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

Identificação de microRNAs e Genes Alvo como Biomarcadores de Câncer de Cabeça e Pescoço sob uma Abordagem de Bioinformática

Saditt Rocio Robles Colonia

Botucatu
2024

Saditt Rocio Robles Colonia

Identificação de microRNAs e Genes Alvo como Biomarcadores de Câncer de Cabeça e Pescoço sob uma Abordagem de Bioinformática

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. Patrícia 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

Robles Colonia, Saditt Rocio.
Identificação de microRNAs e genes alvo como biomarcadores de câncer de cabeça e pescoço sob uma abordagem de bioinformática / Saditt Rocio Robles Colonia. - Botucatu, 2024

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Rogério Antonio de Oliveira
Coorientador: Patrícia Pintor dos Reis
Capes: 90194000

1. Árvores de decisão. 2. Análise de sobrevivência (Biometria). 3. Boca. 4. Laringe. 5. Marcadores bioquímicos.

Palavras-chave: Árvore de decisão; Análise de sobrevivência; Cavidade oral; Laringe; Redes de interação.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Botucatu

ATA DA DEFESA PÚBLICA DA TESE DE DOUTORADO DE SADITT ROCIO ROBLES COLONIA, 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 08:30 horas, no(a) Sala 7 da Central de Aulas 3, realizou-se a defesa de TESE DE DOUTORADO de SADITT ROCIO ROBLES COLONIA, intitulada **Identificação de microRNAs e genes alvo como biomarcadores de câncer de cabeça e pescoço sob uma abordagem de bioinformática**. A Comissão Examinadora foi constituida 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, Prof. Dr. ROBSON FRANCISCO CARVALHO (Participação Presencial) do(a) Departamento de Biologia Estrutural e Funcional / 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. Dr. ANGEL MAURICIO CASTRO GAMERO (Participação Virtual) do(a) Instituto de Ciências da Natureza (ICN) - Universidade Federal de Alfenas/MG, Profa. Dra. HILDETE PRISCO PINHEIRO (Participação Presencial) do(a) Departamento de Estatística / Instituto de Matemática, Estatística e Computação Científica - Universidade Estadual de Campinas. 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

Dedico este trabalho à minha família, com um carinho especial à minha mãe, cujo maior legado para mim foi a valiosa herança da educação.

Agradecimentos

Em primeiro lugar, desejo expressar minha gratidão a Deus, que, nos momentos mais desafiadores da minha vida, sustentou-me e atendeu aos anseios do meu coração.

Agradeço aos meus pais, Manuel e Esperanza, à minha irmã Elianne e aos meus tios, pela constante inspiração, confiança e apoio ao longo desta jornada acadêmica, oferecendo seu apoio e amor incondicional. A meu amigo eterno Alberto Frank Lázaro Aguirre, por sua amizade constante, apoio inabalável e parceria que mantivemos desde a graduação. Desde o momento em que nos conhecemos, compartilhamos o sonho de realizar a Pós-Graduação no Brasil, e é gratificante finalmente alcançá-lo juntos.

Ao meu orientador, Prof. Dr. Rogério Antonio de Oliveira, pela confiança depositada em mim, pela constante motivação proporcionada, pela notável paciência demonstrada e pelo valioso auxílio oferecido ao longo do meu percurso acadêmico. À minha coorientadora, Prof. Dra. Patrícia Pintor dos Reis, pelo direcionamento deste trabalho.

Aos membros da banca examinadora, Prof. Dr. Robson Francisco Carvalho, Profa. Dra. Silvia Helena Modenese Gorla da Silva, Profa. Dra. Hildete Prisco Pinheiro e ao Prof. Dr. Angel Mauricio Castro Gamero pelas valiosas contribuições neste trabalho.

Ao corpo docente do Programa de Pós-Graduação em Biometria, pela oportunidade de ampliar meus conhecimentos e vivenciar experiências enriquecedoras. Aos funcionários, em especial a Luiz, expresso minha sincera gratidão pela atenção e apoio oferecidos.

Quero estender meus agradecimentos especiais a Juliana Gualberto e seus pais, que não apenas me acolheram, mas me consideraram parte de sua família. A Bruno Novaes e sua esposa Renata, pela amizade sincera e pelo constante apoio que sempre me proporcionaram.

Por fim, agradeço a meus amigos Alberto Lázaro, Julián Lázaro, Luis Baselly, Juliana Gualberto, Mario Lucas, Lara Cardeal e Abrantes Mussafo, com os quais compartilhei momentos inesquecíveis. A presença e apoio de todos vocês tornaram esta jornada acadêmica ainda mais significativa e memorável.

Este trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Quem te disse que você não pode? Você não entende que Deus está contigo! Deus está ao nosso lado e somos mais que vencedores.

M.A.A. Fuego de Dios

Não basta fazer análises complicadas, os resultados devem ser úteis na prática e traduzidos para o pesquisador da outra o que os números que encontramos querem dizer.

Luzia Trinca

Resumo

O câncer de cabeça e pescoço (CCP) apresenta uma baixa taxa de sobrevivência, principalmente devido a diagnósticos tardios, falta de compreensão biológica e carência de biomarcadores confiáveis. Neste contexto, os microRNAs (miRNAs ou miRs) e seus alvos surgiram como biomarcadores relevantes para a formação, progressão e prognóstico de diversos tipos de cânceres, incluindo o CCP, sendo objeto de investigação constante. Este estudo teve como objetivo identificar miRNAs e seus genes alvo como potenciais biomarcadores para diagnóstico, prognóstico e tratamento do CCP, utilizando diversas ferramentas de bioinformática. A análise das expressões de miRNA-Seq e mRNA-Seq de pacientes com CCP foi conduzida com base nos dados do projeto Atlas do Genoma do Câncer (TCGA), juntamente com as informações clinicopatológicas correspondentes. As análises abrangeram a expressão diferencial, componentes principais, agrupamento, construção de redes de interação miRNA-mRNA, redes de co-expressão de proteínas, análises de enriquecimento funcional e classificação baseada em árvores, bem como análise de sobrevivência. No contexto da cavidade oral, identificaram-se 55 miRNAs expressos diferencialmente (DEmiRNAs) associados a 103 mRNAs expressos diferencialmente (DEmRNAs), e na laringe, 40 DEmiRNAs associados a 63 DEmRNAs. Alterações transcricionais foram observadas, com destaque para o gene COL1A1, regulado principalmente pelos miR-133a-3p, miR-29c-3p e miR-133b, indicando sua importância na digestão e absorção de proteínas. As árvores de decisão baseadas na classificação destacaram a relevância dos miR-101-3p, miR-455-3p, miR-99a-5p e miR-30a-5p como potenciais biomarcadores no diagnóstico do câncer de cavidade oral, distinguindo o tumor primário do tecido normal adjacente. No câncer de laringe, miR-338-5p, miR-1910-5p, miR-99a-5p, miR-196b-5p e miR-139-3p emergiram como candidatos a biomarcadores independentemente do estadiamento da doença. A análise de riscos proporcionais de Cox univariada identificou cinco miRNAs e uma assinatura combinada de 24 miRNAs associadas à sobrevivência dos pacientes com câncer da cavidade oral. Para o câncer de laringe, nove miRNAs e uma assinatura combinada de 16 miRNAs foram associadas à sobrevivência dos pacientes. Esses resultados destacam o potencial desses miRNAs e seus aalvos como biomarcadores no diagnóstico, prognóstico e compreensão do processo biológico do CCP.

Palavras-chave: Cavidade oral, Laringe, Redes de interação, Árvore de decisão, Análise de sobrevivência.

Abstract

Head and neck cancer (HNC) presents a low survival rate, mainly due to late diagnosis, lack of biological understanding, and a shortage of reliable biomarkers. In this context, microRNAs (miRNAs or miRs) and their targets have emerged as relevant biomarkers for the formation, progression, and prognosis of various types of cancers, including HNC, and have been the subject of constant investigation. This study aimed to identify miRNAs and their target genes as potential biomarkers for the diagnosis, prognosis, and treatment of HNC, using various bioinformatics tools. Analysis of miRNA-Seq and mRNA-Seq expressions in HNC patients was conducted based on data from The Cancer Genome Atlas (TCGA) project, along with corresponding clinicopathological information. The analyses included differential expression, principal components, clustering, construction of miRNA-mRNA interaction networks, protein co-expression networks, functional enrichment analyses, tree-based classification, as well as survival analysis. In the context of the oral cavity, 55 differentially expressed miRNAs (DEmiRNAs) associated with 103 differentially expressed mRNAs (DEmRNAs) were identified, and in the larynx, 40 DEmiRNAs associated with 63 DEmRNAs were identified. Transcriptional alterations were observed, with emphasis on the COL1A1 gene, regulated mainly by miR-133a-3p, miR-29c-3p, and miR-133b, indicating its importance in protein digestion and absorption. Decision trees based on classification highlighted the relevance of miR-101-3p, miR-455-3p, miR-99a-5p, and miR-30a-5p as potential biomarkers in the diagnosis of oral cavity cancer, distinguishing the primary tumor from adjacent normal tissue. In laryngeal cancer, miR-338-5p, miR-1910-5p, miR-99a-5p, miR-196b-5p, and miR-139-3p emerged as biomarker candidates regardless of disease staging. Univariate Cox proportional hazards analysis identified five miRNAs and a combined signature of 24 miRNAs associated with the survival of oral cavity cancer patients. For laryngeal cancer, nine miRNAs and a combined signature of 16 miRNAs were associated with patient survival. These results highlight the potential of these miRNAs and their targets as biomarkers in the diagnosis, prognosis, and understanding of the biological process of HNC.

Keywords: Oral cavity, Larynx, Interaction networks, Decision tree, Survival analysis.

Lista de figuras

Figura 1 – Etapas da carcinogênese: iniciação, promoção, progressão e metástase. Figura criada em <Biorender.com>. 4
Figura 2 – Ilustração da heterogeneidade tumoral (intertumoral e intratumoral). Figura criada em <Biorender.com>. 5
Figura 3 – Ilustração da anatomia da cabeça e pescoço. Figura criada em <Biorender.com>. 7
Figura 4 – Biogênese e mecanismo dos miRNAs. Figura criada em <Biorender.com>. 9
Figura 5 – Nomenclatura dos miRNAs coletados no miRBase. 10
Figura 6 – Ilustração da árvore de decisão. 27
Figura 7 – Ilustração da floresta aleatória. Figura adaptada de Khan *et al.* (2021). . . . 29
Figura 8 – Ilustração da curva ROC juntamente com a Área Sob a Curva ROC. 32
Figura 9 – Ilustração esquemática dos métodos de validação. (A) Método Holdout e (B) validação cruzada *k-fold*. 33
Figura 10 – Fluxograma das etapas utilizadas no estudo. 39
Figura 11 – (A) Gráfico de barras referentes ao tecido normal adjacente e tumor primário nos sítios anatômicos do CCP. (B) Gráfico de barras referentes ao órgão de origem do câncer da cavidade oral. 53
Figura 12 – Gráficos bivariados das características clinicopatológicas. (A) *Boxplots* das idades (anos) dos pacientes por estadiamento patológico. (B) Histograma das idades dos pacientes estratificado por sexo. (C) Histograma dos tempos de censura e falha. (D) Histograma das idades (anos) por estado vital dos pacientes nos sítios anatômicos do câncer de cabeça e pescoço. 54
Figura 13 – Distribuição da expressão de oito miRNAs e oito mRNAs nos sítios anatômicos da cavidade oral (A) e laringe (B), respetivamente. 55
Figura 14 – *Boxplots* da expressão dos miRNAs para os primeiros 112 amostras da cavidade oral (a, c, e) e laringe (b, d, f), respectivamente. 56
Figura 15 – Expressão Diferencial de miRNAs em comparação entre tecidos tumorais primários e normais adjacentes relacionados ao câncer de cavidade oral. (A) O gráfico MA da razão logarítmica (M) versus a média (A) das diferenças entre os dois grupos para cada miRNA. (B) Gráfico de Vulcão dos DEmiRNAs. (C) Gráfico de Barras com os 10 DEmiRNAs mais desregulados. (D) Gráfico de pirulitos divergentes com os 10 DEmiRNAs em relação ao risco relativo. 58
Figura 16 – Gráfico de barras dos 10 DEmiRNAs mais desregulados no câncer da cavidade oral pelo método de edgeR (A) e o teste de Wilcoxon-Mann-Whitney (B), respectivamente. (C) Diagrama de Venn mostrando a sobreposição dos resultados obtidos pelos três métodos (DESeq2, edgeR e o teste de Wilcoxon-Mann-Whitney). 59

Figura 17 – Gráficos de violino ilustrando a expressão dos cinco principais DEmiRNAs superexpressos (A) e dos cinco DEmiRNAs subexpressos (B) ao comparar o tumor primário com o tecido normal adjacente ao tumor, relacionados ao câncer na cavidade oral. 60
Figura 18 – Análise de componentes principais com base na expressão dos DEmiRNA de cada amostra associados ao câncer da cavidade oral. (A) *Scree plot* dos autovalores para cada componente principal. (B) *Scree plot* da variação explicada dos componentes principais. (C) Gráficos de barras para duas dimensões com base na contribuição dos miRNAs. (D) PCA mostra o padrão de expressão dos miRNAs na estratificação do tumor primário e tecido adjacente. 61
Figura 19 – Mapa de calor de agrupamento hierárquico da expressão de miRNA entre as amostras dos dois grupos (tecido tumoral e normal adjacente ao tumor) no câncer da cavidade oral usando a distância Euclideana e o método de ligação completa. 62
Figura 20 – Dendrograma circular hierárquico da separação entre grupos de DEmiRNAs superexpressos e subexpressos no câncer da cavidade oral usando a distância Euclideana e o método de ligação completa. 63
Figura 21 – Matriz de correlação linear de Spearman entre os DEmiRNAs associados ao câncer da cavidade oral. 64
Figura 22 – DEmiRNAs na comparação do tumor primário e normal adjacente relacionados ao câncer da laringe. (A) Gráfico MA da razão logarítmica (M) versus a média (A) das diferenças entre os dois grupos para cada miRNA. (B) Gráfico de Vulcão dos DEmiRNAs. (C) Gráfico de Barras com os 10 DEmiRNAs mais desregulados. (D) Gráfico de pirulitos divergentes com os 10 DEmiRNAs em relação ao risco relativo. 65
Figura 23 – Gráficos de violino da expressão dos 10 primeiros DEmiRNAs (superexpresso (A) e subexpresso (B)) na comparação do tumor primário com o tecido normal adjacente ao tumor associados ao câncer da laringe. 66
Figura 24 – Gráfico de Barras dos 10 DEmiRNAs mais desregulados no câncer da laringe pelo método de edgeR (A) e Wilcoxon-Mann-Whitney (B), respectivamente. (C) Diagrama de Venn da sobreposição dos resultados obtidos pelos três métodos. 67
Figura 25 – Análise de componentes principais com base na expressão dos DEmiRNA de cada amostra associados ao câncer da laringe. (A) *Scree-plot* dos autovalores para cada componente principal. (B) *Scree-plot* da variação explicada dos componentes principais. (C) Gráficos de barras para duas dimensões com base na contribuição dos miRNAs. (D) PCA mostra o padrão de expressão dos miRNAs na estratificação do tumor primário e tecido normal adjacente. 68

Figura 26 – Mapa de calor de agrupamento hierárquico da expressão de miRNA entre as amostras dos dois grupos (tumor primário e normal adjacente ao tumor) no câncer da laringe oral usando a distância Euclideana e o método de ligação completa. 69
Figura 27 – Dendrograma circular hierárquico da separação entre grupos de DEmiRNAs superexpressos e subexpressos no câncer da laringe usando a distância Euclideana e o método de ligação completa. 70
Figura 28 – Matriz de correlação linear de Spearman entre DEmiRNAs no câncer da laringe. 71
Figura 29 – Diagrama de Venn dos DEmiRNAs validados com outras fontes externas para câncer da cavidade oral (A), laringe (B), tipos de cânceres (C) e câncer de cabeça e pescoço (D), respectivamente. 73
Figura 30 – Diagrama de Venn mostra a sobreposição entre (A) Vias KEGG e (B) Anotação GO em câncer da cavidade oral e laringe, respectivamente. 74
Figura 31 – Gráficos de bolhas mostrando as trinta principais vias KEGG e os termos GO associados aos miRNAs desregulados em pacientes com câncer oral (A, B) e laringe (C, D), respectivamente. 75
Figura 32 – Diagrama de Venn mostra a sobreposição dos DEmiRNAs entre o câncer da cavidade oral e da laringe. 76
Figura 33 – Análise de Enriquecimento de vias KEGG em DEmiRNAs sobrepostos entre o câncer da cavidade oral e da laringe. 76
Figura 34 – (A, B) Diagrama de Venn mostrando a sobreposição dos DEmiRNAs entre estadiamento inicial e avançado do câncer da cavidade oral e laringe, respectivamente. (C) Gráficos de violino da expressão dos cinco DEmiRNAs nos estadiamentos avançados do câncer de cavidade oral. (D) Gráfico de violino da expressão dos DEmiRNAs no estadiamento inicial e avançado do câncer de laringe. 78
Figura 35 – *MA-Plot* (A) e *volcano-Plot* (B) dos DEmRNAs entre o tumor primário e tecido de normal adjacente relacionados ao câncer da cavidade oral (A, B) e laringe (C, D), respectivamente. 79
Figura 36 – Diagrama de Venn mostrando a sobreposição entre DEmRNAs e genes alvo dos DEmiRNAs (mirDIP e miRTarBase) associados ao câncer da cavidade oral (A) e da laringe (B), respectivamente. 80
Figura 37 – (A, C) Gráfico de barras dos 10 DEmRNAs mais desregulados (valores FDR muito baixos) e influenciado pelo los valores $\log_2(FC)$ para o câncer da cavidade oral e laringe, respectivamente. (B, C) Análise de componentes principais dos DEmRNAs relacionados ao câncer da cavidade oral e laringe. 81

Figura 38 – *Heatmap* e dendrograma circular dos DEmRNAs entre as amostras dos dois grupos (tecido do tumor primário e normal adjacente) no câncer da cavidade oral (A,B) e laringe (C, D) usando a distância Euclideana e o método de ligação completa para as colunas (amostras) e Ward.D2 para as linhas (DEmRNAs). A seta vermelha em (B) significa que o gene ADD2 é superexpresso e que foi agrupado incorretamente. 82
Figura 39 – Rede regulatória entre DEmiRNAs e DEmRNAs associadas ao câncer da cavidade oral (A) e laringe (B), respectivamente. 83
Figura 40 – Rede regulatória entre DEmiRNAs e DEmRNAs do estadiamento avançado do câncer da cavidade oral (A), inicial e avançado do câncer da laringe (B), respectivamente. 84
Figura 41 – Anotação funcional e análise de enriquecimento dos DEmRNAs. (A, B, C, D) Rede interativa de análise de enriquecimento e gráfico de bolhas exibindo as vias KEGG dos DEmRNAs no câncer da cavidade oral e laringe, respectivamente.(E, F) Gráfico de bolhas de enriquecimento do termo Ontologia Genética para cada um dos termos de instância (processo biológico, componente celular e função molecular) no câncer da cavidade oral e laringe, respectivamente. (G, H) Gráfico de barras da análise de enriquecimento da Ontologia de Doenças no câncer da cavidade oral e laringe, respectivamente. 87
Figura 42 – (A, E) Rede de interação proteína-proteína (PPI) associada ao câncer da cavidade oral e laringe, respectivamente. (B, F) *Hub* genes referentes ao câncer da cavidade oral e laringe, respectivamente. A cor vermelha a amarela representa o grau de conectividade de cima para baixo baseada no algoritmo de *Maximal Clique Centrality* (MCC). (C, G) Rede regulatória entre DEmiRNAs e *hub* genes referentes ao câncer da cavidade oral e laringe, respectivamente. (D, H) Enriquecimento da via KEGG dos *hubs* referentes ao câncer da cavidade oral e laringe, respectivamente. 90
Figura 43 – *Heatmap* baseada na árvore de decisão para classificação de tecidos do câncer da cavidade oral, considerando todos os estadiamentos patológicos. As cores do mapa apresentam o valor relativo da expressão dos miRNAs em cada amostra. 92
Figura 44 – Classificação dos tecidos utilizando CTree e floresta aleatória, abrangendo todos os estadiamentos patológicos (A, B), iniciais (C, D) e avançados (E, F) do câncer da cavidade oral. 95
Figura 45 – *Heatmap* baseada na árvore de decisão para classificação de tecidos do câncer da laringe, considerando todos os estadiamentos patológicos. As cores do mapa apresentam o valor relativo da expressão dos miRNAs em cada amostra. 96
Figura 46 – Classificação dos tecidos utilizando CTree e floresta aleatória, abrangendo todos os estadiamentos patológicos (A, B) e avançados (C, D) do câncer da laringe. 98

