45	p>0,01
Viral carcinogenesis	265	157	46	p>0,01
Small cell lung cancer	100	69	45	p>0,01
Acute myeloid leukemia	69	51	43	p>0,01
Hedgehog signaling pathway	59	45	42	p>0,01
Mitophagy - animal	76	55	45	p>0,01
AGE-RAGE signaling pathway in diabetic complications	115	77	44	p>0,01
Hepatitis B	177	110	45	p>0,01
Hippo signaling pathway	164	103	44	p>0,01
Transcriptional misregulation in cancer	206	125	45	p>0,01
Tight junction	182	112	45	p>0,01
MAPK signaling pathway	329	187	46	p>0,01
Insulin resistance	124	81	44	p>0,01
Yersinia infection	147	93	44	p>0,01
Phosphatidylinositol signaling system	101	68	42	p>0,01
TGF-beta signaling pathway	103	69	43	p>0,01
Longevity regulating pathway	105	70	45	p>0,01
Endometrial cancer	61	45	44	p>0,01
N-Glycan biosynthesis	51	39	40	p>0,01
Glioma	79	55	44	p>0,01
mRNA surveillance pathway	108	71	43	p>0,01
Non-small cell lung cancer	81	56	45	p>0,01
mTOR signaling pathway	177	107	44	p>0,01
RNA transport	199	118	44	p>0,01
Apoptosis	151	93	44	p>0,01
Gap junction	99	65	44	p>0,01

<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Pathogenic Escherichia coli infection	222	129	45	p>0,01
Central carbon metabolism in cancer	74	51	44	p>0,01
Signaling pathways regulating pluripotency of stem cells	156	95	44	p>0,01
Insulin signaling pathway	153	92	44	p>0,01
Relaxin signaling pathway	138	84	43	p>0,01
Neutrophil extracellular trap formation	205	118	42	p>0,01
Ras signaling pathway	241	136	45	p>0,01
Growth hormone synthesis, secretion and action	129	79	44	p>0,01
Platinum drug resistance	75	50	43	p>0,01
Various types of N-glycan biosynthesis	40	30	39	p>0,01
Human papillomavirus infection	406	216	46	p>0,01
Oocyte meiosis	134	81	44	p>0,01
Progesterone-mediated oocyte maturation	104	65	43	p>0,01
Gastric cancer	162	95	44	p>0,01
Choline metabolism in cancer	106	66	44	p>0,01
Sphingolipid signaling pathway	133	80	43	p>0,01
Lysosome	145	86	41	p>0,01
Long-term potentiation	71	47	40	p>0,01
Breast cancer	163	95	45	0,001
Circadian rhythm	31	24	39	0,001
Thermogenesis	252	139	45	0,001
TNF signaling pathway	131	78	44	0,001
Dopaminergic synapse	143	84	43	0,001
Melanoma	76	49	45	0,001
Inositol phosphate metabolism	78	50	43	0,001
Oxytocin signaling pathway	161	93	42	0,001
Endocrine resistance	118	71	45	0,001
Fc gamma R-mediated phagocytosis	101	62	42	0,001
Prolactin signaling pathway	73	47	44	0,001
Apoptosis - multiple species	32	24	40	0,001
Longevity regulating pathway - multiple species	79	50	44	0,001
Wnt signaling pathway	173	98	44	0,002
Ferroptosis	41	29	38	0,002
Metabolic pathways	1634	775	46	0,002
Hepatitis C	173	97	45	0,002
Prion disease	331	174	45	0,002
Non-alcoholic fatty liver disease	177	99	44	0,002
Long-term depression	64	41	43	0,003
Cysteine and methionine metabolism	53	35	39	0,003
Leukocyte transendothelial migration	120	70	43	0,003
Cellular senescence	219	119	45	0,003
Lysine degradation	69	43	41	0,005
Bladder cancer	43	29	42	0,005
Chagas disease	116	67	44	0,005
Platelet activation	136	77	43	0,005
cGMP-PKG signaling pathway	175	96	43	0,006
SNARE interactions in vesicular transport	33	23	40	0,007
VEGF signaling pathway	61	38	40	0,008
Glucagon signaling pathway	114	65	44	0,009
Aldosterone-regulated sodium reabsorption	37	25	40	0,009
Hippo signaling pathway - multiple species	30	21	37	0,010
Carbon metabolism	122	68	43	0,015
DNA replication	38	25	33	0,015
Parathyroid hormone synthesis, secretion and action	118	66	43	0,015
Aminoacyl-tRNA biosynthesis	57	35	37	0,016
Notch signaling pathway	63	38	42	0,017
Autophagy - other	33	22	35	0,019
Mismatch repair	24	17	32	0,019
Human immunodeficiency virus 1 infection	277	142	44	0,019
Phospholipase D signaling pathway	158	85	43	0,019
GnRH signaling pathway	97	55	41	0,019
Circadian entrainment	100	56	43	0,025
Necroptosis	174	92	44	0,026
Protein export	23	16	39	0,030

continuação

				continuação
<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	FDR
Epithelial cell signaling in Helicobacter pylori infection	79	45	44	0,033
Alanine, aspartate and glutamate metabolism	38	24	39	0,034
Measles	161	85	46	0,034
Kaposi sarcoma-associated herpesvirus infection	245	125	45	0,034
Cholinergic synapse	122	66	41	0,036
Cushing syndrome	176	92	44	0,036
Selenocompound metabolism	18	13	31	0,036
Spliceosome	193	100	43	0,037
Vibrio cholerae infection	60	35	43	0,041
Basal cell carcinoma	66	38	40	0,042
Base excision repair	33	21	35	0,042
Terpenoid backbone biosynthesis	22	15	32	0,045
Mannose type O-glycan biosynthesis	24	16	31	0,049
Pyruvate metabolism	43	26	39	0,049
Thyroid cancer	43	26	41	0,049
Fatty acid degradation	43	26	37	0,049

Tabela S7 – Processos biológicos relacionados a Rede PPI - LUSC.

GO-termo	Descrição	Contagem na rede	Força	FDR
GO:0090402	Oncogene-induced cell senescence	2 of 4	2,04	0,04
GO:0060736	Prostate gland growth	3 of 12	1,74	0,01
GO:1902969	Mitotic DNA replication	3 of 12	1,74	0,01
GO:0006268	DNA unwinding involved in DNA replication	4 of 22	1,6	0,00
GO:0006270	DNA replication initiation	4 of 29	1,48	0,00
GO:0090329	Regulation of DNA-templated DNA replication	5 of 57	1,28	0,00
GO:0000082	G1/S transition of mitotic cell cycle	7 of 86	1,25	0,00
GO:0044772	Mitotic cell cycle phase transition	10 of 173	1,1	0,00
GO:0032465	Regulation of cytokinesis	5 of 94	1,07	0,02
GO:0035265	Organ growth	5 of 98	1,05	0,02
GO:0006260	DNA replication	9 of 203	0,99	0,00
GO:0000079	Regulation of cyclin-dependent protein serine/threonine kinase activity	5 of 111	0,99	0,03
GO:0006261	DNA-templated DNA replication	6 of 137	0,98	0,01
GO:2000278	Regulation of DNA biosynthetic process	5 of 124	0,95	0,04
GO:0032147	Activation of protein kinase activity	7 of 203	0,88	0,01
GO:1903047	Mitotic cell cycle process	18 of 537	0,87	0,00
GO:0051302	Regulation of cell division	6 of 188	0,84	0,04
GO:0071900	Regulation of protein serine/threonine kinase activity	12 of 386	0,83	0,00
GO:0071902	Positive regulation of protein serine/threonine kinase activity	7 of 226	0,83	0,02
GO:0045787	Positive regulation of cell cycle	10 of 349	0,8	0,00
GO:0090068	Positive regulation of cell cycle process	7 of 251	0,79	0,03
GO:0022402	Cell cycle process	21 of 835	0,74	0,00
GO:0045860	Positive regulation of protein kinase activity	10 of 416	0,72	0,01
GO:0045859	Regulation of protein kinase activity	15 of 663	0,69	0,00
GO:0051347	Positive regulation of transferase activity	13 of 586	0,69	0,00
GO:0048589	Developmental growth	9 of 412	0,68	0,02
GO:0007049	Cell cycle	25 of 1246	0,64	0,00
GO:0010564	Regulation of cell cycle process	14 of 716	0,63	0,00
GO:0031667	Response to nutrient levels	9 of 461	0,63	0,04
GO:0051726	Regulation of cell cycle	21 of 1108	0,62	0,00
GO:0006259	DNA metabolic process	15 of 785	0,62	0,00
GO:0051301	Cell division	10 of 527	0,62	0,03
GO:0051338	Regulation of transferase activity	17 of 912	0,61	0,00
GO:0051052	Regulation of DNA metabolic process	10 of 541	0,61	0,03
GO:0008283	Cell population proliferation	12 of 712	0,57	0,02
GO:0008285	Negative regulation of cell population proliferation	12 of 713	0,57	0,02
GO:0051276	Chromosome organization	16 of 968	0,56	0,00
GO:0001932	Regulation of protein phosphorylation	18 of 1108	0,55	0,00
GO:0010558	Negative regulation of macromolecule biosynthetic process	23 of 1532	0,52	0,00
GO:0031327	Negative regulation of cellular biosynthetic process	23 of 1592	0,5	0,00
GO:0000122	Negative regulation of transcription by RNA polymerase II	14 of 971	0,5	0,02
GO:0031399	Regulation of protein modification process	22 of 1560	0,49	0,00
GO:0043085	Positive regulation of catalytic activity	17 of 1191	0,49	0,01
GO:0010628	Positive regulation of gene expression	16 of 1146	0,49	0,01
GO:0031401	Positive regulation of protein modification process	14 of 1018	0,48	0,03
GO:0045892	Negative regulation of transcription, DNA-templated	18 of 1322	0,47	0,01
GO:0031324	Negative regulation of cellular metabolic process	30 of 2265	0,46	0,00
GO:0042127	Regulation of cell population proliferation	22 of 1669	0,46	0,00
GO:0044093	Positive regulation of molecular function	21 of 1587	0,46	0,00
GO:0051253	Negative regulation of RNA metabolic process	19 of 1434	0,46	0,01
GO:0006796	Phosphate-containing compound metabolic process	24 of 1877	0,45	0,00
GO:0045934	Negative regulation of nucleobase-containing compound metabolic process	20 of 1562	0,45	0,01
GO:0051172	Negative regulation of nitrogen compound metabolic process	29 of 2403	0,42	0,00
GO:0033554	Cellular response to stress	19 of 1572	0,42	0,01
GO:0051246	Regulation of protein metabolic process	31 of 2622	0,41	0,00
GO:0010605	Negative regulation of macromolecule metabolic process	31 of 2760	0,39	0,00
GO:0031325	Positive regulation of cellular metabolic process	33 of 3114	0,37	0,00
GO:0010604	Positive regulation of macromolecule metabolic process	37 of 3533	0,36	0,00
GO:0051173	Positive regulation of nitrogen compound metabolic process	33 of 3166	0,36	0,00
GO:0046483	Heterocycle metabolic process	30 of 2891	0,36	0,00
GO:0050790	Regulation of catalytic activity	25 of 2370	0,36	0,01
GO:0071310	Cellular response to organic substance	21 of 2019	0,36	0,04