Figura 47 – Curvas de sobrevivência estimada pelo método de Kaplan-Meier de acordo com o câncer da cavidade oral e laringe. 99
Figura 48 – Curva de sobrevivência estimada pelo método de Kaplan-Meier, teste de Log-Rank e modelo de Cox univariado estratificado por estadiamento patológico no conjunto treinamento do câncer da cavidade oral. 102
Figura 49 – Curvas de sobrevivência estimada pelo método de Kaplan-Meier, teste de Log-Rank e modelo de Cox univariado de acordo com a expressão de miRNAs em diferentes estadiamentos patológicos (A) inicial (B) e avançado (C) do câncer da cavidade oral no conjunto de treinamento. 104
Figura 50 – Gráfico dos riscos do modelo multivariado de riscos proporcionais de Cox dos miRNAs associados com a sobrevivência dos pacientes com câncer oral considerando todos os estadiamentos (I - IV) e avançado (III - IV) no conjunto treinamento. Cada miRNA encontra-se categorizada (1: expressão aumentada e 0: expressão diminuída). 106
Figura 51 – Árvore de sobrevivência podada obtida pelo CART© e gráfico de barras da importância de miRNAs obtidos pela floresta aleatória na previsão da sobrevivência dos pacientes considerando todos os estadiamentos (A, B), inicial (C, D) e avançado (E, F) do câncer da cavidade oral no conjunto treinamento. 108
Figura 52 – Curvas de sobrevivências estimada pelo método de Kaplan-Meier para sobrevivência dos pacientes estratificado por estadiamento patológico no câncer da laringe. 109
Figura 53 – Curvas de sobrevivência estimada pelo método de Kaplan-Meier, teste de Log-Rank e modelo de Cox univariado para sobrevivência dos pacientes de acordo com a expressão dos miRNAs considerando todos os estadiamentos (A) e o estadiamento avançado (B) do câncer da laringe. 112
Figura 54 – Gráfico de riscos do modelo multivariado de riscos proporcionais de Cox dos miRNAs associados com a sobrevivência dos pacientes considerando todos os estadiamentos do câncer da laringe. Cada miRNA encontra-se categorizada (1: expressão aumentada e 0: expressão diminuída). 112
Figura 55 – Árvore de sobrevivência podada obtido pelo CART© e fator de importância de miRNAs obtidas via floresta aleatória na previsão da sobrevivência dos pacientes considerando todos os estadiamentos (A, B) e avançado (C, D) no conjunto treinamento do câncer da laringe, respectivamente. 113

Lista de tabelas

Tabela 1 – Tipos de erros de um teste de hipóteses. 19
Tabela 2 – Erros em contrastes de hipóteses múltiplas. 20
Tabela 3 – Matriz de confusão comparando valores preditos e valores observados. . . . 30
Tabela 4 – Métricas de desempenho na classificação. 31
Tabela 5 – Descrição das características clinicopatológicas dos 528 pacientes diagnosticados com câncer de cabeça e pescoço. 40
Tabela 6 – Matriz de expressão do miRNA-Seq. Cada linha corresponde a um miRNA específico e cada coluna uma amostra referente ao tumor primário (exemplo, TCGA-4P-AA8J-01) e tecido sólido normal ou adjacente ao tumor (exemplo, TCGA-CV-6933-11) com seus respectivos níveis de expressão associados ao câncer de cabeça e pescoço. 41
Tabela 7 – Matriz de expressão de genes transcritos codificadores (mRNA-Seq). Cada linha corresponde a um mRNA específico e cada coluna uma amostra referente ao tumor primário (exemplo, TCGA-4P-AA8J-01) e tecido histologicamente normal adjacente ao tumor (exemplo, TCGA-H7-A6C5-11) com seus respectivos níveis de expressão associados ao câncer de cabeça e pescoço. . 41
Tabela 8 – Conjuntos de dados destinados à validação dos DEmiRNAs presentes no banco de dados dbDMEC (versão 3.0) (XU *et al.*, 2022). 46
Tabela 9 – Frequência absoluta (f_a) e percentual (f_p) das características clinicopatológicas dos pacientes com câncer de cabeça e pescoço (CCP) em cada sítio anatômico da doença. 52
Tabela 10 – Análise das curvas de ROC na discriminação do tumor primário e do tecido normal adjacente ao tumor no câncer da cavidade oral e laringe. 91
Tabela 11 – Matriz de confusão e avaliação de desempenho dos modelos baseados nas árvores de decisão (CTree e floresta aleatória) utilizando *Five-Fold Cross-Validation* na classificação binária de tumor primário (1) e tecido normal adjacente (0) em relação ao câncer de cavidade oral com diferentes técnicas de balanceamento. 93
Tabela 12 – Matriz de confusão e avaliação de desempenho dos modelos baseados nas árvores de decisão (CTree e floresta aleatória) utilizando *Five-Fold Cross-Validation* na classificação binária de tumor primário (1) e tecido normal adjacente (0) em relação ao câncer da laringe com diferentes técnicas de balanceamento. 97
Tabela 13 – Comparação do desempenho de modelos multivariados de Cox e árvores de sobrevivência e floresta de sobrevivência aleatória no câncer da cavidade oral e laringe, considerando os estadiamentos da doença. 100

Tabela 14 – Estimativas da regressão de Cox univariado e multivariado das caraterísticas clinicopatológicas do câncer da cavidade oral no conjunto treinamento. . . 101
Tabela 15 – Estimativas de regressão de Cox univariada e multivariada das caraterísticas clinicopatológicas do câncer da laringe. 109
Tabela 16 – Distribuição das características clínico-patológicas em relação ao tecido normal e tumoral nos sítios anatômicos do CCP. 142
Tabela 17 – Estimativas dos parâmetros do modelo Binomial Negativa da expressão dos DEmiRNAs na comparação dos tecidos (tumoral vs. normal) na cavidade oral. 143
Tabela 18 – Estimativas dos coeficientes do modelo Binomial Negativa da expressão dos DEmiRNAs em amostras de tecido tumoral primário em comparação com os tecidos normais adjacentes ao tumor dos pacientes com câncer da laringe. . 144
Tabela 19 – Estimativas dos coeficientes do modelo pelo edgeR em amostras de tecido tumoral primário em comparação com os tecidos normais adjacentes ao tumor dos pacientes com câncer da laringe. 145
Tabela 20 – Estatísticas descritivas da expressão dos miRNAs identificados na cavidade oral comparando os tecidos usando o teste de Wilcoxon-Mann-Whitney. . . 146
Tabela 21 – Estatísticas descritivas da expressão dos miRNAs identificados na laringe comparando os tecidos (tumor primário vs. normal adjacente) usando o teste de Wilcoxon-Mann-Whitney. 147
Tabela 22 – Estimativas dos coeficientes do modelo pelo Deseq2 em amostras de tecido tumoral primário em comparação com os tecidos normais adjacentes ao tumor dos pacientes com câncer da laringe. 148
Tabela 23 – DEmiRNAs identificados com o teste WMW nos estadiamento iniciais e avançados e áreas sob a curva (AUC), especificidade (TNR) e sensibilidade (TPR) na capacidade de discriminação do tumor primário (T) e tecido normal adjacente ao tumor (NAT) na cavidade oral. 149
Tabela 24 – DEmiRNAs identificados com o teste WMW nos estadiamentos iniciais e avançados e áreas sob a curva (AUC), especificidade (TNR) e sensibilidade (TPR) na capacidade de discriminação do tumor primário (T) e tecido normal adjacente ao tumor (NAT) na laringe. 150

Lista de abreviaturas e siglas

DNA	Ácido Desoxirribonucleico
RNA	Ácido ribonucleico
RNA-Seq	Sequenciamento de RNA
ncRNA	RNA não codificante
miRNA	microRNA
miRNA-Seq	Sequenciamento de microRNA
mRNA	RNA mensageiro
mRNA-Seq	Sequenciamento de mRNA
NGS	Sequenciamento da Próxima Geração
AJCC	*American Joint Committee on Cancer*
UTCC	*Union for International Cancer Control*
TNM	Tumor, linfonodos, metástase
TCGA	Genoma Atlas do Câncer
mirDIP	*microRNA Data Integration Portal*
CCP	Câncer de cabeça e pescoço
CCP	Carcinoma de células escamosas de cabeça e pescoço
CCE	Carcinoma de células escamosas
CCEF	Carcinoma de células escamosas fusiformes
CCEB	Carcinoma de células escamosas basalóide
CCEQ	Carcinoma de células escamosas queratinizante
CCENQ	Carcinoma de células escamosas não queratinizado
CART©	Árvores de Classificação e Regressão
MLG	Modelos Lineares Generalizados

Sumário

1 INTRODUÇÃO 1
1.1 Objetivos 2
1.1.1 Objetivo geral 2
1.1.2 Objetivos específicos 3

2 ASPECTOS BIOLÓGICOS DO CÂNCER DE CABEÇA E PESCOÇO E SUA RELAÇÃO COM O TRANSCRIPTOMA 4
2.1 Biologia do câncer 4
2.2 Cânceres de cabeça e pescoço 6
2.3 MiRNAs na regulação da expressão gênica 8
2.4 MiRNAs e genes envolvidos no câncer de cabeça e pescoço 10

3 FERRAMENTAS DA BIOINFORMÁTICA APLICADA À GENÔMICA 12
3.1 Análise de expressão diferencial e predição de alvos 12
3.2 Redes regulatórias e análise de enriquecimento funcional 12

4 MÉTODOS ESTATÍSTICOS APLICADOS À EXPRESSÃO GÊNICA . 14
4.1 Modelos Lineares Generalizados 14
4.1.1 Modelo de regressão de Poisson 15
4.1.2 Modelo de regressão Binomial Negativa 15
4.2 Teste de múltiplas hipóteses 19
4.3 Testes não paramétricos 21
4.3.1 Teste não paramétrico de Wilcoxon-Mann-Whitney (WMW) 21
4.3.2 Teste Qui-quadrado de Pearson 22
4.4 Análise multivariada dos dados 22
4.4.1 Análise de componentes principais 23
4.4.2 Análise de agrupamento – *Clustering* 25
4.5 Árvores de decisão e floresta aleatória 26
4.5.1 Técnicas de reamostragem para dados desbalanceados 30
4.5.2 Métricas de desempenho das árvores de decisão 30
4.6 Análise de sobrevivência 33
4.6.1 Estimador de Kaplan-Meier 35
4.6.2 Modelo de riscos proporcionais de Cox 35
4.6.3 Árvores de regressão para dados censurados 37
4.6.4 Métrica de desempenho para dados censurados 38

5 BANCO DE DADOS E MÉTODOS DE ANÁLISE 39
5.1 Conjunto de dados do TCGA 40
5.2 Análise de expressão diferencial dos miRNAs e mRNAs 42
5.3 Análise de componentes principais e *Clustering* 44
5.4 Validação da expressão dos miRNAs com fontes externas 45
5.5 Análise de enriquecimento e anotação funcional de miRNAs 46
5.6 Predição de alvos, redes de interação miRNA-mRNA e análises de enriquecimento GO e KEGG de mRNAs 46
5.7 Rede de interação proteína-proteína, identificação de *hub* genes, rede regulatória miRNA-*hub* e análise de enriquecimento de vias KEGG . . . 47
5.8 Algoritmos de classificação baseadas em árvores de decisão e florestas aleatórias 48
5.9 Análise de sobrevivência 49

6 RESULTADOS 51
6.1 Análise preliminar dos dados 51
6.1.1 Características clinicopatológicas 51
6.1.2 Perfil de expressão de miRNAs e mRNAs 55
6.2 Análise da expressão diferencial dos miRNAs 57
6.2.1 Identificação de DEmiRNAs no câncer da cavidade oral 57
6.2.2 Identificação de DEmiRNAs no câncer da laringe 64
6.2.3 Validação de miRNAs candidatos com outras fontes de dados 71
6.2.4 Análise de enriquecimento e anotação ontológica de miRNAs 74
6.2.5 Identificação de DEmiRNAs comumente expressos no câncer da cavidade oral e laringe 75
6.2.6 Identificação de DEmiRNAs associados ao estadiamento do câncer da cavidade oral e laringe 77
6.3 Identificação DEmRNAs, predição de alvos, redes de interação miRNA-mRNA, redes de proteína-proteína e análise funcional 79
6.3.1 Rede de interação regulatória miRNA-mRNA 83
6.3.2 Análises de enriquecimento e de vias dos DEmRNAs 85
6.3.3 Rede de interação proteína-proteína e redes regulatórias miRNA-*hub* . . 88
6.4 Discriminação de tecidos por meio de curvas ROC: avaliando a expressão individual de miRNAs 91
6.4.1 Classificadores baseados em árvore de decisão no diagnóstico de câncer da cavidade oral e laringe usando a expressão de miRNAs 92
6.5 MiRNAs como biomarcadores de sobrevivência para câncer da cavidade oral e laringe 99
6.5.1 MiRNAs como preditores de sobrevivência em pacientes com câncer da cavidade oral 101

6.5.2 MiRNAs como preditores de sobrevivência em pacientes com câncer da laringe . . . 109

7 DISCUSSÃO . . . 114

8 CONCLUSÃO . . . 120
8.1 Limitações do estudo . . . 121
8.2 Trabalhos futuros . . . 121

REFERÊNCIAS . . . 123

APÊNDICES 141

APÊNDICE A VARIÁVEIS CLÍNICAS, MIRNA-SEQ E MRNA-SEQ . . . 142
A.1 Análise bivariada das características clinicopatológicas . . . 142
A.2 Identificação de DEmiRNAs pelo DESeq2: oral e laringe . . . 143
A.3 Identificação de DEmiRNAs com edgeR . . . 145
A.4 Identificação de DEmiRNAs pelo WMW: oral e laringe . . . 146
A.5 Identificação de DEmRNAs com DESeq2: oral e laringe . . . 148
A.6 Identificação dos DEmiRNAs nos estadiamento iniciais e avançados do câncer da cavidade oral e laringe . . . 149

1 Introdução

O câncer de cabeça e pescoço (CCP), também conhecido como carcinoma de células escamosas de cabeça e pescoço (CCECP), é considerado um dos tipos de câncer mais agressivos, e complexo no diagnóstico, pois frequentemente é detectado em estadiamentos avançados (MODY *et al.*, 2021; INCA, 2022). Esta doença está associada a altas taxas de incidência e mortalidade, levando a óbito mais de 400 mil pessoas anualmente (SUNG *et al.*, 2021) e prevê-se que até 2030, a incidência possa aumentar em até 30% (JOHNSON *et al.*, 2020). A maioria dos diagnósticos de CCP ocorre nos sítios anatômicos da cavidade oral e laringe, principalmente em homens com mais de 40 anos e em grande parte dos casos, associado ao consumo prolongado de tabaco e a injesta do álcool (NETWORK *et al.*, 2015; SHAW; BEASLEY, 2016; GORMLEY *et al.*, 2022). No entanto, apesar dos avanços nas estratégias terapêuticas, a taxa de sobrevivência dos pacientes com CCP ainda é baixa, devido ao insucesso terapêutico que muitas vezes decorre de diagnósticos tardios, compreensão biológica limitada, dificuldade na avaliação da progressão do cancro devido à localização anatômica, diversidade clinicopatológica, heterogeneidade observada na resposta ao tratamento e à falta de biomarcadores, como os miRNAs clinicamente validados para o diagnóstico precoce (ALSAHAFI *et al.*, 2019; PASI *et al.*, 2020; VAHABI; BLANDINO; AGOSTINO, 2021).

Atualmente, com o avanço das técnicas de sequenciamento de próxima geração (NGS)e o estabelecimento de repositórios de dados públicos, como o *Genomic Data Commons* (GDC), cBioPortal, *UCSC Xena*, *Human Protein Atlas*, *Pan-Cancer Atlas* e *Gene Expression Omnibus* (GEO), têm aumentado significativamente a quantidade de dados genômicos, transcriptômicos e clínicos de diversos tipos de câncer (NETWORK *et al.*, 2015). Como consequência, métodos computacionais e estatísticos, como aprendizado de máquina para mineração e modelagem preditiva generalizável, têm sido amplamente adotados na bioinformática para identificar e caraterizar o RNA-Seq para o diagnóstico e prognóstico do câncer. Além disso, as ferramentas de análise de enriquecimento funcional estão se tornando cada vez mais importantes na decodificação da expressão dos miRNAs e mRNAs que estão relacionados aos processos biológicos e vias de sinalização (CHEN *et al.*, 2019; GARCIA-MORENO; CARMONA-SAEZ, 2020; BUITRAGO; LOVERING; CAPORALI, 2023).

Geralmente, em estudos da expressão do RNA-Seq, empregam-se métodos de expressão diferencial para reduzir a complexidade dos dados e identificar genes expressos diferencialmente

no câncer (MCDERMAID *et al.*, 2019). Esses métodos são complementados por métodos não supervisionados, como técnicas de agrupamento e análise de componentes principais (PCA), para identificar padrões na expressão gênica. Além disso, métodos supervisionados, como regressão linear, regressão logística, árvores de decisão, floresta aleatória, maquinas de vectores de suporte (SVM), redes neuronais, classificador Naive Bayes, k-vizinhos mais próximos, redes Bayesianas, *Lasso*, *Ridge* entre outras são utilizados para construir modelos preditivos. No entanto, as árvores de decisão são um diferencial em relação aos outros algoritmos, pois não requerem especificações de parâmetros complexos. Outro método amplamente utilizado é a análise de sobrevivência que visa examinar o tempo até a ocorrência de um evento de interesse relacionado à doença, sendo complementada por uma série de métodos estatísticos e modelos mais complexos de aprendizado de máquina, como a floresta de sobrevivência aleatória (RSF), utilizado para construir o modelo de sobrevivência com dados genômicos de alta dimensão (MIRZA *et al.*, 2019; LEE; LIM, 2019).

Alguns estudos bem-sucedidos na aplicação desses métodos no diagnóstico e prognóstico do CCP, utilizando a expressão de miRNAs, foram publicados em Nair *et al.* (2015), Cecco *et al.* (2017), Lopez *et al.* (2018), Lu *et al.* (2019), Kumar, Patnaik e Dixit (2020), Panvongsa *et al.* (2021), Piotrowski *et al.* (2021) e Qi *et al.* (2021). No entanto, mesmo que existam estudos promissores sobre miRNAs e mRNAs no CCP, ainda não foram identificados biomarcadores específicos que possam ser utilizados na prática clinica devido à complexidade anatômica desta doença.

Desse modo, verifica-se que existe ainda uma grande necessidade de adotar abordagens integradas de bioinformática para atingir uma compreensão mais abrangente e detalhada da complexidade biológica dos miRNAs e mRNAs associadas ao CCP, o que poderia resultar em avanços significativos no diagnóstico precoce, prognósticos mais precisos e no desenvolvimento de tratamentos personalizados mais eficazes e menos prejudiciais, proporcionando qualidade de vida e possibilidade de cura para os pacientes, classificados com base em sua localização anatômica.

1.1 Objetivos

A seguir são apresentados os objetivos geral e específicos do presente trabalho de pesquisa.

1.1.1 Objetivo geral

Identificar miRNAs e genes alvo que poderiam ser potenciais biomarcadores no diagnóstico, prognóstico para os cânceres de cabeça e pescoço, por meio da aplicação de diferentes ferramentas de bioinformática.

1.1.2 Objetivos específicos

- Descrever as características clinicopatológicas, expressão de miRNAs e mRNAs de pacientes diagnosticados com câncer de cabeça e pescoço;
- Identificar miRNAs e mRNAs diferencialmente expressos no câncer da cavidade oral e laringe usando três métodos diferentes (DESeq2, edgeR e teste de Wilcoxon-Mann-Whitney) e, validar os resultados com outras fontes de dados;
- Identificar miRNAs e seus genes alvo nos estadiamentos iniciais e avançados do câncer da cavidade oral e laringe;
- Aplicar a análise de componentes principais e clusterização hierárquica considerando fatores clinicopatológicos dos pacientes com câncer de cavidade oral e laringe;
- Realizar a análise de enriquecimento funcional das vias KEGG e anotação GO dos miRNAs diferencialmente expressos no câncer da cavidade oral e laringe;
- Predizer os genes candidatos que são alvos de miRNAs com expressão diferencial, criar redes regulatórias, realizar o enriquecimento funcional dos genes alvo e construir redes de interação proteína-proteína no câncer da cavidade oral e laringe;
- Identificar genes centrais (*hub genes*), estabelecer a rede regulatória miRNA-*hub* genes e realizar a análise de enriquecimento de vias KEGG;
- Identificar biomarcadores de miRNAs para o diagnóstico de câncer na cavidade oral e na laringe em diferentes estadiamentos da doença e;
- Identificar os fatores clinicopatológicas e miRNAs associados ao tempo de sobrevivência em pacientes com câncer da cavidade oral e laringe.

Nesse sentido, esta tese está organizada da seguinte forma: O Capítulo 1 contém a introdução e os objetivos gerais e específicos do estudo. A visão geral do CCP e sua relação com os transcritos (miRNAs e mRNAs) é apresentada no Capítulo 2. O Capítulo 3, detalha as ferramentas de bioinformática. No Capítulo 4, aborda-se os métodos estatísticos aplicados à expressão gênica. No Capítulo 5, apresentam-se a descrição do banco de dados e os métodos de bioinformática utilizados na análise dos dados. No Capítulo 6, apresentam-se os resultados do estudo. Finalmente, nos Capítulos 7 e 8, são apresentadas as discussões e conclusões, mostrando as limitações do estudo e trabalhos futuros, respectivamente.