		continuação		
GO-termo	Descrição	Contagem na rede	Força	FDR
GO:0009893	Positive regulation of metabolic process	39 of 3847	0,35	0,00
GO:0006139	Nucleobase-containing compound metabolic process	28 of 2722	0,35	0,01
GO:0009605	Response to external stimulus	24 of 2355	0,35	0,02
GO:0070887	Cellular response to chemical stimulus	26 of 2609	0,34	0,01
GO:0006725	Cellular aromatic compound metabolic process	29 of 2936	0,33	0,01
GO:0009653	Anatomical structure morphogenesis	22 of 2229	0,33	0,05
GO:0031323	Regulation of cellular metabolic process	54 of 5681	0,32	0,00
GO:0034641	Cellular nitrogen compound metabolic process	33 of 3463	0,32	0,00
GO:0006996	Organelle organization	32 of 3470	0,31	0,01
GO:1901360	Organic cyclic compound metabolic process	30 of 3181	0,31	0,01
GO:0065009	Regulation of molecular function	29 of 3085	0,31	0,01
GO:0048523	Negative regulation of cellular process	43 of 4736	0,3	0,00
GO:0080090	Regulation of primary metabolic process	53 of 5899	0,29	0,00
GO:0051171	Regulation of nitrogen compound metabolic process	51 of 5734	0,29	0,00
GO:0010556	Regulation of macromolecule biosynthetic process	35 of 3980	0,28	0,01
GO:0006950	Response to stress	29 of 3358	0,28	0,04
GO:0048518	Positive regulation of biological process	53 of 6207	0,27	0,00
GO:0060255	Regulation of macromolecule metabolic process	53 of 6249	0,27	0,00
GO:0009889	Regulation of biosynthetic process	36 of 4205	0,27	0,01
GO:0019222	Regulation of metabolic process	57 of 6784	0,26	0,00
GO:0044237	Cellular metabolic process	54 of 6568	0,26	0,00
GO:0048522	Positive regulation of cellular process	46 of 5584	0,26	0,00
GO:0007275	Multicellular organism development	35 of 4209	0,26	0,02
GO:0031326	Regulation of cellular biosynthetic process	34 of 4143	0,25	0,03
GO:0019219	Regulation of nucleobase-containing compound metabolic process	33 of 4074	0,25	0,05
GO:0051716	Cellular response to stimulus	48 of 6357	0,22	0,01
GO:0016043	Cellular component organization	41 of 5436	0,22	0,03
GO:0048856	Anatomical structure development	39 of 5117	0,22	0,04
GO:0006807	Nitrogen compound metabolic process	49 of 6643	0,21	0,01
GO:0043170	Macromolecule metabolic process	43 of 5781	0,21	0,03
GO:0071840	Cellular component organization or biogenesis	42 of 5639	0,21	0,03
GO:0032502	Developmental process	42 of 5657	0,21	0,03
GO:0071704	Organic substance metabolic process	55 of 7522	0,2	0,00
GO:0044238	Primary metabolic process	52 of 7156	0,2	0,01
GO:0008152	Metabolic process	57 of 7988	0,19	0,00
GO:0050896	Response to stimulus	55 of 7835	0,19	0,01
GO:0050794	Regulation of cellular process	72 of 11025	0,16	0,00
GO:0050789	Regulation of biological process	74 of 11655	0,14	0,00
GO:0065007	Biological regulation	75 of 12385	0,12	0,01
GO:0009987	Cellular process	85 of 14826	0,1	0,00

Tabela S8 – Maiores correlações negativas entre os pares - LUAD-LUSC.

miRNA	Gene	corr	p	miRNA	Gene	corr	p	miRNA	Gene	corr	p
miR-375-3p	CLCA2	-0,65	p<0,05	miR-29a-3p	HMGCS1	-0,47	p<0,05	miR-181a-5p	ACAP2	-0,43	p<0,05
miR-944	SLC41A2	-0,63	p<0,05	miR-708-5p	DRAM1	-0,47	p<0,05	miR-29a-3p	HMGCR	-0,43	p<0,05
miR-205-5p	ENPP4	-0,62	p<0,05	miR-100-5p	BIRC5	-0,47	p<0,05	miR-944	RAPGEF2	-0,43	p<0,05
miR-375-3p	CDIN1	-0,61	p<0,05	miR-375-3p	ISL2	-0,46	p<0,05	miR-30b-3p	SKA1	-0,43	p<0,05
miR-205-5p	C11orf86	-0,61	p<0,05	miR-30d-5p	PTGFRN	-0,46	p<0,05	miR-149-5p	PRR15L	-0,43	p<0,05
miR-205-5p	HSD17B11	-0,60	p<0,05	miR-375-3p	KIAA1549L	-0,46	p<0,05	miR-30b-5p	HOXA1	-0,43	p<0,05
miR-375-3p	SHOX2	-0,59	p<0,05	miR-181b-5p	BAG4	-0,46	p<0,05	miR-146b-3p	DYNLL1	-0,43	p<0,05
miR-205-3p	TOX3	-0,59	p<0,05	miR-26a-5p	HMGAI	-0,46	p<0,05	miR-30b-5p	CAPZA1	-0,43	p<0,05
miR-205-5p	LPCAT1	-0,58	p<0,05	miR-664b-3p	TMTC3	-0,46	p<0,05	miR-26a-5p	BOD1	-0,43	p<0,05
miR-30d-5p	DBF4	-0,58	p<0,05	miR-19d-5p	E2F7	-0,46	p<0,05	miR-29b-3p	METAP2	-0,43	p<0,05
miR-29a-3p	TMTC3	-0,58	p<0,05	miR-205-5p	JADE1	-0,46	p<0,05	miR-30d-5p	EED	-0,43	p<0,05
miR-29a-3p	SLC16A1	-0,57	p<0,05	miR-195-5p	CHEK1	-0,46	p<0,05	miR-26a-5p	UBE2G1	-0,43	p<0,05
miR-29a-3p	TDG	-0,56	p<0,05	miR-181a-5p	BAG4	-0,46	p<0,05	miR-29a-3p	TAF5	-0,43	p<0,05
miR-375-3p	PAX6	-0,56	p<0,05	miR-205-5p	LYSMD3	-0,46	p<0,05	miR-146b-5p	UBE2G1	-0,43	p<0,05
miR-205-3p	CELF2	-0,55	p<0,05	miR-361-3p	PTDSS1	-0,45	p<0,05	miR-107	PDCD10	-0,42	p<0,05
miR-30d-5p	TWF1	-0,55	p<0,05	miR-140-3p	KANSL2	-0,45	p<0,05	let-7e-5p	CDC25A	-0,42	p<0,05
miR-30d-5p	ITGA6	-0,55	p<0,05	miR-944	DPYSL2	-0,45	p<0,05	miR-100-5p	GMP5	-0,42	p<0,05
miR-205-5p	RBM47	-0,54	p<0,05	miR-30d-5p	RAP1B	-0,45	p<0,05	miR-140-3p	KCMF1	-0,42	p<0,05
miR-205-5p	ERBB3	-0,54	p<0,05	miR-29a-3p	PDHX	-0,45	p<0,05	miR-181b-5p	C2orf69	-0,42	p<0,05
miR-30d-5p	B3GNT5	-0,54	p<0,05	miR-181a-5p	METAP1	-0,45	p<0,05	let-7g-5p	GOLT1B	-0,42	p<0,05
miR-29a-3p	POLE3	-0,53	p<0,05	miR-139-5p	KPNA4	-0,45	p<0,05	let-7e-5p	GOLT1B	-0,42	p<0,05
miR-4709-3p	KPNA4	-0,53	p<0,05	miR-30d-5p	YWHAZ	-0,45	p<0,05	miR-29c-3p	SLC16A1	-0,42	p<0,05
miR-196b-5p	CPM	-0,52	p<0,05	miR-29a-3p	ELOVL4	-0,45	p<0,05	miR-664a-5p	METTL8	-0,42	p<0,05
miR-205-5p	MGAT4A	-0,51	p<0,05	miR-26a-5p	LSM12	-0,45	p<0,05	miR-375-3p	AEBP2	-0,42	p<0,05
miR-26a-5p	MAPK6	-0,51	p<0,05	miR-423-5p	AP2M1	-0,45	p<0,05	miR-205-5p	ERRF1	-0,42	p<0,05
miR-944	SGPP1	-0,51	p<0,05	miR-26a-5p	TWF1	-0,45	p<0,05	miR-29a-3p	MLF1	-0,42	p<0,05
miR-375-3p	YBX1	-0,51	p<0,05	miR-195-5p	WBP11	-0,45	p<0,05	miR-423-5p	CNBP	-0,42	p<0,05
miR-181b-5p	KPNA4	-0,51	p<0,05	miR-195-5p	ATP13A3	-0,45	p<0,05	miR-4728-3p	EXO1	-0,42	p<0,05
miR-195-5p	CDCA4	-0,51	p<0,05	miR-29b-3p	SLC16A1	-0,45	p<0,05	miR-30d-3p	LIN9	-0,42	p<0,05
miR-30d-5p	PFN2	-0,51	p<0,05	miR-30d-5p	ERLIN1	-0,45	p<0,05	miR-30d-5p	GNA13	-0,42	p<0,05
miR-29a-3p	PXYLP1	-0,50	p<0,05	miR-340-5p	FRMD6	-0,45	p<0,05	miR-26a-5p	SSX2IP	-0,42	p<0,05
miR-29a-3p	MLLT11	-0,50	p<0,05	miR-181a-5p	TBPL1	-0,45	p<0,05	miR-139-5p	TMPO	-0,42	p<0,05
miR-140-3p	OLA1	-0,50	p<0,05	miR-146b-5p	SLAH2	-0,45	p<0,05	miR-4728-3p	CDIN1	-0,42	p<0,05
miR-205-5p	SLC4A4	-0,50	p<0,05	miR-26a-5p	CDK2AP1	-0,45	p<0,05	miR-181b-5p	RAP1B	-0,42	p<0,05
miR-30b-5p	ITGA6	-0,50	p<0,05	miR-30d-5p	LCLAT1	-0,44	p<0,05	miR-30d-3p	ORC5	-0,42	p<0,05
miR-181a-5p	KPNA4	-0,50	p<0,05	let-7g-5p	CDC25A	-0,44	p<0,05	miR-29a-3p	RLF	-0,42	p<0,05
miR-4709-3p	NMD3	-0,49	p<0,05	miR-26a-5p	PDHX	-0,44	p<0,05	miR-146b-5p	LNPB	-0,42	p<0,05
miR-29c-3p	TDG	-0,49	p<0,05	miR-3189-3p	CDC25A	-0,44	p<0,05	miR-29b-3p	PXYLP1	-0,42	p<0,05