2 Aspectos biológicos do câncer de cabeça e pescoço e sua relação com o transcriptoma

2.1 Biologia do câncer

O câncer é um grupo complexo e heterogêneo de mais de 200 tipos de patologias genéticas multifatoriais que apresentam diferenças em sua histologia, morfologia, evolução clínica e prognóstico (PECORINO, 2021). No entanto, todas elas se caracterizam pelo crescimento descontrolado das células (oncogenes), que invadem e se espalham a partir do local de origem, para outros locais do corpo devido a fatores físicos, químicos e biológicos, tais como tabagismo, álcool, radiação solar ultravioleta, poluição, vírus, dieta, entre outros (PECORINO, 2021; HESKETH, 2023). Alguns tipos de câncer se disseminam lentamente e apresentam um desenvolvimento gradual, enquanto outros são rápidos e metastáticos (KLEIN; GRANDIS, 2010). Este processo é conhecido como carcinogênese e envolve várias etapas, como a iniciação, promoção e progressão do câncer (Figura 1). Durante a iniciação ocorrem mutações genéticas que tornam as células mais suscetíveis ao câncer. Na promoção, ocorre o crescimento e a multiplicação de células pré-neoplásicas, enquanto, na progressão, acontece a transformação de células pré-neoplásicas em neoplásicas e a invasão para outras partes do tecido, formando a metástase (SIDDIQUI *et al.*, 2015; CARLBERG; VELLEUER, 2021).

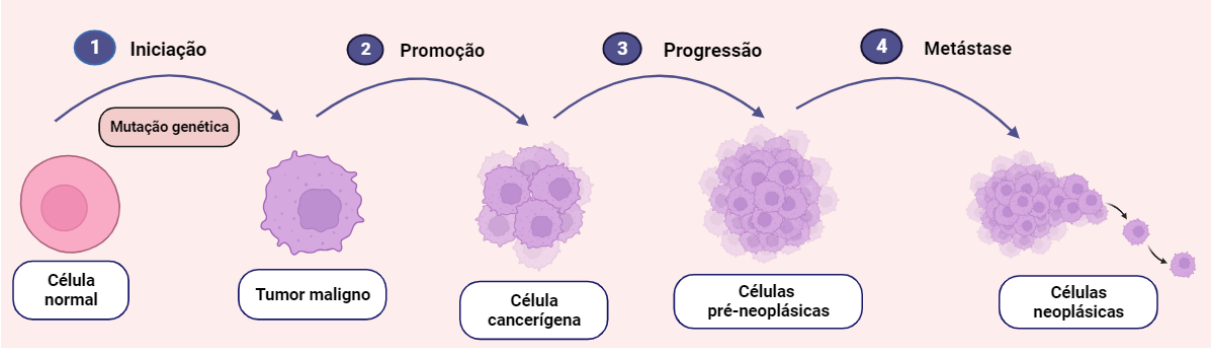


Figura 1 – Etapas da carcinogênese: iniciação, promoção, progressão e metástase. Figura criada em <Biorender.com>.

Por sua vez, destaca-se que não todas as mutações ou exposições a substâncias

cancerígenas desencadeiam necessariamente ao desenvolvimento do câncer, já que diversos tipos de cânceres apresentam suas próprias caraterísticas e fatores de risco (JAMAL-HANJANI *et al.*, 2015). Além disso, a compreensão dos processos envolvidos na carcinogênese são essenciais para o desenvolvimento de novas terapias, detecção precoce e prevenção em qualquer tipo de câncer (QUIRICO; ORSO, 2020).

O câncer se destaca pela sua heterogeneidade em termos de aparência e funcionalidade, desempenhando um papel crucial no desenvolvimento, progressão do tumor e na resposta ao tratamento (LIU; DANG; WANG, 2018; KUMAR *et al.*, 2022). Entretanto, menosprezar essa heterogeneidade pode resultar em erros graves na identificação e escolha da terapia adequada para o tratamento da doença (GUO *et al.*, 2019). De acordo com Kumar *et al.* (2022), as variações genéticas observadas entre diversos tipos de tumores em diferentes pacientes são denominadas como heterogeneidade intertumoral, sendo consideradas como relevante na prática clínica. Além disso, a variação que ocorre entre múltiplos tumores do mesmo tipo, que surgem dentro de um único paciente, é denominada heterogeneidade intrapaciente. Dentro de um mesmo tumor, podem ser identificadas subpopulações de células cancerosas com atributos distintos, tanto em termos de aparência quanto de componentes moleculares, fenômeno que é reconhecido como heterogeneidade intratumoral (Figura 2). Essa complexidade intratumoral pode ser explorada por meio da análise de células individuais (do inglês, *single-cell*) (LEI *et al.*, 2021).

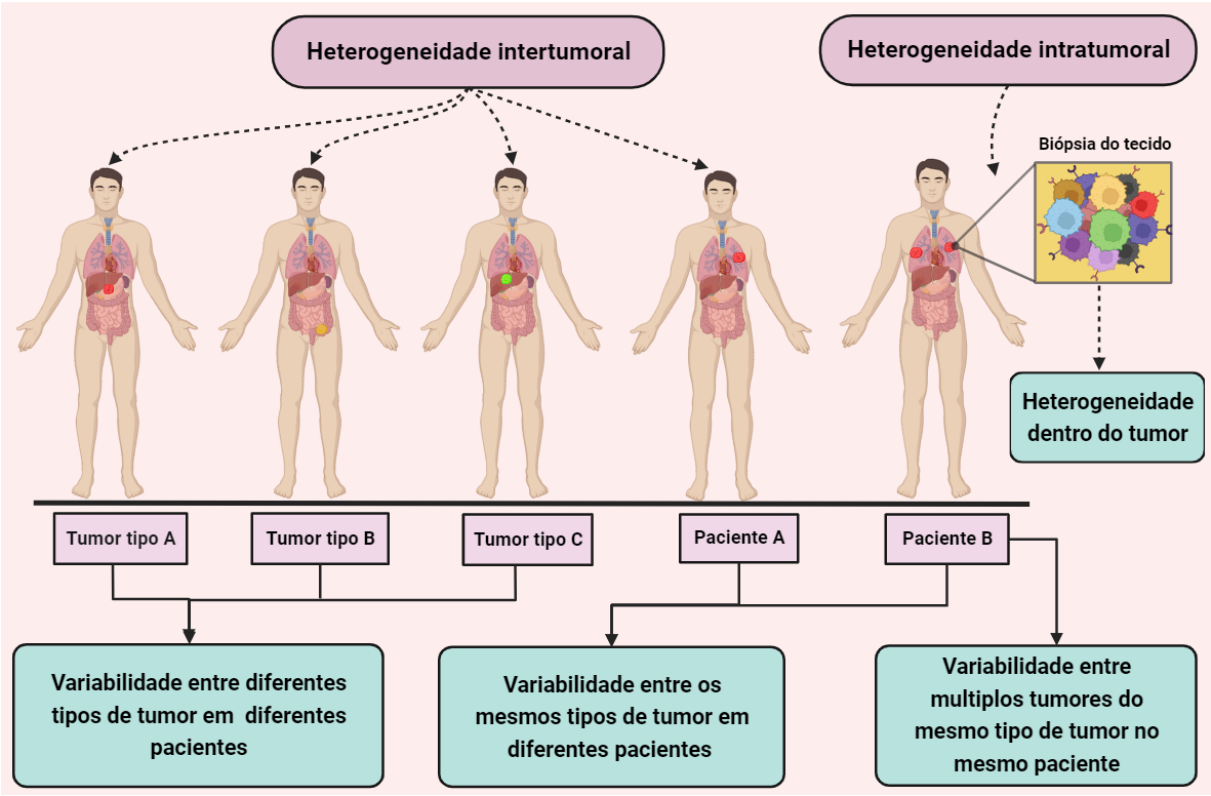


Figura 2 – Ilustração da heterogeneidade tumoral (intertumoral e intratumoral). Figura criada em <Biorender.com>.

2.2 Cânceres de cabeça e pescoço

No ano de 2020, o câncer foi a causa de morte de 9,9 milhões de pessoas em todo o mundo, tornando-se a principal doença responsável pela maioria dos óbitos e a estimativa global foi de cerca de 19,3 milhões de novos casos de câncer (SUNG *et al.*, 2021). Durante este período, o câncer foi responsável pela maioria das mortes na Ásia (58,3%), seguida pela Europa (19,6%), América (14,2%) e teve consequências menores na África e Oceania (SUNG *et al.*, 2021). No Brasil, estima-se para cada ano do triênio 2023–2025, 704 mil casos novos de câncer, excluindo o câncer de melanoma (INCA, 2022). Dentro deste cenário, o câncer de cabeça e pescoço tem aumentado nas últimas décadas, ocupando a sétima posição no ranking mundial de doenças malignas nos últimos anos, mostrando que a estimativa do câncer da cavidade oral e laringe representaram 2% e 1% de novos casos neoplásicos, respectivamente (BRAY *et al.*, 2020; SUNG *et al.*, 2021; MODY *et al.*, 2021). Nos Estados Unidos, esse tipo de câncer representa quase 4% de todos os cânceres e a doença é diagnosticada duas vezes mais em homens com mais de 50 anos do que em mulheres (SIEGEL *et al.*, 2021). Por sua vez, no Brasil, os cânceres de cavidade oral e laringe são os mais comuns, com cerca de 16 mil e 8 mil novos casos estimados a cada ano no triênio 2023-2025, respectivamente. Neste período, de acordo com o Instituto Nacional de Câncer (INCA, 2022) a incidência desses cânceres prevalece em homens com mais de 40 anos nas regiões Sudeste e Centro-Oeste. Nos estados de São Paulo, Minas Gerais, Rio de Janeiro e Santa Catarina, são esperados 4300, 1620, 1560 e 1050 novos casos de câncer de cavidade oral, e 1980, 810, 700 e 520 novos casos de câncer de laringe, respectivamente.

O câncer de cabeça e pescoço envolve um grupo de cânceres heterogêneos, geneticamente complexos e de difícil tratamento que surgem no epitélio mucoso, que incluem as regiões da cavidade oral (como lábios, gengivas, mucosa bucal, palato duro, língua anterior, assoalho bucal e trígono retromolar), nasofaringe (localizada atrás da cavidade nasal), orofaringe (incluindo amígdalas, tonsilas palatinas, tonsilas linguais, base da língua, palato mole, úvula e parede posterior da faringe), hipofaringe (que se estende desde o osso hioide até a cartilagem cricoide) que formam parte da faringe, laringe (supraglótica, glótica e subglótica), cavidade nasal e seios paranasais (como seios maxilares, etmoidais, esfenoidais e frontais), além das glândulas salivares (PORCHERI; MITSIADIS, 2020; CHOW, 2020; JOHNSON *et al.*, 2020) (Figura 3).

A maioria dos cânceres de cabeça e pescoço com origem nas células epiteliais são classificados como carcinoma de células escamosas de cabeça e pescoço (CCECP). O carcinoma de células escamosas (CCE), em particular, representa mais de 90% de todos os casos de tumores malignos nesta área (CHOW, 2020). Dentre as suas variantes, estão inclusos o carcinoma verrucoso ou fusiforme (CCEF), basaloide (CCEB), queratinizante (CCEQ) e não-queratinizante (CCENQ), que frequentemente se desenvolvem na mucosa do trato aerodigestivo superior (GRANDIS; WHO, 2017). Neste contexto, os sítios anatômicos mais afetados são a cavidade oral e a laringe, principalmente em homens adultos (SHAW; BEASLEY, 2016; INCA, 2022). Na cavidade oral, a língua é o local mais comumente afetado, seguido pelo assoalho da boca,

que apresenta um prognóstico pior em relação a outros locais e resulta em menores taxas de sobrevivência nessa região (RODRIGUES *et al.*, 2019).

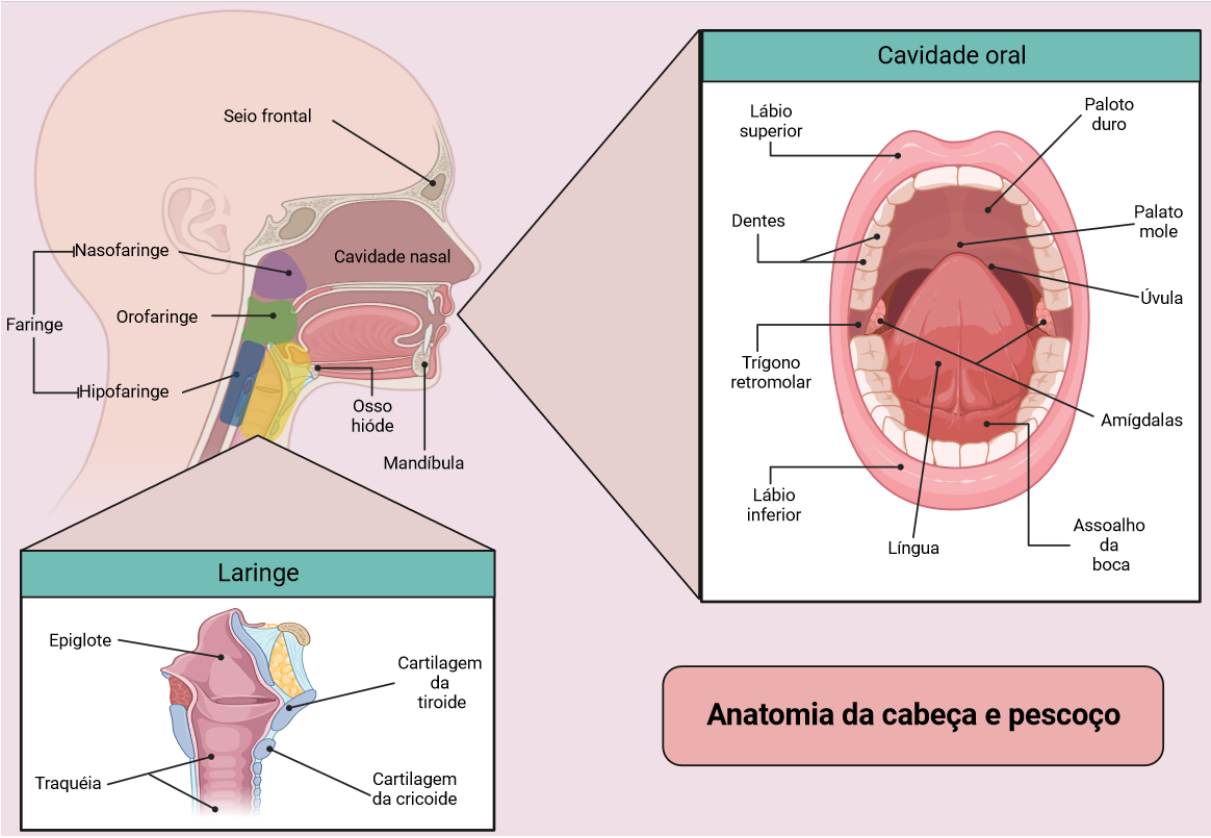


Figura 3 – Ilustração da anatomia da cabeça e pescoço. Figura criada em <Biorender.com>.

Dentre os fatores etiológicos, pesquisas têm demonstrado que o tabagismo e etilismo são fatores de risco bem estabelecidos para o desenvolvimento do CCP, podendo ser afetados por vários aspectos, incluindo o contexto socioeconômico e o nível de industrialização de uma região específica (AUPÉRIN, 2020; GOYAL *et al.*, 2023). Notavelmente, outros fatores podem ser relevantes com base na localização anatômica; tais como, a infeção do vírus papiloma humano (HPV) para orofaringe (YANG *et al.*, 2019; SALAHUDDIN *et al.*, 2023). Por sua vez, esses fatores foram associados à taxa de sobrevivência dos pacientes (DENISSOFF *et al.*, 2022; SALAHUDDIN *et al.*, 2023).

Na maioria dos casos de câncer, inclusive no CCP, mais de 60% dos pacientes são diagnosticados nos estadiamentos avançados, resultando em um prognóstico desfavorável e causando sequelas mais graves (KONINGS *et al.*, 2020; KOWALSKI *et al.*, 2020). No entanto, o tratamento para esta doença específica continua sendo um desafio, devido à sua localização anatômica e à intrincada heterogeneidade tumoral e dependendo do estadiamento, diferentes abordagens terapêuticas são empregadas (PURAM; ROCCO, 2015). Estes tratamentos podem ser a cirurgia, quimioterapia, radioterapia e quimiorradioterapia, os quais aumentam a sobrevivência e reduzem efeitos colaterais, mas também podem acarretar sequelas que comprometem a qualidade

de vida dos pacientes (NIGRO *et al.*, 2017; ALSAHAFI *et al.*, 2019; CHOW, 2020). Tanto nos estadiamentos iniciais quanto avançados, a radioterapia é uma das principais alternativas de tratamento (COLEVAS *et al.*, 2018). Especificamente no caso do câncer de cavidade oral, a cirurgia é a primeira escolha terapêutica nos estadiamentos inicias e avançados (ANDERSON *et al.*, 2021). Na região orofaríngea, as opções de tratamento podem incluir tanto intervenções cirúrgicas quanto radioterapia, sendo que esta última desempenha um papel crucial na preservação das funções da laringe (GALBIATTI *et al.*, 2013; ADELSTEIN *et al.*, 2017; BOZEC *et al.*, 2019).

O CCP se baseia no sistema de estadiamento TNM (Tumor primário, Linfonodos regionais, Metástase à distância) do *American Joint Committee on Cancer* (AJCC) e da *Union for International Cancer Control* (UTCC), que é uma ferramenta amplamente empregada no prognóstico de pacientes, planejamento e avaliação de tratamentos (LYDIATT *et al.*, 2017). As edições mais prevalentes são a 4ª, 7ª e 8ª, sendo que a última traz alterações substanciais em relação à sétima edição, específicas para cada localização anatômica e estadiamento tumoral (AMIN *et al.*, 2017; BRIERLEY; GOSPODAROWICZ; WITTEKIND, 2017). Em geral, os tumores nos estadiamentos iniciais (TNM estadiamento I e II) são caracterizados por terem um tamanho que varia entre 2 e 4 centímetros (cm), sem evidência de comprometimento dos gânglios linfáticos ou metástases. Quando o tumor excede 5cm e se dissemina apenas para os gânglios linfáticos circundantes (TNM estadiamento III) ou para outras partes do corpo (TNM estadiamento IV, subdividido em IVa, IVb e IVc, que são conhecidos como doença extirpável avançada, não extirpável e metastática, respectivamente) (CHOW, 2020).

Apesar dos avanços constantes no diagnóstico, no entendimento dos marcadores moleculares e nas melhorias das abordagens terapêuticas, o diagnóstico do CCP ainda ocorre em estadiamentos avançados devido à complexidade da doença, que apresenta elevados níveis de heterogeneidade inter e intratumoral que levam ao insucesso terapêutico (JAMAL-HANJANI *et al.*, 2015; CANNING *et al.*, 2019). Neste contexto, os diferentes expressões dos miRNAs e mRNAs têm se mostrado promissores como potenciais biomarcadores[1] diagnósticos (para identificar doenças específicas), biomarcadores prognósticos (para prever a probabilidade de eventos clínicos, recorrência ou progressão de doenças) e biomarcadores preditivos (para avaliar a resposta a tratamentos terapêuticos) contribuindo para uma melhor compreensão da regulação gênica e dos processos biológicos relacionados ao CCP e a outros tipos de cânceres (PENG; CROCE, 2016; AHMAD; IMRAN; AHSAN, 2023).

2.3 MiRNAs na regulação da expressão gênica

Os miRNAs são pequenas moléculas de ácido ribonucleico (RNA), não codificantes (ncRNA), compostas por 19 a 24 nucleotídeos de comprimento que auxiliam no processo de regulação da expressão gênica, sendo envolvidos em diversos processos biológicos e

[1] Sinal de um processo biológico normal, patológico ou resposta de algum regime terapêutico específico que pode ser detetado em células ou fluidos biológicos (CALIFF, 2018; AHMAD; IMRAN; AHSAN, 2023).

patológicos, incluindo proliferação celular, diferenciação, apoptose, resposta ao estresse, imunidade, envelhecimento e regulação transcricional, entre outros (COWLAND; HOTHER; GRØNBAEK, 2007; WALAYAT; YANG; XIAO, 2018).

A biogênese dos miRNAs (Figura 4) começa com a transcrição de regiões genômicas situadas dentro ou entre genes codificadores de proteínas pela *RNA POLIMERASE II*, que resulta na produção de moléculas primárias de miRNA (pri-miRNA) e são processadas no núcleo pelo complexo enzimático *DROSHA/DGCR8* em precursores de miRNA (pre-miRNA) de aproximadamente 70 nucleotídeos. Os pre-miRNAs são transportados para o citoplasma pela proteína *EXPORTIN-5*, onde são processados por *DICER1* para formar miRNAs maduros de fita dupla com cerca de 22 nucleotídeos em consequência incorporados ao complexo *RISC* (*RNA-Induced Silencing Complex*), que separa as duas fitas e induz a repressão ou degradação do mRNA [2] (RNA mensageiro) alvo (COWLAND; HOTHER; GRØNBAEK, 2007; WALAYAT; YANG; XIAO, 2018). Dessa forma, a principal diferença entre miRNAs primários, precursores e maduros está relacionada ao comprimento, estrutura e etapas de processamento. Além disso, um único miRNA é capaz de regular vários mRNAs alvo com funções distintas que interagem em uma rede complexa, e diversos miRNAs podem agir em conjunto para controlar a expressão de um mRNA (WEISS; ITO, 2017).