miR-196b-5p	CELF2	-0,49	p<0,05	miR-195-3p	PDCD10	-0,44	p<0,05	miR-26a-5p	EIF5	-0,42	p<0,05
miR-26a-5p	PAK2	-0,49	p<0,05	miR-664a-3p	MYNN	-0,44	p<0,05	let-7g-5p	HMGAI	-0,42	p<0,05
miR-30b-5p	TWF1	-0,49	p<0,05	miR-26a-5p	B3GNT5	-0,44	p<0,05	let-7e-5p	MAPK6	-0,42	p<0,05
miR-26a-5p	E2F7	-0,49	p<0,05	miR-375-3p	PPARGC1B	-0,44	p<0,05	miR-30a-5p	R3HDM1	-0,42	p<0,05
miR-26a-5p	RNC2	-0,49	p<0,05	miR-944	MACC1	-0,44	p<0,05	miR-361-3p	PSMD2	-0,42	p<0,05
miR-26a-5p	SEN5	-0,49	p<0,05	miR-1266-5p	ITGA6	-0,44	p<0,05	miR-1266-5p	NTRK2	-0,42	p<0,05
miR-195-5p	KIF23	-0,49	p<0,05	let-7g-5p	SEN5	-0,44	p<0,05	miR-4709-3p	U2SURP	-0,42	p<0,05
miR-29b-3p	TMTC3	-0,49	p<0,05	miR-30b-5p	ERLIN1	-0,44	p<0,05	miR-4709-3p	BTBD10	-0,42	p<0,05
miR-30d-5p	ATG5	-0,49	p<0,05	miR-664a-3p	MAPK6	-0,44	p<0,05	miR-139-5p	YWHAQ	-0,42	p<0,05
miR-29a-3p	CCSAP	-0,49	p<0,05	miR-4709-3p	CNBP	-0,44	p<0,05	miR-100-5p	CDC25A	-0,42	p<0,05
miR-29a-3p	TMEM183A	-0,49	p<0,05	miR-1266-5p	PRIMA1	-0,44	p<0,05	miR-30d-5p	GCLC	-0,42	p<0,05
miR-30d-5p	R3HDM1	-0,48	p<0,05	miR-944	UBL3	-0,44	p<0,05	miR-326	CNBP	-0,42	p<0,05
miR-26a-5p	EIF2S1	-0,48	p<0,05	miR-26a-5p	CDK8	-0,44	p<0,05	miR-29a-3p	TET3	-0,42	p<0,05
miR-5698	AP2M1	-0,48	p<0,05	miR-29a-3p	POGLUT2	-0,44	p<0,05	miR-195-5p	TBPL1	-0,42	p<0,05
let-7g-5p	MAPK6	-0,48	p<0,05	miR-3189-3p	IPPK	-0,44	p<0,05	miR-375-3p	UST	-0,42	p<0,05
miR-205-3p	PCDHAC1	-0,48	p<0,05	miR-30b-5p	R3HDM1	-0,44	p<0,05	miR-664a-3p	KCMF1	-0,42	p<0,05
miR-29a-3p	YBX3	-0,48	p<0,05	miR-149-5p	KCNK5	-0,44	p<0,05	miR-4728-3p	ZNF367	-0,42	p<0,05
miR-205-5p	SLC35A1	-0,48	p<0,05	miR-1266-5p	SLC2A1	-0,44	p<0,05	miR-375-3p	KLF5	-0,42	p<0,05
miR-181a-5p	DCUN1D1	-0,48	p<0,05	miR-3065-3p	DBF4	-0,44	p<0,05	miR-205-5p	TAPT1	-0,42	p<0,05
miR-29a-3p	FEM1B	-0,48	p<0,05	miR-29a-3p	METAP2	-0,43	p<0,05	miR-29a-3p	COL4A5	-0,42	p<0,05
miR-181b-5p	NAA50	-0,48	p<0,05	miR-1287-5p	DSG1	-0,43	p<0,05	miR-340-5p	PNO1	-0,42	p<0,05
miR-664a-3p	STRAP	-0,48	p<0,05	miR-30d-5p	KPNA3	-0,43	p<0,05	miR-2110	TTL	-0,42	p<0,05
miR-205-5p	MGRN1	-0,48	p<0,05	miR-30d-5p	LRRC8D	-0,43	p<0,05	miR-151b	LYPD3	-0,42	p<0,05
miR-29a-3p	NUDT11	-0,48	p<0,05	miR-30b-3p	YWHAQ	-0,43	p<0,05	miR-26a-5p	KLHL42	-0,42	p<0,05
miR-30b-5p	B3GNT5	-0,48	p<0,05	miR-140-3p	GOLT1B	-0,43	p<0,05	miR-375-3p	YWHAZ	-0,42	p<0,05
miR-30d-5p	CAPZA1	-0,48	p<0,05	miR-26a-5p	PPP3R1	-0,43	p<0,05	miR-29a-3p	RNF138	-0,42	p<0,05
miR-664a-3p	NRAS	-0,47	p<0,05	miR-30b-5p	RAP1B	-0,43	p<0,05	miR-26a-5p	ERLIN1	-0,42	p<0,05
miR-4709-3p	SNAI2	-0,47	p<0,05	miR-139-5p	ATP5MC3	-0,43	p<0,05	miR-1910-5p	GPR39	-0,41	p<0,05
miR-205-3p	ZDHHC21	-0,47	p<0,05	miR-195-5p	YWHAQ	-0,43	p<0,05	miR-423-5p	CDC42SE1	-0,41	p<0,05
miR-30d-5p	TFDP1	-0,47	p<0,05	miR-146b-5p	KCMF1	-0,43	p<0,05	miR-139-3p	RAB10	-0,41	p<0,05
miR-181b-5p	ACAP2	-0,47	p<0,05	miR-29a-3p	COL7A1	-0,43	p<0,05	miR-29a-3p	CDK6	-0,41	p<0,05
miR-30d-3p	PDK1	-0,47	p<0,05	miR-29c-3p	TMEM183A	-0,43	p<0,05	let-7g-5p	HOXA1	-0,41	p<0,05
miR-29c-3p	TMTC3	-0,47	p<0,05	miR-1287-5p	CDCA4	-0,43	p<0,05	miR-31-5p	ATP8A1	-0,41	p<0,05
miR-664b-3p	TFRC	-0,47	p<0,05	miR-29b-3p	FEM1B	-0,43	p<0,05	miR-203a-3p	SLC4A4	-0,41	p<0,05
miR-181b-5p	TBPL1	-0,47	p<0,05	miR-205-5p	TMEM144	-0,43	p<0,05	miR-26a-5p	JAG1	-0,41	p<0,05
miR-30d-5p	RARG	-0,47	p<0,05	miR-21-5p	MED21	-0,43	p<0,05	miR-146b-5p	SLC38A2	-0,41	p<0,05
miR-181b-5p	METAP1	-0,47	p<0,05	miR-181b-5p	KPNA1	-0,43	p<0,05	miR-3189-3p	BICD2	-0,41	p<0,05
miR-29b-3p	TDG	-0,47	p<0,05	miR-326	RARG	-0,43	p<0,05	miR-205-3p	TMEM144	-0,41	p<0,05
miR-29a-3p	BLMH	-0,47	p<0,05	miR-3065-3p	MLLT11	-0,43	p<0,05	miR-29a-3p	NUP160	-0,41	p<0,05

										continuação	
miRNA	Gene	corr	p	miRNA	Gene	corr	p	miRNA	Gene	corr	p
miR-181b-5p	ATP2B1	-0,41	p<0,05	miR-30b-3p	SENP5	-0,41	p<0,05	let-7g-5p	NRAS	-0,40	p<0,05
miR-139-5p	CTDSPL2	-0,41	p<0,05	miR-29b-3p	POGLUT2	-0,41	p<0,05	miR-30b-5p	GNA13	-0,40	p<0,05
miR-30d-5p	ATP2B1	-0,41	p<0,05	miR-1-3p	HSPD1	-0,41	p<0,05	miR-92b-3p	RAP1B	-0,40	p<0,05
miR-30b-5p	ATG5	-0,41	p<0,05	miR-29a-3p	MORF4L1	-0,41	p<0,05	miR-29a-3p	PRKAB2	-0,40	p<0,05
miR-195-5p	TBL1XR1	-0,41	p<0,05	miR-140-3p	RNF7	-0,41	p<0,05	miR-21-5p	SRSF3	-0,40	p<0,05
miR-181a-5p	RAP1B	-0,41	p<0,05	miR-29a-3p	ISG20L2	-0,41	p<0,05	miR-340-5p	DCUN1D1	-0,40	p<0,05
miR-139-3p	SF3B5	-0,41	p<0,05	miR-30d-5p	RAB32	-0,41	p<0,05	miR-375-3p	TSC22D2	-0,40	p<0,05
let-7e-5p	RIOK3	-0,41	p<0,05	miR-181b-5p	LRRC8D	-0,41	p<0,05	miR-181d-5p	METAP1	-0,40	p<0,05
miR-196b-5p	PRTG	-0,41	p<0,05	miR-5698	DVL3	-0,41	p<0,05	miR-26a-5p	SLC38A2	-0,40	p<0,05
miR-29a-3p	NKRF	-0,41	p<0,05	miR-375-3p	SYNCRIP	-0,41	p<0,05	miR-26a-5p	NAB1	-0,40	p<0,05
let-7a-5p	RRM2	-0,41	p<0,05	miR-4709-3p	ALG10	-0,41	p<0,05	miR-181a-5p	C2orf69	-0,40	p<0,05
miR-30b-5p	PTGFRN	-0,41	p<0,05	miR-146b-3p	TCPI1	-0,41	p<0,05	miR-26a-5p	SERBP1	-0,40	p<0,05
miR-34a-5p	NAA50	-0,41	p<0,05	miR-93-5p	STK17B	-0,41	p<0,05	miR-664a-3p	OSBPL8	-0,40	p<0,05
miR-29a-3p	ZBTB5	-0,41	p<0,05	miR-146b-5p	LIN7C	-0,41	p<0,05	miR-9-5p	SGMS2	-0,40	p<0,05
miR-26a-5p	ADM	-0,41	p<0,05	miR-3189-3p	TAF4B	-0,41	p<0,05	miR-30b-5p	LCLAT1	-0,40	p<0,05
miR-140-3p	DCUN1D1	-0,41	p<0,05	miR-26a-5p	EZH2	-0,41	p<0,05	miR-340-5p	ACTR3	-0,40	p<0,05
miR-664a-5p	MAPK6	-0,41	p<0,05	miR-195-3p	GOLT1B	-0,41	p<0,05	miR-29b-3p	TMEM183A	-0,40	p<0,05
miR-139-5p	DENR	-0,41	p<0,05	miR-30b-5p	TFDP1	-0,40	p<0,05	miR-4652-5p	ATP11A	-0,40	p<0,05
miR-5698	FOXE1	-0,41	p<0,05	miR-205-5p	CASD1	-0,40	p<0,05	miR-664a-5p	CDK6	-0,40	p<0,05