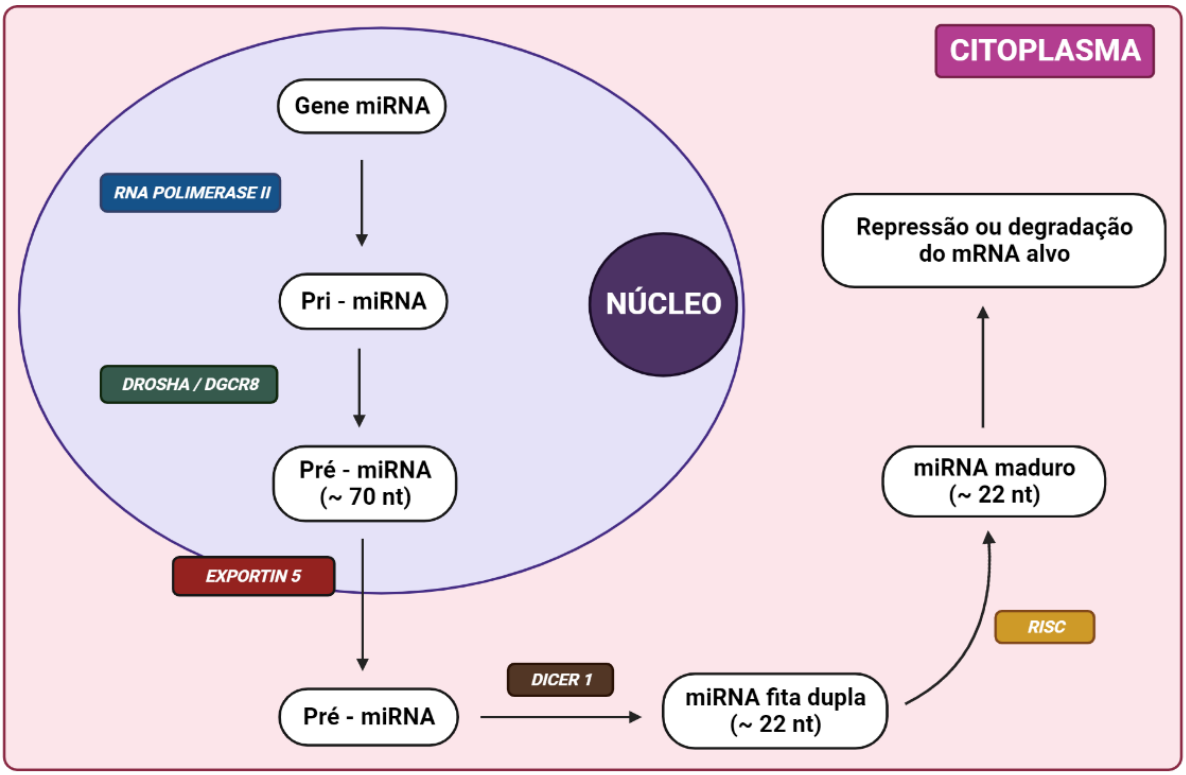


Figura 4 – Biogênese e mecanismo dos miRNAs. Figura criada em <Biorender.com>.

Desde a descoberta dos primeiros miRNAs, *lin-4* por Lee, Feinbaum e Ambros (1993)

[2] O mRNA é uma molécula de RNA que se origina do ácido desoxirribonucleico (DNA) e contém informações genéticas essenciais para a produção de proteínas (DÍEZ; LÓPEZ-NOVOA, 1993).

e *let-7* por Reinhart *et al.* (2000), diversas espécies foram estudadas e tiveram seus miRNAs sequenciados. Atualmente, no repositório público miRBase (versão 22; <www.mirbase.org>) contém 87424 miRNAs catalogados em 291 espécies, por exemplo, no genoma humano existem 1917 pre-miRNAs e 2656 miRNAs de sequências maduras. Esse repositório foi estabelecido em 2002 e tem como principal função gerenciar a nomenclatura, coordenadas e contexto do genoma, referências da literatura, dados de expressão de sequenciamento profundo e anotações orientadas pela comunidade (KOZOMARA; BIRGAOANU; GRIFFITHS-JONES, 2019).

Em 2003, foram estabelecidas normas para a nomenclatura dos miRNAs e seus genes (AMBROS *et al.*, 2003). Na Figura 5, apresenta-se um esquema do significado das diferentes partes do nome dos miRNAs, a qual a numeração é sequencial, refletindo a ordem em que foram descobertos (nomenclatura). O nome começa com um prefixo abreviado de 3 ou 4 letras para designar a espécie ("hsa": *Homo Sapiens*, "rno": *Rattus Norvegicus*, "mmu": *Mus Musculus*, entre outros). As sequências precursoras são chamadas de "mir", enquanto os maduros são chamados de "miR". Consequentemente, dois miRNAs maduros diferentes são gerados a partir de braços opostos do mesmo precursor, denotados da seguinte forma: hsa-miR-99-3p (braço 3p) e hsa-miR-99-5p (braço 5p). Posteriormente, sequências parálogas, cujos miRNAs maduros diferem em apenas uma ou duas posições, são indicadas por sufixos de letras (por exemplo, hsa-miR-26a e hsa-miR-26b). Assim, os transcritos com regiões diferentes no genoma podem produzir miRNAs maduros idênticos, sendo designados com sufixos numéricos, ou seja, hsa-miR-194-1 e hsa-miR-194-2 (GRIFFITHS-JONES *et al.*, 2006).

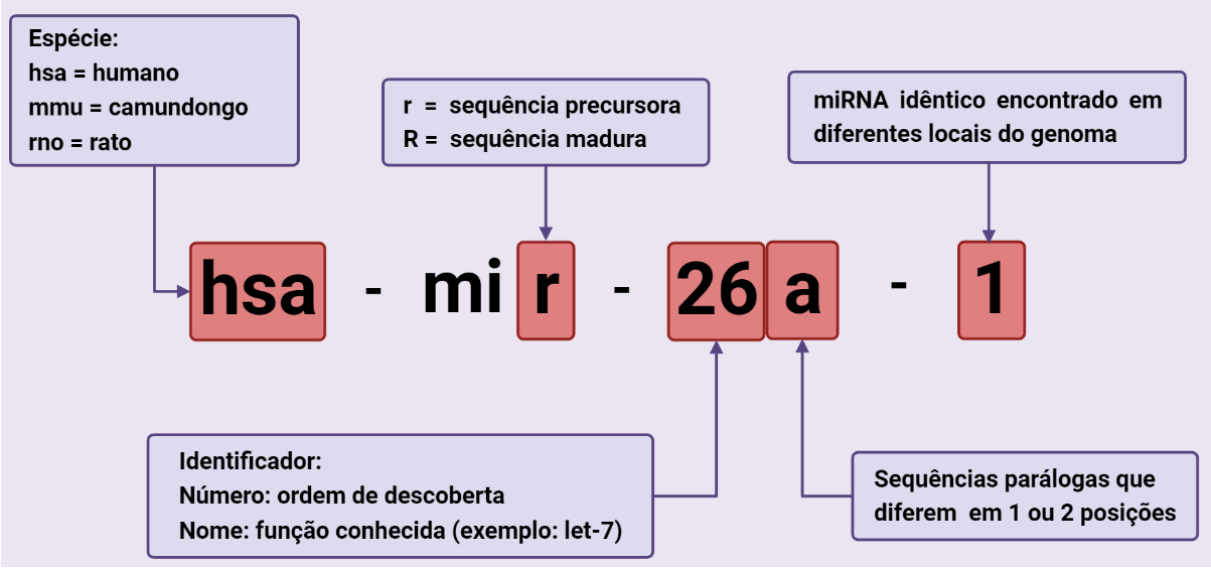


Figura 5 – Nomenclatura dos miRNAs coletados no miRBase.

2.4 MiRNAs e genes envolvidos no câncer de cabeça e pescoço

Estudos recentes, como Cecco *et al.* (2017), Lopez *et al.* (2018), Lu *et al.* (2019), Panvongsa *et al.* (2021), Piotrowski *et al.* (2021) e Qi *et al.* (2021), demostraram que o perfis

de expressão de miRNAs está desregulada em CCP em comparação com tecidos normais ou adjacentes ao tumor e podem ser afetados pelas caraterísticas clinicopatológicas, tais como o sexo, estadiamento e a idade. A desregulação dos miRNAs tem atraído considerável atenção e está sendo cada vez mais estudada devido ao seu papel crucial na regulação da expressão gênica e no desenvolvimento de doenças, especialmente no contexto clínico.

Os miRNAs podem desempenhar funções distintas, atuando como supressores tumorais (TS-miR) ou promotores do câncer (oncomiRs), dependendo de sua expressão e dos seus alvos, como destacado por Alwani e Baj-Krzyworzeka (2021). Os TS-miR são caracterizados por uma redução na expressão (subexpressos) em tumores malignos, resultando na inibição de oncogenes. Por outro lado, os oncomiRs são frequentemente superexpressos em tumores malignos, promovendo a proliferação celular e inibindo a função de genes supressores tumorais, conforme indicado por Ahmad *et al.* (2013).

Dentre os miRNAs frequentemente vinculados ao câncer, incluindo no CCP, o miR-21 é comumente encontrado em tumores sólidos que desempenha um papel crucial na promoção da proliferação e invasão celular (IRIMIE-AGHIORGHIESEI *et al.*, 2019). Outro miRNA relevante é o miR-34a, cuja desregulação demonstrou desempenhar papéis essenciais na tumorigênese e em processos associados à progressão do câncer, como a transição epitelial-mesenquimal, invasão e metástase (LI *et al.*, 2021; KALFERT *et al.*, 2020). Este miRNA também desempenha um papel crucial na supressão do crescimento, induzindo a parada e senescência do ciclo celular, por meio do direcionamento de genes associados à proliferação (WANG *et al.*, 2017).

Além disso, o miR-375 emerge como um biomarcador frequentemente subexpresso no câncer e é proposto como um potencial indicador para a detecção precoce do CCP (CRIMI *et al.*, 2020). Da mesma forma, o miR-155 está implicado no desenvolvimento de vários tipos de câncer, incluindo o CCP (D'SOUZA; KUMAR, 2020). Evidências crescentes sugerem que a expressão aberrante do miR-155 neste tipo de câncer está fortemente correlacionada com a sobrevivência do paciente e a atividade metastática das células cancerígenas (THOMAIDOU *et al.*, 2022).

O CCP não apenas envolve a regulação de miRNAs, mas também é marcado por uma série de alterações genômicas em genes como TP53, CDKN2A, NOTCH1, MET, EGFR, CCND1 e PIK3CA, que desempenham um papel crucial na resistência à quimioterapia, terapias moleculares direcionadas e no prognóstico do paciente (NETWORK *et al.*, 2015; VAHABI; BLANDINO; AGOSTINO, 2021). Especificamente, o gene TP53 é frequentemente mutado no CCP e está associado às células do sistema imunológico (WEI *et al.*, 2023b). Além disso, tumores relacionados ao tabagismo exibem um padrão quase universal de mutações na perda de função no gene TP53, junto com a inativação de CDKN2A (NETWORK *et al.*, 2015).

3 Ferramentas da bioinformática aplicada à genômica

Os avanços nas tecnologias de RNA-Seq têm possibilitado a análise abrangente dos transcritos em células, tecidos ou órgãos específicos, incluindo RNAs codificantes e não codificantes. Essas análises transcriptômicas são fundamentais para identificar genes e avaliar seus níveis de transcrição (ASSIS *et al.*, 2022). Para isso, são empregadas diversas ferramentas de bioinformática especializadas em expressão genética e enriquecimento funcional.

3.1 Análise de expressão diferencial e predição de alvos

A expressão gênica diferencial usando dados de RNA-seq e miRNA-Seq é um método comum para descobrir mecanismos de regulação específicos sob determinados ambientes, estadiamento, tecidos ou estímulos (SONESON; DELORENZI, 2013; MCDERMAID *et al.*, 2019). Para compreender o papel dos miRNAs, é importante identificar os genes que eles regulam por meio da predição de genes alvo candidatos (CHEN; WANG, 2020). Para este propósito, existem várias ferramentas computacionais que podem fornecer predições de genes alvo potenciais para um ou mais miRNAs, sendo recomendado utilizar apenas os genes que são preditos como alvos de miRNAs por múltiplas ferramentas. Essas ferramentas incluem o TargetScan (<https://www.targetscan.org/vert_80/>), miRWalk (<http://mirwalk.umm.uni-heidelberg.de>), DIANA-microT-CDS (<https://dianalab.e-ce.uth.gr/html/dianauniverse/index.php?r=microT_CDS>), mirDIP (<https://ophid.utoronto.ca/mirDIP/>), miRTarBase (<https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/index.php>) e miRDB (<https://mirdb.org/>).

3.2 Redes regulatórias e análise de enriquecimento funcional

A construção da rede de interação é uma representação visual das interações esperadas entre os miRNAs e seus genes alvo, sendo uma etapa crucial para compreender o papel dos miRNAs como reguladores da expressão gênica (O'BRIEN *et al.*, 2018). Esse método envolve a identificação dos miRNAs diferencialmente expressos (DEmiRNAs) e dos mRNAs diferencialmente expressos (DEmRNAs), bem como a previsão dos genes alvo potenciais

regulados pelos miRNAs expressos. Em geral, o processo começa com o uso de ferramentas computacionais para prever os genes alvo potenciais de um ou mais miRNAs. Em seguida, é essencial confirmar se um gene é de fato um alvo do miRNA. Isso pode ser realizado por meio do aumento da expressão do miRNA para diminuir o nível de expressão do gene alvo, ou pela inibição ou silenciamento do miRNA para aumentar a expressão do gene alvo (KUHN *et al.*, 2008). Por fim, os pares de miRNA-mRNA identificados são utilizados para construir uma rede regulatória potencial entre eles.

Após a confirmação da interação, é viável proceder à análise de enriquecimento funcional, uma técnica computacional aplicada em bioinformática e genômica para interpretar conjuntos de genes ou proteínas que demonstram enriquecimento em relação a categorias funcionais específicas, como vias metabólicas, processos biológicos e funções celulares, obtidas de experimentos de RNA-Seq. Essa análise pode ser conduzida tanto de forma direta quanto indireta.

A análise de enriquecimento funcional de forma direta se concentra na Ontologia Genética (GO, do inglês: *Gene Ontology*) e nas vias da Enciclopédia de Genes e Genomas de Kyoto (KEGG, do inglês: *Kyoto Encyclopedia of Genes and Genomes*) diretamente afetados pelo miRNA e, a análise funcional indireta se concentra no GO e nas vias KEGG afetados pelos genes alvo do miRNA. Ambas as análises são úteis para compreender a regulação gênica e os impactos funcionais dos miRNAs em processos biológicos (GODARD; EYLL, 2015).

A anotação GO (<http://www.geneontology.org>) é um recurso que fornece informações sobre genes e suas funções genéticas e usa a ontologia para aprimorar o conhecimento biológico e o KEGG (<https://www.kegg.jp/>) é um banco de dados para a interpretação qualitativa de sequências genômicas e outros dados biológicos, incluindo informações sistemáticas, genômicas e químicas, bem como uma categoria adicional de informações de saúde específicas para humanos. Para auxiliar nessas análises, existem várias ferramentas de bioinformática disponíveis. Algumas delas são a DIANA-miRPath (<http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=mirpath/index>), miEAA (<https://ccb-compute2.cs.uni-saarland.de/mieaa2/>), Enrichr (<https://maayanlab.cloud/Enrichr/>), Toppgene (<https://toppgene.cchmc.org>) e shinyGO (<http://bioinformatics.sdstate.edu/go/>).

Além das redes miRNA-mRNA, também é possível construir redes de interações proteína-proteína (PPI) que desempenham um papel importante na identificação de relações e na compreensão de propriedades biológicas (ATHANASIOS *et al.*, 2017). No entanto, as redes PPI possuem um pequeno número de nós de proteínas altamente conectados (conhecidos como *hubs*) e muitos nós mal conetados. Os genes *hubs* são os mais importantes na organização da estrutura da rede global (HE; ZHANG, 2006). Essas redes PPI podem ser construídas usando o STRING (<https://string-db.org/>) que é um recurso *web* que serve como banco de dados biológicos, concentrando-se em interações conhecidas e previstas entre proteínas. Esta plataforma abrange uma variedade de fontes de informação, como dados experimentais, métodos de previsão computacional e coleções de textos públicos.

4 Métodos estatísticos aplicados à expressão gênica

Este capítulo apresenta uma visão geral dos métodos estatísticos utilizados para estudar a expressão gênica em câncer.

4.1 Modelos Lineares Generalizados

Os modelos lineares generalizados (MLGs) foram introduzidos por Nelder e Wedderburn (1972) e são uma classe de modelos estatísticos amplamente utilizada nas diferentes áreas científicas e em análise da expressão gênica. O MLG é uma generalização flexível da regressão linear que permite relacionar a média da variável resposta com as variáveis preditoras por meio de uma função de ligação (MENARD, 2002). Estes modelos pressupõem que a variável resposta possua uma das várias distribuições pertencentes à família exponencial. De acordo com Turkman e Silva (2000) os MLGs são formados por três componentes:

1) O componente aleatório corresponde a uma variável resposta Y com observações independentes $(Y_1, Y_2, \cdots, Y_n)$ que adota uma distribuição da família exponencial (Binomial, Poisson, Binomial Negativa, Gama, Normal entre outros). Além disso, a esperança matemática é denotada por μ, isto é, $E(Y_i) = \mu_i$.

2) O componente sistemático, também conhecido como preditor linear, denotado por η corresponde ao vetor de n componentes, ou seja,

$$\eta_i = \sum_{j=p} \beta_j x_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \tag{4.1}$$

em que x_{ij} é o valor do j-ésimo preditor no i-ésimo indivíduo.

3) A função de ligação $g(.)$ relaciona a esperança matemática da variável resposta com o preditor linear, dada por

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \tag{4.2}$$

Existem muitas funções de ligação, uma delas é a função de ligação de identidade $g(\mu) = \mu$ que modela diretamente a média, para dados de contagem, a função de ligação mais aplicada

é a log, chamada de log-linear já que é a função de ligação canônica no caso Poisson. Em dados binários ("sucesso" ou "falha"), a função de ligação é a $logit$ que transforma a média em uma escala linear.

A seguir apresentam-se três modelos pertencentes à família exponencial: o modelo de Poisson e o modelo Binomial Negativa, que são úteis para modelar dados de contagem, e o modelo Logístico utilizado para dados binários.

4.1.1 Modelo de regressão de Poisson

O modelo de regressão de Poisson modela dados discretos (contagens) com valores inteiros não negativos. Nesse contexto, Turkman e Silva (2000) e Casella e Berger (2010) mencionam que a função de probabilidade dessa distribuição é denotada pela variável aleatória (v.a), $Y_i \sim Poisson(\mu_i)$, dada por

$$f(y_i;\mu_i) = \frac{\mu_i^{y_i}\exp(-\mu_i)}{y_i!}, \text{para } y_i \in Z_0^+;\ \mu_i > 0, \tag{4.3}$$

em que $y_i = 0, 1, 2\cdots$, é a quantidade de sucessos e, o parâmetro μ é a taxa de ocorrência. A média e variância é denotada por: $E(Y_i) = Var(Y_i) = \mu_i$.

De acordo com Monroy, Rivera e Dávila (2018), o modelo de regressão de Poisson, portanto, é dado pelo logaritmo da média para a função de ligação de modo que o modelo log-linear com uma variável explicativa pode ser expresso como

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \tag{4.4}$$

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}). \tag{4.5}$$

Em certas situações em que há correlação entre as observações, falta de homogeneidade, presença de dados atípicos ou excesso de zeros, ocorre o fenômeno conhecido como sobredispersão, onde a variância é maior que a média (SALINAS-RODRÍGUEZ; MANRIQUE-ESPINOZA; SOSA-RUBÍ, 2009). Originalmente, a distribuição de Poisson era utilizada para modelar dados de contagem de RNA-Seq (sequenciamento de RNA), mas a variabilidade biológica não era bem estimada devido ao problema da sobredispersão (ROBINSON; SMYTH, 2008; ANDERS; HUBER, 2010; LANGMEAD; HANSEN; LEEK, 2010). Para lidar com essa situação, foi proposto o uso da distribuição Binomial Negativa para modelar esses dados (ROBINSON; SMYTH, 2008).

4.1.2 Modelo de regressão Binomial Negativa

A distribuição Binomial Negativa é um modelo estatístico discreto que descreve o número de sucessos em uma sequência de tentativas independentes e idênticas, antes de ocorrer um número

fixo de falhas (CASELLA; BERGER, 2010). Esta distribuição pode ser obtida com a utilização de misturas das distribuições Poisson (μ_i) e Gama (α, β), amplamente aplicada em dados biológicos.