Tabela S9 – Vias enriquecidas para DE-miRNAs *down* -LUAD-LUSC.

<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Renal cell carcinoma	70	57	16	p<0,01
Proteoglycans in cancer	220	130	16	p<0,01
Shigellosis	268	150	16	p<0,01
Pathways in cancer	555	270	16	p<0,01
Autophagy - animal	146	93	16	p<0,01
Ubiquitin mediated proteolysis	142	90	16	p<0,01
Protein processing in endoplasmic reticulum	194	114	16	p<0,01
FoxO signaling pathway	139	88	16	p<0,01
Focal adhesion	213	122	16	p<0,01
Salmonella infection	277	149	16	p<0,01
Colorectal cancer	88	61	16	p<0,01
Prostate cancer	101	66	16	p<0,01
Cell cycle	129	79	16	p<0,01
Alzheimer disease	426	206	16	p<0,01
p53 signaling pathway	75	52	16	p<0,01
Adherens junction	79	54	16	p<0,01
Regulation of actin cytoskeleton	224	120	16	p<0,01
Spinocerebellar ataxia	145	85	16	p<0,01
Pancreatic cancer	78	53	16	p<0,01
Hippo signaling pathway	164	93	16	p<0,01
Pathways of neurodegeneration - multiple diseases	539	247	16	p<0,01
Neurotrophin signaling pathway	124	74	16	p<0,01
Bacterial invasion of epithelial cells	80	53	16	p<0,01
Apelin signaling pathway	140	81	16	p<0,01
Phosphatidylinositol signaling system	101	62	16	p<0,01
AGE-RAGE signaling pathway in diabetic complications	115	68	16	p<0,01
Chronic myeloid leukemia	79	51	16	p<0,01
Rap1 signaling pathway	214	111	16	p<0,01
EGFR tyrosine kinase inhibitor resistance	82	52	16	p<0,01
Huntington disease	339	162	16	p<0,01
Hepatocellular carcinoma	177	94	16	p<0,01
Longevity regulating pathway	105	62	16	p<0,01
Thyroid hormone signaling pathway	137	76	16	p<0,01
Hepatitis B	177	93	16	p<0,01
Small cell lung cancer	100	59	16	p<0,01
Oocyte meiosis	134	74	16	p<0,01
Glioma	79	49	16	p<0,01
HIF-1 signaling pathway	112	64	16	p<0,01
Apoptosis	151	81	16	p<0,01
mTOR signaling pathway	177	92	16	p<0,01
Endocytosis	311	147	16	p<0,01
Signaling pathways regulating pluripotency of stem cells	156	83	16	p<0,01
Tight junction	182	94	16	p<0,01
PI3K-Akt signaling pathway	372	170	16	p<0,01
Human papillomavirus infection	406	183	16	p<0,01
Yersinia infection	147	78	16	p<0,01
ErbB signaling pathway	86	51	16	p<0,01
Long-term potentiation	71	44	16	p<0,01
MAPK signaling pathway	329	152	16	p<0,01
Axon guidance	186	94	16	p<0,01
Apoptosis - multiple species	32	24	16	p<0,01
Hepatitis C	173	88	16	p<0,01
Pathogenic Escherichia coli infection	222	108	16	p<0,01
Insulin signaling pathway	153	79	16	p<0,01
Amyotrophic lateral sclerosis	408	181	16	p<0,01
Endometrial cancer	61	38	16	p<0,01
Insulin resistance	124	66	16	p<0,01
N-Glycan biosynthesis	51	33	16	p<0,01
Dopaminergic synapse	143	74	16	p<0,01
Parkinson disease	282	130	16	p<0,01
Prolactin signaling pathway	73	43	16	p<0,01
AMPK signaling pathway	130	68	16	p<0,01
Acute myeloid leukemia	69	41	16	p<0,01
Relaxin signaling pathway	138	71	16	p<0,01

<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	<i>FDR</i>
Mitophagy - animal	76	44	16	p<0,01
Fc gamma R-mediated phagocytosis	101	55	16	p<0,01
Central carbon metabolism in cancer	74	43	16	p<0,01
Cellular senescence	219	104	16	p<0,01
Non-small cell lung cancer	81	46	16	p<0,01
Endocrine resistance	118	62	16	p<0,01
Platinum drug resistance	75	43	16	p<0,01
Sphingolipid signaling pathway	133	68	16	p<0,01
Circadian rhythm	31	22	16	p<0,01
TNF signaling pathway	131	67	16	p<0,01
Inositol phosphate metabolism	78	44	16	p<0,01
Breast cancer	163	80	16	p<0,01
Wnt signaling pathway	173	84	16	p<0,01
Fluid shear stress and atherosclerosis	149	74	16	p<0,01
Ras signaling pathway	241	111	16	p<0,01
Growth hormone synthesis, secretion and action	129	65	16	p<0,01
TGF-beta signaling pathway	103	54	15	p<0,01
mRNA surveillance pathway	108	56	16	p<0,01
Bladder cancer	43	27	16	p<0,01
Melanoma	76	42	16	p<0,01
Gastric cancer	162	78	16	p<0,01
Progesterone-mediated oocyte maturation	104	54	16	p<0,01
Choline metabolism in cancer	106	54	16	0,001
cGMP-PKG signaling pathway	175	82	16	0,001
Transcriptional misregulation in cancer	206	94	16	0,001
Cushing syndrome	176	82	16	0,001
Oxytocin signaling pathway	161	76	16	0,001
RNA transport	199	91	16	0,001
Longevity regulating pathway - multiple species	79	42	16	0,001
Thermogenesis	252	111	16	0,001
Viral carcinogenesis	265	116	16	0,001
Gap junction	99	50	16	0,001
GnRH signaling pathway	97	49	16	0,002
Glucagon signaling pathway	114	56	16	0,002
Various types of N-glycan biosynthesis	40	24	14	0,002
Parathyroid hormone synthesis, secretion and action	118	57	15	0,002
Measles	161	74	16	0,003
Hippo signaling pathway - multiple species	30	19	14	0,003
Melanogenesis	107	52	16	0,003
Long-term depression	64	34	15	0,003
Kaposi sarcoma-associated herpesvirus infection	245	106	16	0,003
Circadian entrainment	100	49	16	0,004
Prion disease	331	138	16	0,004
Renin secretion	72	37	16	0,005
Platelet activation	136	63	16	0,005
Chagas disease	116	55	16	0,005
Human immunodeficiency virus 1 infection	277	116	16	0,008
Cysteine and methionine metabolism	53	28	14	0,010
Cholinergic synapse	122	56	16	0,010
Metabolic pathways	1634	599	16	0,011
Non-alcoholic fatty liver disease	177	77	16	0,012
Hedgehog signaling pathway	59	30	16	0,014
PD-L1 expression and PD-1 checkpoint pathway in cancer	101	47	16	0,015
Endocrine and other factor-regulated calcium reabsorption	57	29	15	0,016
Epithelial cell signaling in Helicobacter pylori infection	79	38	16	0,016
Thyroid cancer	43	23	16	0,017
Leukocyte transendothelial migration	120	54	16	0,019
T cell receptor signaling pathway	115	52	16	0,019
Estrogen signaling pathway	167	72	16	0,019
Protein export	23	14	15	0,019
Human cytomegalovirus infection	306	124	16	0,019
Aldosterone-regulated sodium reabsorption	37	20	15	0,023
Alcoholism	195	82	16	0,024
RNA polymerase	40	21	14	0,029

continuação

				continuação
<i>KEGG pathway</i>	<i>Term Genes</i>	<i>Target Genes (n)</i>	<i>miRNAs (n)</i>	FDR
cAMP signaling pathway	226	93	16	0,030
Aminoacyl-tRNA biosynthesis	57	28	15	0,030
Mismatch repair	24	14	14	0,030
Gastric acid secretion	80	37	16	0,035
Phospholipase D signaling pathway	158	67	16	0,035
Adrenergic signaling in cardiomyocytes	161	68	16	0,037
Glutamatergic synapse	117	51	16	0,042
Basal cell carcinoma	66	31	16	0,044
VEGF signaling pathway	61	29	16	0,044
Arrhythmogenic right ventricular cardiomyopathy	79	36	16	0,047
Lysine degradation	69	32	15	0,048

Tabela S10 – Vias enriquecidas para DE-miRNAs *up* - LUAD-LUSC.