Seja a variável resposta Y_i que segue distribuição Binomial Negativa com parâmetros μ_i e k, denotada por $Y_i \sim BN(\mu_i, k)$, então a função de probabilidade é dada por,

$$f(y_i, \mu_i, k) = \frac{\Gamma(k + y_i)}{\Gamma(k)\Gamma(y_i + 1)} \left(\frac{\mu_i}{\mu_i + k}\right)^{y_i} \left(\frac{k}{\mu_i + k}\right)^{k} \tag{4.6}$$

com $y_i = 0, 1, 2, \cdots p$, sendo a média e a variância dada por,

$$E(Y_i) = \mu_i \tag{4.7}$$

$$Var(Y_i) = \mu_i + \frac{\mu_i^2}{k} \tag{4.8}$$

em que $\frac{1}{k}$ é um parâmetro da dispersão que é estimado juntamente com os parâmetros da regressão, de modo que, se $\frac{1}{k} \longrightarrow 0$, então, $Var(Y_i) \longrightarrow \mu_i$, o modelo se torna equivalente ao modelo de Poisson. Por outro lado, para um valor fixo de k esta distribuição pertence à família exponencial, de modo que um MLG Binomial Negativa pode ser definido e uma função de ligação canônica é utilizada (HARDIN *et al.*, 2007; HILBE, 2014). Este modelo pode ser ajustado usando várias ferramentas estatísticas, mas para análise dos dados de RNA-Seq, existem uma série de implementações no repositório *Bioconductor* (GENTLEMAN *et al.*, 2004) para o R (R Core Team, 2022), incluindo, os pacotes DESeq2 (LOVE; HUBER; ANDERS, 2014) e edgeR (ROBINSON; MCCARTHY; SMYTH, 2010). A diferença entre ambos métodos é a normalização incorporada no modelo (*Trimmed Mean of M-values, TMM*, para edgeR e *Median of Ratios*, para DESeq2) que mostram um bom desempenho consistente em comparação a outros algoritmos de normalização. A continuação descreve-se o modelo Binomial Negativa implementado no DESeq2.

Considera-se Y_{ij} o número de leituras do gene (*counts*) i na amostra j que seguem uma distribuição Binomial Negativa dada por

$$Y_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

em que μ_{ij} é a média esperada das contagens observadas para o gene i na amostra j, e o parâmetro de dispersão α_i define a relação entre a variância da contagem observada e seu valor médio. Em experimentos de RNA-Seq, o número de repetições é geralmente baixo, o que requer a incorporação de mais hipóteses no modelo para obter estimativas confiáveis (ANDERS; HUBER, 2010). Dessa forma, a média (μ_{ij}) é o produto de uma condição dependente do parâmetro $q_{i,j}$ que é proporcional à concentração verdadeira esperada do gene i para a amostra j com um fator de tamanho (número total de leituras) s_j, dada por

$$\begin{aligned} \mu_{ij} &= q_{ij} \times s_j \\ q_{ij} &= \frac{\mu_{ij}}{s_j}, \end{aligned} \tag{4.9}$$

e o objetivo dos fatores de tamanho s_j é fazer que as contagens das diferentes amostras sejam comparáveis devido a que estas foram sequenciadas em diferentes profundidades. No entanto, a variância é definida pela soma da média e variação bruta dos dados, isto é

$$V(Y_{ij}) = \mu_{ij} + \alpha_i \times \mu_{ij}^2, \tag{4.10}$$

Para estimar os diferentes parâmetros do modelo, os dados brutos precisam ser organizados em uma tabela de contagem de dimensões $g \times m$, em que g é o número de genes e m as amostras (ANDERS; HUBER, 2010). Antes de processar os dados, é necessário realizar a normalização para garantir que o número total de leituras entre e dentro das amostras possa ser comparado de forma consistente. Isso se deve ao fato de que alguns genes altamente expressos podem ter um impacto significativo na contagem total de leituras de expressão (MARIONI *et al.*, 2008; ANDERS; HUBER, 2010; ROBINSON; MCCARTHY; SMYTH, 2010; GLUSMAN *et al.*, 2013; DILLIES *et al.*, 2013; LOVE; HUBER; ANDERS, 2014). Para isso, o estimador do fator de tamanho, $\hat{s}_j$, que é definido pela mediana das razões de contagens observadas divididas por uma amostra de pseudo-referência obtida tomando a média geométrica entre as amostras é utilizado, isto é

$$\hat{s}_j = \text{mediana}_i \times \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{\frac{1}{m}}}, \tag{4.11}$$

em que k são as contagens observadas e m é o número de repetições.

Assim, assumindo que q_{ij} está relacionado aos preditores por meio da relação log-linear, é dada por

$$\log[E(Y_{ij})] = \log_2(q_{ij}) = \beta_0 + \beta_i x_j + \varepsilon \tag{4.12}$$

e, por consequência,

$$E(Y_{ij}) = q_{ij} = 2^{(\beta_0 + \beta_i x_j)} \tag{4.13}$$

sendo β_i os coeficientes estimados e x_j os preditores. Devido à natureza logarítmica da função de ligação deste modelo, ao realizar a estimativa dos coeficientes, é necessário aplicar a exponenciação aos coeficientes do modelo, bem como aos intervalos de confiança correspondentes.

Entretanto, na análise de expressão gênica, os coeficientes estimados são exibidos da seguinte maneira:

$$\hat{\beta}_1 = \log_2\left(\frac{\text{Média do tratamento}(\bar{T})}{\text{Média do controle}(\bar{C})}\right) = \log_2(\textit{Fold Change}) \tag{4.14}$$

em que $\hat{\beta}_0$ é o nível de expressão log_2 na referência (amostras de controle); $\hat{\beta}_1$ é conhecido como $\log_2(\boldsymbol{FC})$ que é uma medida importante que mede e/ou compara as diferenças na expressão gênica entre duas condições experimentais diferentes (por exemplo, diferença de expressão no gene entre caso doente e o saudável) usando as medidas de tendencia central (média ou mediana). É

uma maneira de expressar essa diferença em termos de um fator multiplicativo que é mais fácil de interpretar do que uma mudança aditiva, sendo frequentemente usada na análise de dados de expressão gênica em experimentos de *microarray* e RNA-Seq. No entanto, a base do logaritmo pode ser 2, 10 ou o logaritmo natural (ln) e a escolha da base é em grande parte uma questão de interpretação conveniente, mas de acordo com Smyth, Yang e Speed (2003) é adequado usar $\log_2$ para as estimativas dos coeficientes β_j de maneira que sejam expressos em unidades de um aumento de $2\log_2(FC)$ para o gene i em cada x_j, porém o *FC* não representa mudança, mas não há um critério estabelecido a serem considerados nos estudos de RNA-Seq.

Por exemplo, se um gene apresenta uma expressão 2 vezes maior na condição experimental A em relação à condição experimental B, então o $\hat{\beta}_1$ será igual a 1, o que significa que a expressão desse gene na condição A é 2 elevado à potência 1 (2^1) vezes maior do que na condição B. Se, por outro lado, um gene apresentar uma expressão 4 vezes menor na condição A em relação à condição B, então o $\hat{\beta}_1$ será igual a -2, o que significa que a expressão desse gene na condição A é 2 elevado à potência -2 (2^{-2}) vezes menor do que na condição B. Assim, $\hat{\beta}_1 = 0$ representa igualdade de expressão ao comparar A com B (SMYTH; YANG; SPEED, 2003; DALMAN *et al.*, 2012).

Nos estudos da expressão gênica, os resultados são representados nos gráficos MA e de vulcão. O gráfico MA foi originalmente desenvolvido para visualizar os resultados de microarranjo de cDNA (DNA complementar), mas também é útil para análises de RNA-Seq já que permite uma visão global da relação entre a mudança de expressão entre as condições, a força média de expressão dos genes (M) e a capacidade do algoritmo em detectar expressão gênica diferencial (A). Geralmente, os genes com valor de p ajustado são coloridos para facilitar a identificação (LOVE; HUBER; ANDERS, 2014). Por outro lado, o gráfico de vulcão é um tipo de diagrama de dispersão bilogarítmico que apresenta o $\log_2(FC)$ (taxa de variação entre o valor máximo e mínimo entre as duas condições observadas) no eixo das abscissas e o logaritmo negativo do valor de p do $\log_2(FC)$ no eixo das ordenadas. Dessa forma, os pontos de interesse com alta significância estatística e acentuada variação de magnitude são geralmente encontrados na região superior do gráfico (LI, 2012).

Os coeficientes do MLG Binomial Negativa são ajustados sob a hipótese nula ($H_0 : \beta_i = 0$) com a finalidade de verificar se os coeficientes são significativamente diferentes de zero, para isso, a estatística do teste de Wald (1943) (W) é usada para testar a significância de cada um dos coeficientes do modelo,

$$W = Z_0 = \frac{\hat{\beta}_i}{\widehat{EP}(\hat{\beta}_i)} \sim N(0,1), \tag{4.15}$$

em que $\widehat{EP}(\hat{\beta}_j)$ é o erro padrão assintótico do coeficiente estimado $\hat{\beta}_j$. Sob a hipótese nula, Z_0 segue uma distribuição normal padrão.

O intervalo de confiança (IC) a $100(1-\alpha)\%$ para β_i, é dada por:

$$[\hat{\beta}_i \pm z_{1-\frac{\alpha}{2}} \times \widehat{EP}(\hat{\beta}_i)] \tag{4.16}$$

em que $z_{1-\frac{\alpha}{2}}$ é a quantil da distribuição normal padrão.

Para maiores informações com respeito a este modelo procurar em Anders e Huber (2010) e Love, Huber e Anders (2014).

4.2 Teste de múltiplas hipóteses

O teste de múltiplas hipóteses ou comparações múltiplas está relacionado ao problema das múltiplas inferências para o controle do erro tipo I, devido à utilização de uma grande quantidade de variáveis, ou seja, existe um aumento de erros de estimativa conforme mais testes são efetuados simultaneamente (ABDI *et al.*, 2007; FILHO, 2015). Neste contexto, na análise da expressão diferencial, há muitos contrastes a realizar, em que normalmente existem milhares de genes, pelo que os métodos de comparações múltiplas são normalmente necessários (FOULKES, 2009).

Como a detecção de genes diferencialmente expressos envolve a realização de um grande número de testes estatísticos, tendo várias hipóteses testadas, a probabilidade conjunta de que o erro tipo I seja cometido aumenta expressivamente (BENJAMINI; HOCHBERG, 1995). O erro tipo I, também denominado de falso positivo, é o erro que se comete ao rejeitar a hipótese nula quando a mesma é verdadeira (CASELLA; BERGER, 2010). Os tipos de erros são mostrados na Tabela 1.

Tabela 1 – Tipos de erros de um teste de hipóteses.

Situação real	Decisão	
	Não rejeitar H_0 (Não significativo)	Rejeitar H_0 (Significativo)
H_0 Verdadeira	Decisão correta	Erro tipo I (α)
H_0 Falsa	Erro tipo II (β)	Decisão incorreta

Tradicionalmente, os testes estatísticos baseiam-se no erro tipo I, para que o nível de teste, α, geralmente com um valor de 0,05, que mede a probabilidade de rejeitar incorretamente a hipótese nula (adequado para uma única comparação). Maiormente o valor de p, definido como a probabilidade de observar um resultado igual ou mais extremo que a estatística do teste, é comparado com α para tomar a decisão de rejeitar ou não a hipótese nula (CASELLA; BERGER, 2010).

Se considerarmos o problema de contrastar m hipóteses nulas por $H_0^1, H_0^2, \cdots H_0^m$ em cada uma das m tentativas poderiam ocorrer os dois tipos de erros reconhecidos que são mostrados na Tabela 2. Dessa maneira, Verhoeven, Simonsen e McIntyre (2005) explicaram na Tabela 2 a forma de identificar os possíveis resultados de uma série de hipóteses.

Tabela 2 – Erros em contrastes de hipóteses múltiplas.

Situação real	Decisão		Total
	Não rejeitar H_0 (Não significativo)	Rejeitar H_0 (Significativo)	
H_0 Verdadeira	U	V	m_0
H_0 Falsa	T	S	$m - m_0$
Total	$m - r$	r	m

Considere m como o número de hipóteses testadas; m_0 é o número de hipóteses nulas verdadeiras; $m - m_0$ é o número de hipóteses alternativas verdadeiros; V é o número de falsos positivos (erro tipo I); S é o número de positivos verdadeiros; T é o número de falsos negativos (erro de tipo II); U é o número de negativos verdadeiros e; r é o número de hipóteses nulas rejeitadas.

A maioria dos métodos de ajuste para comparações múltiplas utiliza as taxas $FWER$ *(family-wise error rate)*, correção de Bonferroni e a taxa de falsas descobertas (en inglês, *false discovery rate*, ou FDR), que, como em testes de hipóteses simples, tentam controlar o erro tipo I ou controlar V (ou seja, de rejeitar a hipótese nula quando ela é verdadeira). O $FWER$ define-se como a probabilidade de cometer pelo menos un erro de tipo I, então tem-se que

$$FWER = P(V \geq 1), \tag{4.17}$$

além disso, o $FWER$ pode ser definido com mais precisão em termos de se a hipótese nula é verdadeira (BENJAMINI; HOCHBERG, 1995). Porém, o método mais simples e tradicional para controlar o $FWER$ é chamado de correção de Bonferroni que estabelece um novo critério de significância, dada por

$$\alpha_{ind} = \frac{\alpha_{ref}}{m}, \tag{4.18}$$

em que α_{ref} é a probabilidade desejada de no máximo um falso positivo, m é o número de testes realizados (equivalente ao número de genes existentes) e α_{ind} é o α individual a ser testado para cada valor de p (PROVINCE, 2001). De acordo com Filho (2015), a correção de Bonferroni torna-se ultraconservadora à medida que o número de testes aumenta ficando mais improvável encontrar valores significativos. Nos experimentos de RNA-Seq são estudados uma grande quantidade de genes, pelo qual a correção de Bonferroni é inviável diminuindo o poder do teste para a detecção de genes diferencialmente expressos (ROSA; ROCHA; FURLAN, 2007). Neste caso, um método alternativo é a utilização do FDR que baseia-se na proporção esperada de hipóteses nulas incorretamente rejeitadas, dada por

$$\text{FDR} = E\left(\frac{V}{V+S}\right) = \frac{E(V)}{S}, \tag{4.19}$$

em que $V + S$ é o número total de resultados considerados significativos. Porém, deseja-se manter o valor do FDR abaixo de α, mas este número não é calculado facilmente.

Neste sentido, o método FDR de Benjamini e Hochberg (1995) (FDR-BH) é a mais utilizada para lidar com o problema de testes múltiplos referente à taxa de falsas descobertas. Este método baseia-se nos valores de p ordenados $p_1 \leq p_2, \cdots \leq p_m$ resultantes das m hipóteses nulas independentes testadas simultaneamente $H_0^1, H_0^2, \cdots, H_0^m$ e, sejam, $p_{(1)}, p_{(2)}, \cdots, p_{(i)}$ os valores ordenados, defina-se

$$p_{(i)}^{adj} \leq p_{(i)} \frac{m}{i}. \tag{4.20}$$

Assim, para controlar FDR-BH a um nível de significância $p_{(i)}^{adj} = 5\%$, o ponto de corte será o $p(i)$ com maior i que satisfaça a condição $5\% \leq p_{(i)} \frac{m}{i}$, ou seja, serão rejeitadas as hipóteses com valores de p menores ou iguais a $p(i)$.

4.3 Testes não paramétricos

Os testes não paramétricos são métodos estatísticos que se concentram em medidas de tendência central como a mediana e, não requerem a especificação de uma distribuição paramétrica subjacente para os dados em questão. Estes testes são frequentemente utilizados em situações em que não é possível fazes suposições sobre a distribuição dos dados, quando apresentam desvios significativos em relação a normalidade ou quando o tamanho da amostra é pequeno (KRZYWINSKI; ALTMAN, 2014). No entanto, eles podem ter menos poder estatístico do que os testes paramétricos, quando as suposições paramétricas são atendidas.

4.3.1 Teste não paramétrico de Wilcoxon-Mann-Whitney (WMW)

O teste U de Wilcoxon, que é uma alternativa ao teste t de Student, foi originalmente proposto por Wilcoxon (1945) para comparar as médias de duas amostras com o mesmo tamanho. Mais tarde, esta técnica foi generalizada por Mann e Whitney (1947) e apresentado como o teste de Mann-Whitney ou simplesmente como teste U para amostras com tamanhos diferentes. Assim como o teste de Wilcoxon, o teste de Wilcoxon-Mann-Whitney (WMW) são baseados em postos, o que os torna mais robustos em comparação com os testes paramétricos.

Segundo Gibbons e Chakraborti (2014) a estatística do teste U de Wilcoxon-Mann-Whitney é dada por:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{4.21}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{4.22}$$

sendo que a estatística U é determinada pelo mínimo entre U_1 e U_2, em que U_1 e U_2 são os valores estatísticos de Mann-Whitney para cada grupo, e R_1 e R_2 são as somas dos postos do grupo 1 e grupo 2, respectivamente. Quando o tamanho da amostra é maior que 20 ($n_2 > 20$), a estatística U segue uma distribuição Normal, permitindo a utilização de um teste de hipótese para verificar

a diferença entre as medianas dos grupos, isto é,

$$\begin{cases} H_0 : \text{As medianas dos dois grupos são iguais} \\ H_1 : \text{As medianas dos dois grupos são diferentes} \end{cases} \tag{4.23}$$

e a estatística de teste, para este caso, é dada por:

$$Z = \frac{U - \mu_U}{\sigma_U} = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \sim N(0, 1). \tag{4.24}$$

Este procedimento é comumente empregado na análise de dados RNA-Seq para comparar os níveis de expressão gênica entre dois grupos de amostras (por exemplo, um grupo de pacientes saudáveis e outro com uma doença específica). No entanto, é importante, ter em mente que esse teste não é adequado para avaliar diferenças em todos os conjuntos de dados, e em alguns casos, pode ser necessário recorrer a outras técnicas estatísticas para avaliar as discrepâncias entre as amostras.

4.3.2 Teste Qui-quadrado de Pearson

O teste Qui-quadrado de Pearson (Q_P) desenvolvido por Pearson (1900) é um teste não paramétrico (livre de distribuição) amplamente utilizado para analisar diferenças entre grupos quando as variáveis de interesse são nominais ou dicotômicas e dispostos em uma tabela de contingências bidimensional com s linhas e r colunas (MCHUGH, 2013; RANA; SINGHAL *et al.*, 2015). Nesse contexto, as hipóteses a serem testadas são:

$$\begin{cases} H_0 : \text{Não existe associação entre as variáveis em estudo} \\ H_1 : \text{Existe associação entre as variáveis em estudo} \end{cases} \tag{4.25}$$

e a estatística do teste de Q_P do Qui-Quadrado de Pearson é dada por:

$$Q_P = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \tag{4.26}$$

em que n_{ij} e e_{ij} são as frequências observadas e esperadas, para testar a hipótese nula de não associação (independência) contra hipótese alternativa de associação (dependência). Quando algum valor esperado é menor do que 5 ou a amostra seja menor do que 25 o Teste Exato de Fisher (1922) deve ser utilizado.

4.4 Análise multivariada dos dados

A análise multivariada é composta por diversas ferramentas estatísticas que tem ampla aplicabilidade em ciência de dados, biológicas, bioinformática e outras áreas, pois permite uma análise simultânea de duas ou mais variáveis. Em contraste com a análise univariada, que lida

com uma única variável por vez, a análise multivariada permite que os pesquisadores explorem as relações entre variáveis, identifiquem padrões, redução da dimensionalidade dos dados, tendências e teste de hipóteses (HAIR *et al.*, 2004; JOHNSON; WICHERN, 2014).

De acordo com Johnson e Wichern (2014), os dados multivariados podem ser representados por um arranjo matricial $\boldsymbol{X}$ com p variáveis (colunas) medidas em n elementos (linhas) da seguinte forma:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

em que a matriz $\boldsymbol{X}$ contém os dados observados de todas as variáveis.

4.4.1 Análise de componentes principais

Entre as muitas técnicas adotadas para explorar dados multivariados, a análise de componentes principais (PCA) é uma técnica estatística multivariada que utiliza uma transformação ortogonal para identificar os padrões subjacentes de um conjunto de variáveis correlacionadas (RENCHER, 2002). O objetivo da PCA é encontrar um novo conjunto de variáveis, chamadas de componentes principais, que resuma a maior parte da variabilidade dos dados originais facilitando a visualização e interpretação dos dados (HONGYU; SANDANIELO; JUNIOR, 2016; MANLY; ALBERTO, 2016). Este método é frequentemente utilizado nas aplicações genômicas para a avaliação de qualidade e na análise exploratória em dados de alta dimensão, como RNA-Seq (YEUNG; RUZZO, 2001).

De acordo com Mingoti (2007) e Johnson e Wichern (2014), a PCA transforma o conjunto de variáveis correlacionadas $\mathbf{X}' = \{X_1, X_2, \cdots, X_p\}$ em um conjunto de variáveis não correlacionadas $\mathbf{Y}' = \{Y_1, Y_2, \cdots, Y_p\}$, em que cada Y_i (componente principal) é uma combinação linear das variáveis em $\boldsymbol{X}$. Dessa forma os componentes principais $\mathbf{Y}'$ são, então, as combinações lineares:

$$\begin{aligned} Y_1 &= \boldsymbol{a}_1'\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + a_{1p}X_p \\ Y_2 &= \boldsymbol{a}_2'\boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + a_{2p}X_p \\ &\vdots \\ Y_p &= \boldsymbol{a}_p'\boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + a_{pp}X_p \end{aligned}$$

em que $\boldsymbol{a}_i'$ é o vetor que determina a i-ésima combinação linear. Logo, obtém-se:

$$Var(\boldsymbol{Y}_i) = \boldsymbol{a}_i' \sum \boldsymbol{a}_i, \quad i = 1, 2, \cdots, p,$$

$$Cov(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \boldsymbol{a}_i' \sum \boldsymbol{a}_j, \quad i \neq j.$$

no qual $\sum$ apresenta a matriz dos autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_i \geq 0$.

Assim, as componentes principais são combinações lineares não correlacionadas com a variância máxima. Considerando $\boldsymbol{a}_i$ como os autovetores da matriz de covariância $\sum$, tem-se que o i-ésimo componente principal é dado por:

$$Y_i = \boldsymbol{a}_i' \boldsymbol{X} = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p,$$

em que $i = 1, .., p$ e com variância e covariância dada por:

$$Var(Y_i) = \boldsymbol{a}_i' \sum \boldsymbol{a}_i = \boldsymbol{a}_i' \lambda_i \boldsymbol{a}_i = \lambda_i$$

$$Cov(Y_i, Y_j) = \boldsymbol{a}_i' \sum \boldsymbol{a}_j = \boldsymbol{a}_i' \lambda_j \boldsymbol{a}_j = 0,$$

em que $\boldsymbol{a}_i$ e $\boldsymbol{a}_j$ são ortogonais para $i \neq j$. Dessa forma, a variabilidade total dos componentes principais é equivalente à variabilidade total das variáveis originais, conforme expresso por:

$$traço(\boldsymbol{S}) = \sigma_{11} + \sigma_{22} + \cdots \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i).$$

Portanto, a proporção da variância explicada pelas primeiras componentes principais p é dada por:

$$\sum_{i=1}^{p} \frac{\lambda_i}{p}.$$

Assim, os elementos são organizados em ordem decrescente com base na quantidade de variação que explicam nos dados originais. A primeira componente principal explica a maior parte da variação, seguido pela segunda componente e assim sucessivamente.

Após a identificação dos principais componentes, torna-se essencial determinar quantos deles condensam as informações das variáveis originais em um modelo simplificado, utilizando critérios para estabelecer o número ideal de componentes principais para capturar a maior parte da variação nos dados. Nesse contexto, a retenção dos componentes é fundamentada na porcentagem de contribuição da variabilidade total, geralmente atingindo 80% ou mais dessa variabilidade total. Um critério amplamente adotado é o método de Kaiser (KAISER, 1958), que considera todos os $\lambda_i > 1$ como a variação total dos dados. Outro método comum é a abordagem visual por meio do gráfico de *scree-plot* ou do teste de Cattell (CATTELL, 1966). Esse método envolve a observação do gráfico dos autovalores em relação ao número de dimensões, com o objetivo de identificar o ponto em que os autovalores começam a apresentar uma tendência linear decrescente.

Entretanto, em estudos de expressão gênica, são amplamente utilizados métodos como o *Elbow* e a análise paralela de *Horn's* de (HORN, 1965; BUJA; EYUBOGLU, 1992). O método de Horn's consiste em comparar os autovalores gerados por meio de uma PCA em vários conjuntos de dados aleatórios (ou seja, variáveis não correlacionadas) que possuem o mesmo número de variáveis e observações que o conjunto de dados observacionais. Isso é feito para gerar autovalores ajustados para componentes ou fatores, levando em consideração a inflação induzida pelo erro

amostral. Nesse contexto, os autovalores são retidos quando são maiores que 1, conforme a seguinte condição:

$$\lambda_p - (\bar{\lambda}_p^r - 1), \tag{4.27}$$

em que λ_p é o p-ésimo autovalor dos dados observados (para $p = 1$), e $\bar{\lambda}_p^r$ é o autovalor médio correspondente do número de iterações de conjuntos de dados aleatórios simulados.

Além disso, é essencial destacar a influência da escala de medição das variáveis, que podem apresentar valores em escalas distintas, sobre a identificação dos componentes. Assim, torna-se crucial realizar a normalização (padronização) das variáveis originais, pois estas estão em diferentes escalas. A técnica mais comum para obter os componentes envolve a padronização das variáveis originais utilizando a expressão abaixo, que calcula o escore Z:

$$Z_i = \frac{X_i - \mu_i}{\sigma_{ii}}, \text{ para } i = 1, 2, \cdots, p \tag{4.28}$$

em que μ_i e σ_{ii} são, respectivamente, média e desvio padrão da variável aleatória X_i.

4.4.2 Análise de agrupamento – *Clustering*

A análise de agrupamento é uma técnica de clusterização da análise multivariada, pois nenhuma suposição é feita sobre o número de grupos ou a estrutura do grupo e que visa agrupar indivíduos e/ou variáveis a partir de dados observados com base na similaridade ou dissimilaridade (RENCHER, 2002; HAIR *et al.*, 2009). A medida de similaridade e/ou dissimilaridade entre objetos é um aspecto importante que deve ser considerado ao agrupar dados. Dentre as diversas medidas de similaridade existentes (cada um produz agrupamentos diferentes) destaca-se a Distância Euclidiana, sendo a mais utilizada na prática (HAIR *et al.*, 2009).

Segundo Hair *et al.* (2009) e Johnson e Wichern (2014), a Distância Euclidiana mede a distancia entre dois pontos (x_i, y_j) baseado em suas coordenadas em um espaço de p dimensões e é definida pela seguinte equação:

$$d(x_i, y_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - y_{jk})^2}. \tag{4.29}$$

A distância de dissimilaridade $(\boldsymbol{D})$ é dada por uma matriz de dimensão $n \times n$:

$$\boldsymbol{D} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ d(x_1, y_2) & 0 & \cdots & 0 \\ d(x_1, y_3) & d(x_2, y_3) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ d(x_1, y_n) & d(x_2, y_n) & \cdots & 0 \end{bmatrix},$$

em que d_{ij} representa a distância entre o elemento amostral x_i e y_j.

Essa distância é influenciada por diferentes escalas nas variáveis, portanto, é necessário padronizar os dados ao usar essa métrica. Após escolher a medida de distância, é necessário selecionar o algoritmo para formação dos agrupamentos. Existem dois tipos de métodos: hierárquicos e não hierárquicos (CORMACK, 1971).

Os agrupamentos hierárquicos visam agrupar objetos semelhantes em conjuntos maiores, formando uma estrutura hierárquica que envolve duas etapas. A primeira refere-se à estimação de uma medida de distância entre os indivíduos e a segunda, à adoção de uma técnica de formação de grupos possibilitando a construção de uma árvore, chamado de dendrograma (RENCHER, 2002; SIEGMUND; LAIRD; LAIRD-OFFRINGA, 2004; MINGOTI, 2007; JOHNSON; WICHERN, 2014).

No entanto, esse tipo de agrupamento pode ser diferenciado através do método de ligação que utilizam, ou seja, a forma de recalcular as distâncias entre os grupos. Os métodos mais conhecidos são apresentados em Gauch e Jr (1982), Rencher (2002), Hair *et al.* (2009) e Johnson e Wichern (2014), os quais são: métodos de ligação simples (*single linkage*, vizinho mais próximo ou mínima distância), ligação completa (*complete linkage*, vizinho mais distante ou máxima distância), ligação média (*average linkage*, distância média), método Ward e Ward.D2. Estes métodos possuem características e desempenhos diferentes quando comparados entre si. Na maioria dos casos, o método de ligação completa oferece melhores resultados em comparação com o método de ligação simples (D'HAESELEER, 2005). Crispim *et al.* (2020) mostrou que os métodos de ligação completa, média e Ward possuem os melhores desempenhos de forma geral. Em alguns casos, a seleção do melhor método pode ser realizada usando a correlação cofenética (CARVALHO; MUNITA; LAPOLLI, 2019). Em dados de RNA-Seq, Zhang *et al.* (2018) e Jaskowiak, Costa e Campello (2018) concluíram que o método de ligação completa produz os melhores resultados.

4.5 Árvores de decisão e floresta aleatória

As árvores de decisão são algoritmos de aprendizado de máquina () que requerem poucas suposições estatísticas, criam regras de divisão para formar grupos homogêneos e podem analisar múltiplas variáveis, sejam numéricas ou categóricas mesmo quando apresentam multicolinearidade (KINGSFORD; SALZBERG, 2008; HAN; KAMBER; PEI, 2012; VILLE, 2013). Embora uma única árvore de decisão possa oferecer previsões precisas, a combinação dos resultados de várias árvores, conhecido como floresta aleatória (do inglês, Random Forest) (RFT) pode aumentar ainda mais sua precisão (BREIMAN, 2001; ALFARO; GÁMEZ; GARCÍA, 2018; KHAN *et al.*, 2021).

Uma árvore de decisão é composta por nós, ramos e folhas (Figura 6). A construção da árvore começa no topo da estrutura, chamado de nó raiz, que é o primeiro nó a ser criado e dá origem aos nós intermediários, incluindo o próprio nó raiz, que formam os nós de decisão e

nas extremidades da árvore, estão os nós folhas que representam os valores de predição para a variável dependente da amostra (BRIEMAN *et al.*, 1984; HAN; KAMBER; PEI, 2012; KHALID *et al.*, 2022).

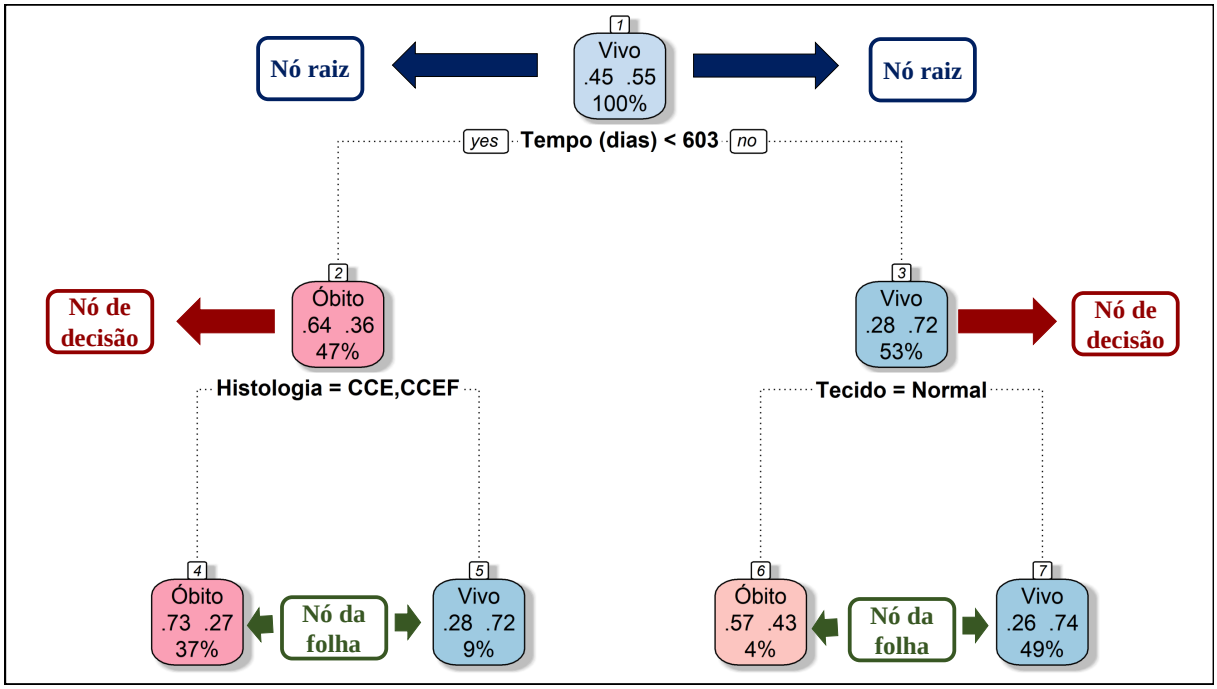


Figura 6 – Ilustração da árvore de decisão.

Diversos algoritmos são aplicados na criação de árvores, operando de forma recursiva para selecionar o melhor atributo a fim de dividir os dados e expandir os nós folha da árvore, até que um critério de parada seja alcançado. Exemplos destes algoritmos incluem o *ID3 - Iterative Dichotomizer 3*, proposto por Quinlan (1986), o *C4.5* de Quinlan (1993), que busca maximizar o ganho de informação, e o CART (*Classification and Regression Trees*), proposto por Brieman *et al.* (1984) que utiliza a impureza Gini para problemas de classificação ou a soma de quadrados para regressão.

Para problemas de classificação, o processo inicial envolve o cálculo da entropia de um conjunto de entrada (S) com duas classes, em um cenário binário, isto é,

$$\textit{Entropia}(S) = \sum_{i=1}^{2} -p_i \log_2(p_i), \tag{4.30}$$

em que $p_i = \frac{n_i}{n}$ representa a probabilidade de uma amostra pertencer à classe i.

Logo, o ganho de informação ao selecionar um atributo A é a redução esperada na entropia, que é dado por:

$$GI(S, A) = \sum_{x \in P(A)} \frac{|S_x|}{|S|} \textit{Entropia}(S_x), \tag{4.31}$$

em que $P(A)$ é o conjunto de valores que o atributo A pode assumir, x um elemento pertence à esse conjunto e S_x o subconjunto de S formado pelo dados, em que $A = x$.

Para árvores de regressão, a impureza é definida pelo conceito de soma de quadrados:

$$SQ(s,t) = SQ_{Total} - (SQ_{Esquerda} + SQ_{Direita}), \tag{4.32}$$

em que t é um nó da árvore de regressão; s é um ponto de divisão do nó; $SQ(s,t)$ é a diferença de soma de quadrados entre o nó pai e seus filhos, SQ_{Total} é a soma de quadrados calculada em todo o nó t; $SQ_{Esquerda}$ é a soma de quadrados calculada para os y_i ($y_i < s$) e $SQ_{Direita}$ é a soma de quadrados calculada para os y_i ($y_i > s$), para decidir o melhor ponto de divisão binária do nó basta escolher o s que minimize $SQ(s,t)$.

No entanto, a árvore de decisão resultante, tanto para classificação quanto para regressão, pode se tornar excessivamente complexa, incluindo partições que não contribuem significativamente para a predição. Para evitar esse problema, é comum realizar uma poda, que consiste em remover partições da árvore baseada no custo-complexidade (BRIEMAN *et al.*, 1984; MONARD; BARANAUSKAS, 2003).

Entretanto, Hothorn, Hornik e Zeileis (2006) propuseram o algoritmo de árvores baseado na inferência condicional (CTree), referindo-se ao *MOB* (*Model-Based Recursive Partitioning*) de Zeileis, Hothorn e Hornik (2008), para contornar os problemas associados aos algoritmos tradicionais. O CTree é uma classe de árvores de decisão não paramétricas para classificação e regressão, que utiliza a teoria estatística inferencial, mais especificamente testes de hipóteses baseados em permutações aleatórias, para selecionar as variáveis usadas no particionamento recursivo (ponto de divisão). A poda é baseada em um teste estatístico que aprimora a capacidade preditiva da árvore de decisão (HOTHORN; HORNIK; ZEILEIS, 2006).

Com base no método de árvores de decisão, foi desenvolvido o *bagging*, uma técnica utilizada para diminuir a variância das previsões ao escolher diversas amostras com reposição para a construção das distintas árvores de decisão (BREIMAN, 2001).

A técnica conhecida como *bagging*, quando aplicada especificamente a árvores aleatórias, é denominada como florestas aleatórias (RFT). Consiste na criação de um conjunto de B árvores, projetadas para serem não viesadas e sem a necessidade de poda. Em cada árvore, um subconjunto de covariáveis é selecionado a partir do conjunto inicialmente disponível nos nós, e com base nisso, uma árvore de decisão é ajustada. Este processo é repetido B vezes, utilizando amostras *bootstrap* da amostra original. Após os B ajustes, a convergência das árvores é analisada, resultando na classificação automática das observações, conforme ilustrado na Figura 7 (KHAN *et al.*, 2021).

No entanto, ao criar árvores com o algoritmo RFT, algumas observações não são utilizadas na construção de cada árvore. Estas observações são referidas como *out-of-bag* (OOB). Dessa forma, é possível obter previsões para uma observação usando apenas as árvores para as quais essa observação não foi utilizada na construção. Isso resulta em uma classificação para cada observação, e a taxa de erro pode ser estimada com base nessas previsões (BREIMAN, 2001; JANITZA; HORNUNG, 2018). Adicionalmente, o RFT realiza uma classificação com base na importância das variáveis para prever a variável resposta, utilizando as medidas *Mean Decrease*

Gini (MDG) e *Mean Decrease Accuracy* (MDA) (BREIMAN, 2001; JANITZA; HORNUNG, 2018). O MDG é uma métrica que avalia a contribuição de cada variável para a homogeneidade dos nós e a eficácia na floresta aleatória resultante. Quanto maior a precisão média de diminuição ou a pontuação média de Gini de diminuição, maior é a relevância da variável no modelo. Essa medida é derivada do cálculo do índice de Gini dada por:

$$MDG = 1 - \sum_{j=1}^{c} (p_j)^2, \tag{4.33}$$

em que $j = 1, 2, \cdots c$ é o número de classes da variável resposta categórica e p_j é a frequência relativa da classe c no modelo.

No entanto, o MDA é uma métrica que avalia a precisão de um modelo para excluir cada variável. Quanto mais afeta a precisão, mais importante é a variável para a classificação exitosa. É determinado permutando aleatoriamente as observações de cada variável sem afetar as demais, e calculando a diferença entre a quantidade corretamente classificada antes e depois da permutação em cada árvore do modelo, isto é,

$$MDA = \frac{1}{B} \sum_{b=1}^{B} (R_{OOB} - R_{perm}), \tag{4.34}$$

em que R_{OOB} é o número de classificações corretas antes da permutação e R_{perm} é o número de classificações corretas após permutação (BREIMAN, 2001; JANITZA; HORNUNG, 2018).

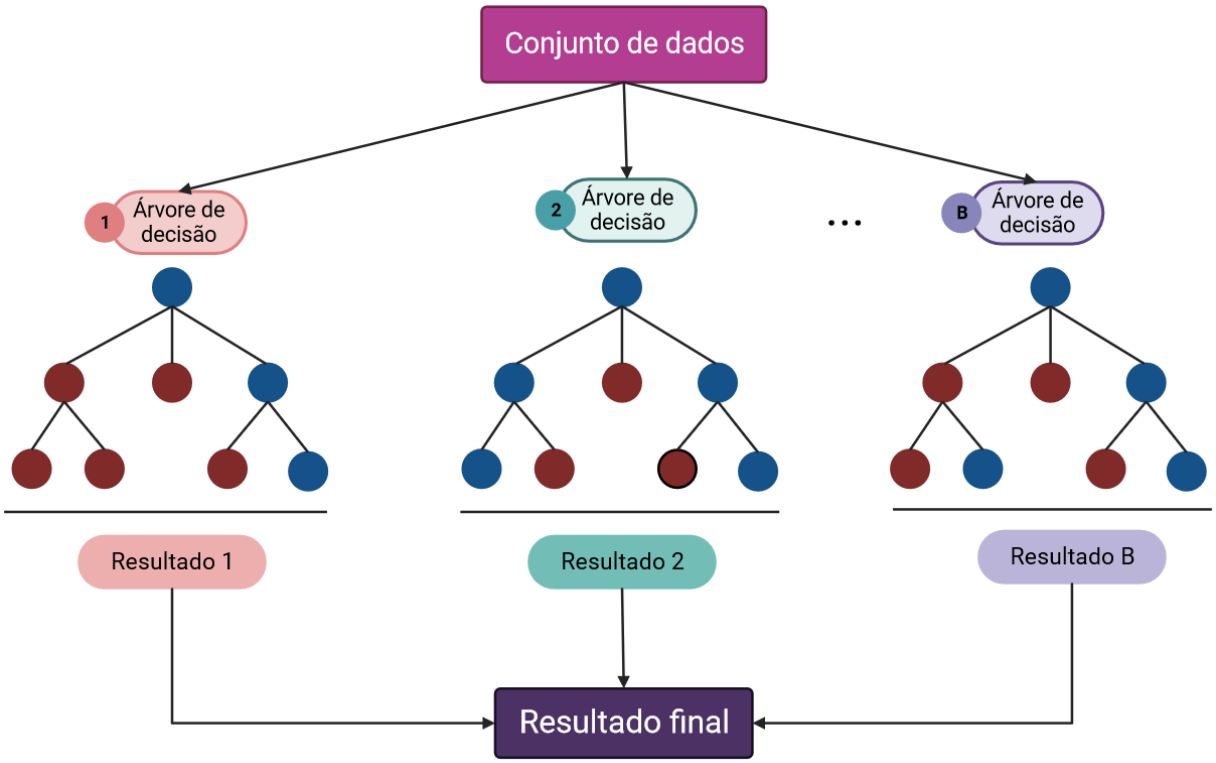


Figura 7 – Ilustração da floresta aleatória. Figura adaptada de Khan *et al.* (2021).