KEGG pathway	Term Genes	Target Genes (n)	miRNAs (n)	FDR
FoxO signaling pathway	139	39	3	p<0,01
Mitophagy - animal	76	27	3	p<0,01
Neurotrophin signaling pathway	124	35	2	p<0,01
Autophagy - animal	146	38	3	p<0,01
Bladder cancer	43	18	3	p<0,01
p53 signaling pathway	75	25	3	p<0,01
Ubiquitin mediated proteolysis	142	37	2	p<0,01
Thyroid hormone signaling pathway	137	36	3	p<0,01
Colorectal cancer	88	27	3	p<0,01
Prolactin signaling pathway	73	24	3	p<0,01
mRNA surveillance pathway	108	30	3	p<0,01
Cellular senescence	219	48	3	p<0,01
Kaposi sarcoma-associated herpesvirus infection	245	52	3	p<0,01
MAPK signaling pathway	329	64	3	p<0,01
Prostate cancer	101	28	3	p<0,01
Focal adhesion	213	46	3	p<0,01
Cell cycle	129	32	3	p<0,01
Endocrine resistance	118	30	3	p<0,01
Hepatitis C	173	39	3	p<0,01
Chronic myeloid leukemia	79	23	3	p<0,01
Fluid shear stress and atherosclerosis	149	35	3	p<0,01
Protein processing in endoplasmic reticulum	194	42	3	p<0,01
Bacterial invasion of epithelial cells	80	23	3	p<0,01
Renal cell carcinoma	70	21	2	p<0,01
Hippo signaling pathway	164	37	3	p<0,01
mTOR signaling pathway	177	39	2	p<0,01
Proteoglycans in cancer	220	45	3	p<0,01
Pathways in cancer	555	91	3	p<0,01
Sphingolipid signaling pathway	133	31	3	p<0,01
AGE-RAGE signaling pathway in diabetic complications	115	28	3	p<0,01
Shigellosis	268	51	3	p<0,01
Longevity regulating pathway	105	26	3	p<0,01
EGFR tyrosine kinase inhibitor resistance	82	22	2	p<0,01
AMPK signaling pathway	130	30	2	p<0,01
Salmonella infection	277	52	3	p<0,01
Endometrial cancer	61	18	3	p<0,01
Hepatitis B	177	37	3	p<0,01
Pancreatic cancer	78	21	3	p<0,01
Spinocerebellar ataxia	145	32	3	p<0,01
Endocytosis	311	56	3	0,001
TGF-beta signaling pathway	103	25	3	0,001
Wnt signaling pathway	173	36	2	0,001
Thyroid cancer	43	14	3	0,001
RNA degradation	101	24	3	0,001
RNA transport	199	39	3	0,001
Axon guidance	186	37	3	0,001
ErbB signaling pathway	86	21	2	0,002
Signaling pathways regulating pluripotency of stem cells	156	32	3	0,002
Amyotrophic lateral sclerosis	408	67	3	0,002
Hepatocellular carcinoma	177	35	3	0,002
Acute myeloid leukemia	69	18	3	0,002
PI3K-Akt signaling pathway	372	62	3	0,002
Apoptosis	151	31	3	0,002
Transcriptional misregulation in cancer	206	39	3	0,002
Measles	161	32	3	0,003
Gastric cancer	162	32	3	0,003
Adherens junction	79	19	3	0,003
Choline metabolism in cancer	106	23	2	0,004
Insulin signaling pathway	153	30	3	0,004
Non-small cell lung cancer	81	19	2	0,004
Platinum drug resistance	75	18	3	0,004
Regulation of actin cytoskeleton	224	40	3	0,005
Melanoma	76	18	3	0,005
Circadian rhythm	31	10	1	0,006

continuação				
KEGG pathway	Term Genes	Target Genes (n)	miRNAs (n)	FDR
Breast cancer	163	31	3	0,006
Hedgehog signaling pathway	59	15	3	0,006
Viral carcinogenesis	265	45	3	0,006
Thermogenesis	252	43	3	0,007
Glioma	79	18	3	0,007
Pathways of neurodegeneration - multiple diseases	539	80	3	0,008
Human papillomavirus infection	406	63	3	0,008
Oocyte meiosis	134	26	3	0,009
TNF signaling pathway	131	25	3	0,014
Huntington disease	339	53	3	0,015
Rap1 signaling pathway	214	36	3	0,019
Small cell lung cancer	100	20	3	0,020
Alzheimer disease	426	63	3	0,022
T cell receptor signaling pathway	115	22	3	0,022
Apoptosis - multiple species	32	9	3	0,023
C-type lectin receptor signaling pathway	116	22	3	0,024
Progesterone-mediated oocyte maturation	104	20	2	0,029
Parkinson disease	282	44	3	0,029
Yersinia infection	147	26	3	0,030
HIF-1 signaling pathway	112	21	3	0,031
Longevity regulating pathway - multiple species	79	16	3	0,036
Tight junction	182	30	2	0,045
JAK-STAT signaling pathway	168	28	3	0,048
Insulin resistance	124	22	3	0,048
Cushing syndrome	176	29	3	0,049