4.5.1 Técnicas de reamostragem para dados desbalanceados

Em cenários de classificação binária, depara-se frequentemente com o desafio do desbalanceamento de classes, que ocorre quando a distribuição dos valores da variável de resposta é desproporcional, sendo a classe relacionada ao evento de interesse significativamente menor do que a outra classe (não evento), gerando um desafio complexo no qual os algoritmos de aprendizado convencionais podem não atingir um desempenho satisfatório (THABTAH *et al.*, 2020). Para abordar essa questão, diversas técnicas amplamente utilizadas são empregadas para equalizar a distribuição de dados entre as diferentes classes. Estas técnicas são:

- *Undersampling* ou *downsampling* é um método não heurístico que visa equilibrar a distribuição de classes através da eliminação aleatória de exemplos de classes majoritárias (KOTSIANTIS; PINTELAS, 2003).
- *Upsampling* ou *oversampling* visa equilibrar aleatoriamente a distribuição de classes em um conjunto de dados, adicionando exemplos à classe minoritária (LING; LI, 1998).
- SMOTE (*Synthetic Minority Oversampling Method*) cria instâncias adicionais da classe minoritária por meio da interpolação dos valores das instâncias minoritárias mais próximas dentro de um conjunto específico (CHAWLA *et al.*, 2002).
- ROSE (*Random Over-Sampling Examples*) integra métodos de *undersampling* e *Upsampling* para ampliar a quantidade de dados, especialmente em amostras da classe minoritária, que ajuda o classificador a estimar uma regra de classificação mais preciso, pois a mesma atenção será dirigida a ambas as classes (LUNARDON; MENARDI; TORELLI, 2014).

4.5.2 Métricas de desempenho das árvores de decisão

Um dos principais propósitos ao desenvolver um modelo de classificação reside na sua aplicação no âmbito da previsão. Dessa maneira, torna-se viável avaliar a eficácia do ajuste ao lidar com dados que não foram incluídos no processo de construção do modelo. Em modelos de classificação, é frequente empregar a matriz de confusão como uma ferramenta para analisar os acertos do modelo (FAWCETT, 2006), conforme exemplificado na Tabela 3.

Tabela 3 – Matriz de confusão comparando valores preditos e valores observados.

Valor observado	Valor predito	
	Sucesso (1)	Falha (0)
Sucesso (1)	Verdadeiro Positivo (TP)	Falso Negativo (FN)
Falha (0)	Falso Positivo (FP)	Verdadeiro Negativo (TN)

A partir da análise da matriz de confusão, é possível estabelecer diversas métricas para avaliar o desempenho da classificação. As mais comuns incluem a acurácia (indicativa da taxa de classificação correta), sensibilidade e especificidade, que medem a precisão em relação a tipos específicos de acertos (FAWCETT, 2006). Além destas, diversas outras medidas são amplamente empregadas e estão detalhadas na Tabela 4.

Tabela 4 – Métricas de desempenho na classificação.

Medidas	Notação	Fórmula	Intervalo
Área sob a curva	AUC	$\int_{t_0}^{t_1} ROC(t)dt$	[0 - 1]
Área delimitada pela curva ROC.			
Especificidade	TNR	$\frac{TN}{TN+FP}$	[0 - 1]
Capacidade de detectar corretamente os casos negativos.			
Sensibilidade / *Recall*	TPR	$\frac{TP}{TP+FN}$	[0 - 1]
Capacidade de detectar corretamente os casos positivos.			
Acurácia	ACC	$\frac{TP+TN}{TP+TN+FP+FN}$	[0 - 1]
Avalia a taxa de acertos do modelo.			
Acurácia balanceada	BAL ACC	$\frac{TPR+TNR}{2}$	[0 - 1]
Capacidade de prever corretamente todas as classes.			
Precisão	PR	$\frac{TP}{TP+FP}$	[0 - 1]
Proporção de instâncias classificadas como positivas que são verdadeiramente positivas.			
F1-score	F1	$2 \times \frac{TNR \times TPR}{TNR+TPR}$	[0 - 1]
Capacidade de manter um bom equilíbrio entre a precisão e a sensibilidade.			
Capa de Cohen	KAP	$\frac{2\times(TP+TN-FP\times FN)}{(TP+FP)(FP+TN)+(TP+FN)(FN+TN)}$	[0 - 1]
Mede a proporção de concordância.			

Uma maneira adicional de avaliar a qualidade ou semelhança da classificação de um modelo em relação às classes reais é por meio da curva ROC (do inglês, *Receiver Operating Characteristic*), representada na Figura 8. Esse gráfico analisa a taxa de verdadeiros positivos no eixo vertical (sensibilidade) e a taxa de falsos positivos ($1-$ especificidade) no eixo horizontal, conforme varia o ponto de corte das probabilidades estimadas pelo classificador (MARZBAN, 2004; FAN; UPADHYE; WORSTER, 2006; FAWCETT, 2006).

A partir dessa curva, surgiu uma métrica que serve como indicador único para avaliar a capacidade preditiva dos modelos de classificação: a Área Sob a Curva ROC (do inglês, *Area Under Curve*), que é uma medida de desempenho aplicável a modelos de classificação binária (FAWCETT, 2006). As regras para interpretar valores de AUC são as seguintes: AUC $= 0,5$ é sem discriminação; $0,6 \leq$ AUC $< 0,7$ existe uma discriminação aceitável; $0,7 \leq$ AUC $< 0,8$ existe uma discriminação boa; AUC $\geq 0,8$ discriminação excelente (MARZBAN, 2004; FAN; UPADHYE; WORSTER, 2006).

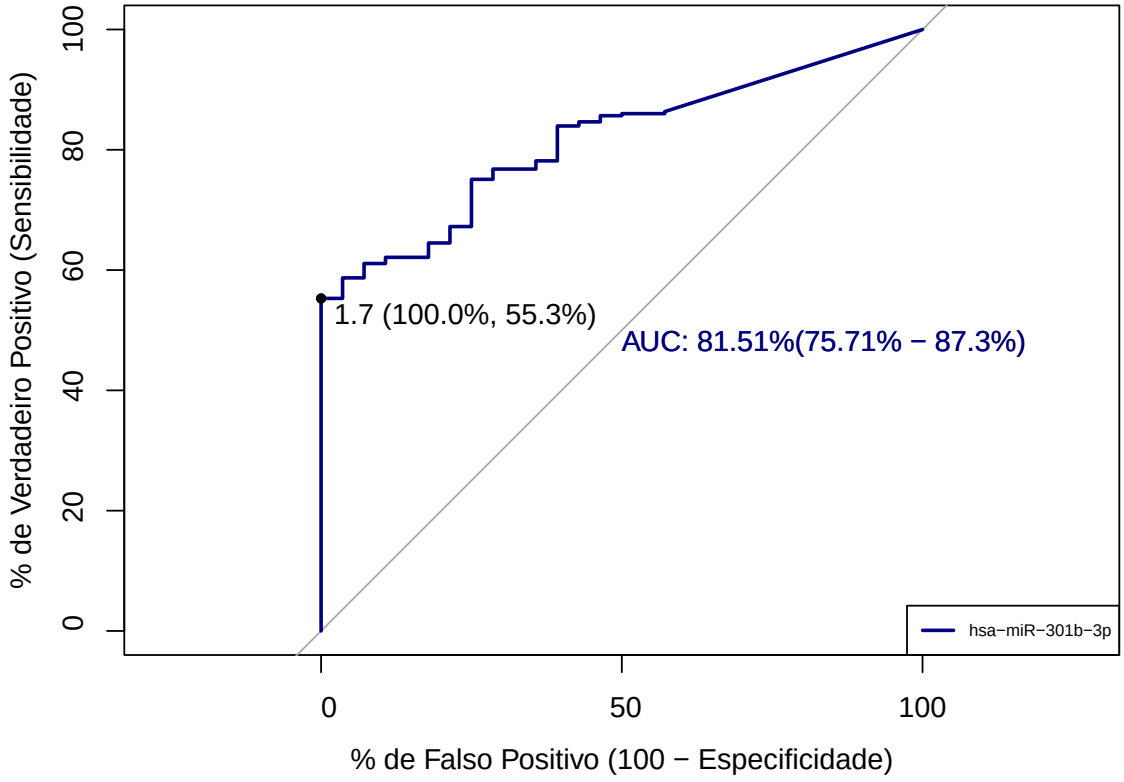


Figura 8 – Ilustração da curva ROC juntamente com a Área Sob a Curva ROC.

A avaliação da capacidade preditiva do modelo pode ser realizada por meio de dois métodos de validação amplamente utilizados: *Holdout* e *k-fold* (YAKA *et al.*, 2022). No método *Holdout*, o conjunto de dados é dividido em dois conjuntos distintos: treinamento e teste. O modelo é então treinado no conjunto de treinamento e avaliado no conjunto de teste para medir seu desempenho. Por outro lado, na validação cruzada *k-fold*, o conjunto de dados é dividido em k partições equitativas. Em cada iteração, uma dessas partições é designada para teste, enquanto as demais são utilizadas para treinamento. Os resultados de todas as iterações são agregados para gerar previsões abrangentes para o conjunto de dados completo. Embora a validação cruzada forneça uma avaliação imparcial, o *Holdout* pode introduzir viés devido à aleatoriedade no processo de divisão (YAKA *et al.*, 2022). Estes métodos estão ilustrados na Figura 9.

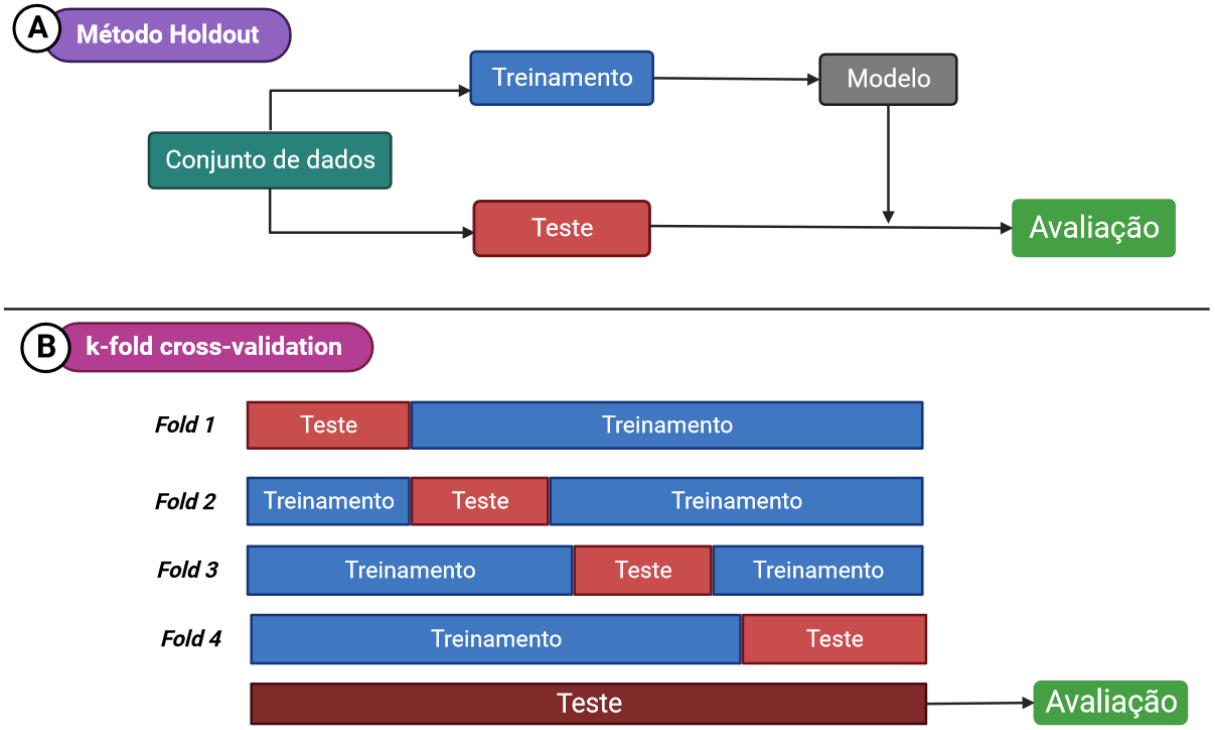


Figura 9 – Ilustração esquemática dos métodos de validação. (A) Método Holdout e (B) validação cruzada *k-fold*.

4.6 Análise de sobrevivência

A análise de sobrevivência é um tema relevante na pesquisa do câncer, pois possibilita a predição do tempo até a morte ou progressão do tumor, bem como fornece informações potenciais sobre os fatores que causam a doença (COLOSIMO; GIOLO, 2006). Em particular, a expressão gênica tem sido amplamente utilizada nesta área, já que as alterações na regulação dos genes são comuns no câncer, e aplicadas em diversos métodos de sobrevivência, como por exemplo, o modelo de risco proporcional de Cox (MALENOVÁ; ROWSON; BOEVA, 2021).

Análise de sobrevivência é uma coleção de métodos estatísticos usados para abordar questões relacionadas ao tempo de acompanhamento medido desde um início definido até a ocorrência de um determinado evento conhecido como tempo de falha, por exemplo, diagnóstico de doença, recuperação ou outra experiência de interesse relacionada a um indivíduo ou objeto em estudo (KLEINBAUM; KLEIN *et al.*, 2012; MAIMON; ROKACH, 2014). Em alguns casos, os períodos de tempo são considerados censurados devido à presença de informações incompletas, ou seja, quando o acompanhamento do paciente é interrompido antes do término do estudo e o evento de interesse não é observado (COLOSIMO; GIOLO, 2006; WIENKE, 2010).

De acordo com Colosimo e Giolo (2006), Lemeshow, May e Hosmer (2011) e Kleinbaum, Klein *et al.* (2012), a censura à direita (tempo de falha é maior do que o tempo de estudo) é caracterizada pela presença de três mecanismos de censura: censura do tipo I, censura do tipo

II e censura aleatória. A censura do tipo I ocorre quando o estudo é finalizado em um período de tempo estabelecido. Já a censura do tipo II acontece quando o estudo é interrompido após um número predefinido de falhas. Por fim, na censura aleatória, o paciente é retirado do estudo antes da ocorrência da falha por algum motivo. Além desses, existem outras formas de censura, como a censura à esquerda, que ocorre quando o tempo observado é maior ou igual ao tempo de falha, e a censura intervalar, que é útil para estudar eventos cuja hora exata de ocorrência é desconhecida, mesmo que os dados estejam disponíveis antes e depois da ocorrência do evento.

Neste contexto, Colosimo e Giolo (2006) representam a variável resposta pelo par (t_i, δ_i), sendo t_i ($i = 1, 2, \cdots, n$) o tempo registrado do i-ésimo individuo e a variável indicadora de falha ou censura, isto é,

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

Um dos principais desafios decorrentes da censura é a complexidade, e às vezes a impossibilidade, de aplicar métodos estatísticos convencionais ao tratar dados. Assim, em estudos de sobrevivência, três funções são usadas para caracterizar a distribuição do tempo, tais como, a função de sobrevivência, a função de risco e a função de densidade de probabilidade. Para isso considera-se, uma variável aleatória T não negativa e contínua para o tempo de falha ($T \geq 0$), e sejam $f(t)$ e $F(t)$ as probabilidades da função de densidade e da função de distribuição cumulativa de T, respectivamente. Então, a função de sobrevivência, $S(t)$, a função de risco, $h(t)$, e a função de risco cumulativo, $\Lambda(t)$ são especificadas como:

$$S(t) = 1 - F(t) = P(T \geq t) = \int_t^{\infty} f(u)du, \tag{4.35}$$

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \tag{4.36}$$

$$\Lambda(t) = \int_0^t h(u)du. \tag{4.37}$$

Além disso, as funções previamente definidas permitem a relação da função de risco com a função de sobrevivência, ou seja,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log(S(t)). \tag{4.38}$$

Logo, as funções $f(t)$, $S(t)$, $h(t)$ e $\Lambda(t)$ estão relacionadas entre si (COLOSIMO; GIOLO, 2006; KLEINBAUM; KLEIN *et al.*, 2012; LEE; LIM, 2019).

Na análise de sobrevivência, a função $h(t)$ é extremamente útil para estimar $S(t)$ devido a que representa a probabilidade condicional de experimentar um evento específico, dado que a sobrevivência persistiu até o momento em questão. Ademais, a função $h(t)$ leva em conta as informações censuradas, considerando a probabilidade condicional de que essas observações tenham sobrevivido até o momento em questão, apesar da censura (COLOSIMO; GIOLO, 2006; LEE; LIM, 2019).

4.6.1 Estimador de Kaplan-Meier

O estimador não paramétrico de Kaplan-Meier (KM) proposto por Kaplan e Meier (1958) é frequentemente empregado em estudos clínicos para estimar a função de sobrevivência a partir de dados censurados. O estimador da probabilidade de sobrevivência até o tempo t é representado por $\hat{S}(t)$, enquanto a curva de sobrevivência é obtida ao plotar $S(t)$ em função de t. O estimador de KM é definido como segue:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right), \tag{4.39}$$

em que d_j é o número de eventos no tempo t_j ($j = 1, \cdots, k$), k representa os tempos distintos e ordenados de falha; n_j é o número total de indivíduos em risco antes do tempo t_j (COLOSIMO; GIOLO, 2006; KLEINBAUM; KLEIN *et al.*, 2012).

O teste de Log-rank (MANTEL *et al.*, 1966) é um exemplo mais abrangente do teste estatístico Qui-Quadrado, amplamente utilizado em estudos de sobrevivência para dados censurados com o objetivo de testar se há diferença significativa entre as curvas de Kaplan-Meier entre dois ou mais grupos em cada ponto de tempo. As hipóteses do teste são formuladas por:

$$\begin{cases} H_0 : \text{Não há diferença entre as curvas de Kaplan-Meier} \\ H_1 : \text{Há diferença entre as curvas de Kaplan-Meier} \end{cases} \tag{4.40}$$

e a estatística do teste de Log-Rank é

$$\text{Log-rank} = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{N-1}, \tag{4.41}$$

em que N é o número de grupos da variável, O_i é o número de falhas observadas para o grupo i e, E_i é o número esperado de falhas para o grupo i. Por exemplo, o número esperado de falhas para o grupo 1 é calculado por:

$$E_1 = \sum_{j=1}^{t} \left(\frac{r_{1j}}{r_{ij} + r_{2j}} \right) \times (e_{1j} + e_{2j}) \tag{4.42}$$

em que no tempo t_j, r_{1j} é o número de pacientes em risco no grupo 1, r_{2j} é o número de pacientes em risco no grupo 2, e_{1j} é o número de falhas no grupo 1, e_{2j} é o número de falhas no grupo 2 (MANTEL *et al.*, 1966; COLOSIMO; GIOLO, 2006).

4.6.2 Modelo de riscos proporcionais de Cox

O método semi-paramétrico de riscos proporcionais de Cox, também conhecido como modelo de Cox ou regressão de Cox (COX, 1972), é um método estatístico utilizado para analisar a relação entre variáveis explicativas e o tempo até o evento de interesse em estudos de sobrevivência (WALTERS, 1999). Matematicamente, o modelo de Cox pode ser descrito como:

$$h(t|x) = h_0(t) \times \exp\{b_1 x_1 + b_2 x_2 + \cdots + b_p x_p\} = h_0(t) \times \exp\{X'\beta\}, \tag{4.43}$$

em que $h(t)$ é a função de risco no tempo t determinada por um conjunto de covariáveis $(x_1, x_2, \cdots, x_p)$, X é o vetor das variáveis explicativas em que, os coeficientes de regressão são desconhecidos e corresponde o das covariáveis na sobrevivência, determinado por um vetor de componentes dos parâmetros ajustáveis $\boldsymbol{\beta}$, cujo impacto é medido por seus respectivos coeficientes $(b_1, b_2, \cdots, b_p)$ e $h_0(t)$ é o componente não paramétrico do tempo, chamada de função constante ou risco basal correspondente a valor de risco se todos os $\boldsymbol{x_i}$ forem iguais a zero.

Este modelo requer o pressuposto de riscos proporcionais, ou seja, este modelo só é valido se for verificada essa pressuposição, isto é, o *hazard ratio* (HR), ou razão de riscos proporcionais é a razão das taxas de falhas entre dois indivíduos i e j é,

$$\text{HR} = \frac{h_i(t|x_i)}{h_j(t|x_j)} = \frac{h_0(t) \times \exp\{x_i'\beta\}}{h_0(t) \times \exp\left\{x_j'\beta\right\}} = \exp\left\{x_i'\beta - x_j'\beta\right\}, \tag{4.44}$$

que não depende de t. Logo, o HR acima de 1 (ou, $\beta > 0$) indica uma covariável que está positivamente associada à probabilidade do evento e, portanto, negativamente associada ao tempo de sobrevivência; HR < 1 (ou, $\beta < 0$) em caso contrario.

Os coeficientes de regressão $\boldsymbol{\beta}$ são estimados usando a propriedade de invariância do estimador de máxima verossimilhança parcial (EMVP) que pode ser resolvido através do método interativo de Newton-Raphson. Dessa forma o intervalo de confiança (IC) é dado por: $100(1-\alpha)\%$ para $\boldsymbol{\beta}$ é $[\hat{\beta} \pm z_{1-\frac{\alpha}{2}} \times EP(\hat{\beta})]$ (COX, 1972; COLOSIMO; GIOLO, 2006).

Para a seleção do melhor modelo, existem diversas estratégias de modelagem para um modelo multivariado sendo os métodos *forward*, *backward* e *stepwise* os mais utilizadas pela comunidade científica. Porém, existem outros fatores que são considerados na seleção dos modelos devido a que existem variáveis com importância biológica que deveriam permanecer no modelo não sendo estatisticamente significativas pelo qual, o método de *stepwise* é geralmente utilizado já que respeita a permanência das variáveis de importância biológica, juntamente com o critério de informação de Akaike (1974) (AIC):

$$\text{AIC} = -2\sum_{i=1}^{n} \ln L(\hat{\mu}_i, y_i) + 2 \times p, \tag{4.45}$$

em que y_i é o i-ésimo valor da resposta e $\hat{\mu}_i$ é a estimativa de y_i quando se ajusta um modelo de p parâmetros por meio da maximização da função de log-verossimilhança.

A verificação do pressuposto de riscos proporcionais pode ser realizada de forma visual por meio de gráficos, como o gráfico $\log(-\log(S(t)))$ em relação ao tempo, em que se verifica a pressuposição de que as curvas de sobrevivência apresentam paralelismo. Caso as curvas se cruzem, o pressuposto de riscos proporcionais não é válido. Outra abordagem proposta por Grambsch e Therneau (1994), baseada em resíduos ponderados, oferece uma maneira adicional de avaliar esse pressuposto. As hipóteses associadas a essa abordagem é formulado da seguinte

maneira:

$$\begin{cases} H_0 : \text{o pressuposto de riscos proporcionais é válido} \\ H_1 : \text{o pressuposto de riscos proporcionais não é válido.} \end{cases} \tag{4.46}$$

A hipótese nula é rejeitada se o valor da prova global for inferior ao nível de significância estabelecido. Outro método de validação envolve a análise do gráfico dos resíduos de Schoenfeld (1982), que é útil após o ajuste de um modelo de regressão de Cox.

4.6.3 Árvores de regressão para dados censurados

As árvores de regressão para dados censurados à direita, também conhecidas como árvores de sobrevivência são modelos não paramétricos para análise de eventos que ocorrem ao longo do tempo, ou seja, para modelar o tempo até que um evento de interesse ocorra, como morte ou falha (BERTSIMAS *et al.*, 2022). Esses métodos foram propostos, estudados e avaliados por muitos autores utilizando os algoritmos CART© e CTree que permitem a análise de múltiplos fatores simultaneamente e a identificação dos fatores mais relevantes para explicar um determinado evento de interesse. Ciampi *et al.* (1986) sugeriram usar o teste Log-rank para comparação dos dois grupos de observação nos nós filhos e, posteriormente os autores, LeBlanc e Crowley (1993) introduziram um algoritmo baseado na estatística Log-rank por meio do custo-complexidade. Além disso, este método adota o modelo de riscos proporcionais de Cox (LEBLANC; CROWLEY, 1992).

Da mesma forma que na classificação, foi desenvolvida a floresta de sobrevivência aleatória (RSF) proposta por Ishwaran *et al.* (2008), que incorpora muitas das ideias úteis do RF promovidas por Breiman (2001). O interesse principal reside na maneira de selecionar um subconjunto aleatório de variáveis como candidatos à divisão de um nó. A abordagem tradicional de seleção de variáveis em florestas aleatórias baseia-se na importância da variável (VIMP) que é fundamentado na taxa de erro *Out-Of-Bag* (OOB) e varia consideravelmente, dependendo dos dados em problemas de alta dimensão. Dessa forma a VIMP de uma variável Y_j é definida como

$$VIMP(Y_j) = \frac{1}{B}\sum_{b=1}^{B}(\widetilde{errOOB}_b^j - errOOB_b), \tag{4.47}$$

em que $errOOB_b$ é o erro de previsão nos dados OOB. Ademais, os valores $VIMP > 0$ indicam variáveis com capacidade preditiva, enquanto valores zero ou negativos identificam variáveis não preditivas a serem filtradas (BREIMAN, 2001).

Como alternativa, conforme descrito no estudo de Ishwaran *et al.* (2008), a profundidade mínima é introduzida como uma nova medida de seleção de variáveis de alta dimensão. Esta avalia o grau de predição de uma variável pela sua profundidade em relação ao nó raiz de uma árvore. Em uma única árvore de decisão, as variáveis importantes provavelmente aparecerão mais próximas da raiz da árvore, enquanto as variáveis com menos importância estarão frequentemente mais próximas das folhas. Desta forma, a importância da variável pode ser estimada pela profundidade

média em que ela aparece em todas as árvores da floresta. Um valor menor da profundidade mínima corresponde a uma variável mais preditiva, e a eficácia da utilização da profundidade mínima é bem demonstrada como um problema de sobrevivência de alta dimensão.

Contudo, a RSF têm a capacidade de lidar com conjuntos de dados extensos e um grande número de preditores. Além disso, conseguem obter estabilidade notável ao combinar os resultados de várias árvores. No entanto, é importante notar que a combinação dessas árvores pode comprometer significativamente a interpretação intuitiva de uma única árvore (KANTIDAKIS *et al.*, 2020).

4.6.4 Métrica de desempenho para dados censurados

O desempenho do modelo para dados censurados é avaliado por meio do índice de concordância (C-index) de Harrell *et al.* (1982), também conhecido como índice de concordância de Harrel (ou C de Harrell) que foi desenvolvido para dados com censura à direita, como o modelo de riscos proporcionais de Cox. A fórmula do C-index pode ser explicada da seguinte forma: primeiro, ela calcula o número de pares de indivíduos que são concordantes, ou seja, os indivíduos com menor probabilidade de sobrevivência morrem antes dos indivíduos com maior probabilidade de sobrevivência, definido como *concordance*. Em seguida, ela divide esse número pelo número total de pares de indivíduos possíveis, denominada *permissible*.

No entanto para RSF, a estimação do OOB é baseada no C-index, que é definido como

$$OOB = 1 - (C - index), \tag{4.48}$$

e

$$C - index = \frac{Concordance}{Permissible}. \tag{4.49}$$

Portanto, OOB e C-index são indicadores de desempenho de modelos de sobrevivência que variam de 0 a 1. Um valor de OOB próximo de zero indica que o modelo é capaz de fazer boas previsões para dados novos. Um valor de C-index próximo de um indica que o modelo é capaz de classificar corretamente todos os pares de indivíduos em ordem de risco. Um valor de C-index abaixo de 0,5 indica que o modelo é pior que uma escolha aleatória (ISHWARAN *et al.*, 2008).

5 Banco de dados e métodos de análise

Neste capítulo, detalha-se a obtenção do banco de dados, procedimentos e os métodos utilizados. Na Figura 10, apresenta-se o fluxograma das etapas realizadas durante o estudo.

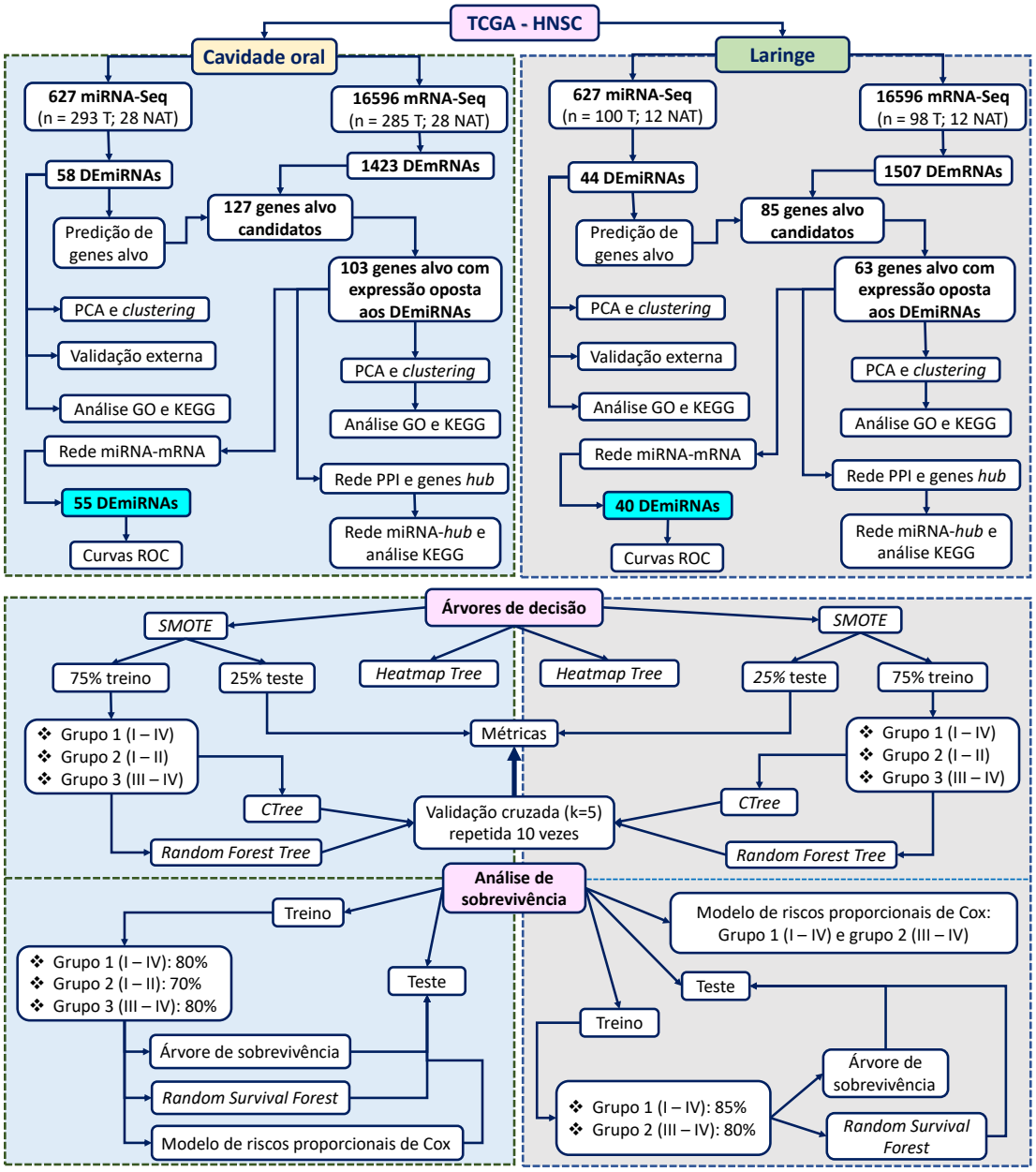


Figura 10 – Fluxograma das etapas utilizadas no estudo.

5.1 Conjunto de dados do TCGA

Os dados utilizados neste estudo foram obtidos do repositório Atlas do Genoma do Câncer (TCGA) por meio do *National Cancer Institute Genomic Data Commons* (GDC), que oferece acesso aberto a perfis transcriptômicos da quantificação de expressão para diferentes tipos de cânceres, bem como dados clínicos de pacientes obtidos por tecnologias de sequenciamento da próxima geração (NGS). Estes dados estão disponibilizados publicamente para a comunidade científica com o objetivo de melhorar a prevenção, diagnóstico, prognóstico e tratamento do câncer, podendo ser baixados manualmente através deste portal (<https://portal.gdc.cancer.gov/>) ou via *scripts* por meio de diferentes linguagens de programação, como `R` e `Python`.

Os dados brutos (*reads*) de sequenciamento de RNA e de miRNAs (miRNA-Seq de nível 3) e as contagens de RNA-seq (transcritos codificadores e não codificadores), juntamente com suas informações clinicopatológicas associadas ao CCP foram baixados em setembro de 2022 a partir da opção *Head and Neck Cancer* do TCGA usando os pacotes `GDCRNATools` (LI *et al.*, 2018) e `TCGAbiolinks` (COLAPRICO *et al.*, 2016) do repositório de código aberto `Bioconductor` versão 3.15 (<https://www.bioconductor.org>) com pacote de instalação `BiocManager` (GENTLEMAN *et al.*, 2004; HUBER *et al.*, 2015) por meio do *software* `R` (R Core Team, 2022). Estes dados não requerem a aprovação do Comitê de Ética por serem dados de domínio público.

Inicialmente, foram disponibilizados três arquivos:

i) Dados clínicos de 528 pacientes diagnosticados com CCP entre os anos 1992 e 2013, com acompanhamento até 2015 e informações clínicas detalhadas, tais como a identificação do paciente, idade, raça, sexo, ano do diagnóstico, entre outros (Tabela 5);

Tabela 5 – Descrição das características clinicopatológicas dos 528 pacientes diagnosticados com câncer de cabeça e pescoço.

Características	Tipo	Descrição
patient_id	Carácter	Identificação de cada paciente
prior_malignancy	Fator	História do paciente de diagnóstico de câncer anterior
primary_diagnosis	Fator	Diagnóstico primário
ajcc_pathologic_stage	Fator	Estadiamento patológico da doença
day_to_last_follow_up	Numérico	Dias até o último acompanhamento
year_of_diagnosis	Numérico	Ano de diagnóstico
ajcc_pathologic_t	Fator	Classificação do tamanho tumoral
ajcc_pathologic_n	Fator	Grau de espalhamento tumoral
ajcc_pathologic_m	Fator	Classificação de distância metastática
tissue_ou_organ_of_origin	Fator	Sítio anatômico do tumor
race	Fator	Raça
gender	Fator	Sexo
vital_status	Fator	Status de sobrevivência
age_at_diagnosis	Numérico	Idade
year_of_diagnosis	Numérico	Ano do diagnóstico patológico inicial do câncer
icd_10_code	Carácter	Classificação Internacional de Doenças (CID10)
pharmaceutical_treatmet_or_therapy	Numérico	Tratamento farmacêutico ou radioterapia

ii) a matriz de expressão composta por 567 amostras de tecido e 2588 miRNA-Seq, os quais estão anotados no miRbase v.21 (Tabela 6);

Tabela 6 – Matriz de expressão do miRNA-Seq. Cada linha corresponde a um miRNA específico e cada coluna uma amostra referente ao tumor primário (exemplo, TCGA-4P-AA8J-01) e tecido sólido normal ou adjacente ao tumor (exemplo, TCGA-CV-6933-11) com seus respectivos níveis de expressão associados ao câncer de cabeça e pescoço.

Nº	ID miRNA	miRNAs	TCGA-4P-AA8J-01	TCGA-CV-6933-11	···	TCGA-CV-6955-11	TCGA-WA-A7H4-01
1	MIMAT0000062	let-7a-5p	157574	420187	···	227562	150782
2	MIMAT0004481	let-7a-3p	246	123	···	180	850
3	MIMAT0010195	let-7a-2-3p	26	5	···	3	48
4	MIMAT0000063	let-7b-5p	60535	166891	···	139214	115110
5	MIMAT0004482	let-7b-3p	155	232	···	191	193
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2584	MIMAT0027368	miR-6733-3p	1	1	···	2	1
2585	MIMAT0027410	miR-6755-5p	4	2	···	10	7
2586	MIMAT0027422	miR-6761-5p	2	1	···	1	4
2587	MIMAT0027513	miR-6806-3p	2	2	···	8	5
2588	MIMAT0027541	miR-6820-3p	1	1	···	0	1

iii) a matriz de expressão composta por 548 amostras de tecido e 60616 transcritos (considerando os que codificam proteínas e os que não codificam), mas apenas foram utilizados os genes codificadores de proteínas, que totalizaram 19962 mRNA-Seq (Tabela 7).

Tabela 7 – Matriz de expressão de genes transcritos codificadores (mRNA-Seq). Cada linha corresponde a um mRNA específico e cada coluna uma amostra referente ao tumor primário (exemplo, TCGA-4P-AA8J-01) e tecido histologicamente normal adjacente ao tumor (exemplo, TCGA-H7-A6C5-11) com seus respectivos níveis de expressão associados ao câncer de cabeça e pescoço.

Nº	ID gene	mRNAs	Tipo	TCGA-4P-AA8J-01	TCGA-BA-4074-01	···	TCGA-H7-A6C5-11
1	ENSG00000000003	TSPAN6	protein_coding	1224	8046	···	9577
2	ENSG00000000005	TNMD	protein_coding	0	0	···	0
3	ENSG00000000419	DPM1	protein_coding	1252	8073	···	1882
4	ENSG00000000457	SCYL3	protein_coding	208	654	···	874
5	ENSG00000000460	C1orf112	protein_coding	174	1500	···	146
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
19958	ENSG00000288661	AL451106.1	protein_coding	0	0	···	0
19959	ENSG00000288669	AC008763.4	protein_coding	0	1	···	0
19960	ENSG00000288671	AC006486.3	protein_coding	0	0	···	0
19961	ENSG00000288674	AL391628.1	protein_coding	3	2	···	5
19962	ENSG00000288675	AP006621.6	protein_coding	43	29	···	19

Previamente ao desenvolvimento do estudo, os dados clinicopatológicos foram sincronizados com a matriz de expressão de miRNA-Seq e mRNA-Seq, respectivamente. Por outro lado, pacientes sem informação clínica, miRNAs e mRNAs com níveis de expressão de zero em todas as amostras foram excluídos e mantidos aqueles com expressão em pelo menos uma amostra histologicamente normal adjacente ao tumor (NAT) diferente de zero para evitar falsos positivos e manter a variabilidade nesse grupo.

Após aplicação do filtro, para o primeiro banco de dados restaram 488 amostras no CCP e um total de 627 tipos de miRNAs para cada amostra. Já o segundo banco de dados apresentou 473 amostras, além de 16596 mRNAs (genes codificantes de proteínas) e 14 características clínicopatológicas para ambos casos, tais como: sexo, idade (anos), raça/cor da pele, malignidade anterior, consumo de álcool, diagnóstico primário (CCE, CCEF, CCEB, CCEQ e CCENQ), tamanho patológico do tumor primário (T: TX, T0 e T1-T4), linfonodos regionais patológicos (N: NX, N0 e N1-N3), ausência e presença de metástase patológica (M: MX, M0 e M1), estadiamento patológico do tumor (I, II, III, IVA, IVB e IVC) conforme determinado pelo sistema de estadiamento do AJCC da 4ª a 7ª edição, tipo de tratamento (farmacêutico ou radioterapia), estado vital (vivo ou morto) e sobrevivência geral (tempo em dias da observação do paciente desde o diagnóstico do tumor até a morte do paciente). Ademais, para a identificação da localização anatômica do tumor, utilizaram-se os códigos de Classificação Internacional de Doenças (CID10), sendo eles: cavidade oral (C01-C06; C14), orofaringe (C09-C10), hipofaringe (C12-C13) e laringe (C32).

5.2 Análise de expressão diferencial dos miRNAs e mRNAs

As análises estatísticas apenas se concentraram nos sítios anatômicos da cavidade oral e laringe devido a que o CCP inclui cânceres muito heterogêneos em termos de sua localização, sendo recomendado analisar separadamente.

Neste estudo, foram utilizados três métodos distintos para a análise de expressão diferencial: DESeq2 e edgeR, ambos baseados no modelo de regressão Binomial Negativa, além do teste de Wilcoxon-Mann-Whitney com normalização dos fatores de tamanho (s_j) utilizando edgeR. Estes métodos foram adotadas com o intuito de assegurar uma análise ampla e confiável na identificação de miRNAs e mRNAs desregulados no câncer da cavidade oral e laringe, proporcionando uma compreensão mais precisa e aprofundada das alterações genéticas associadas a essa doença.

a) **O modelo de regressão Binomial Negativa (DESeq2 e edgeR)** foi utilizado para reduzir a dimensionalidade dos dados, comparando os valores de expressão de cada miRNA e mRNA entre tumor primário (T) e o tecido histologicamente normal adjacente ao tumor (NAT) em ambos os sítios anatômicos (cavidade oral e laringe):

$$Y_{ij}\ NB(q_{ijk}, \sigma^2_{ij}). \tag{5.1}$$

em que Y_{ij} é matriz de expressão bruta (*counts*) para os $i = 1, 2, \cdots, 627$ tipos de miRNAs, $j = 1, 2, \cdots 321$ e $j = 1, 2, \cdots 112$ amostras da cavidade oral e laringe, respectivamente; e, para os $i = 1, 2, \cdots, 16596$ tipos de mRNA e $j = 1, 2, \cdots 313$ amostras de tecido referentes ao câncer da cavidade oral e $j = 1, 2, \cdots 108$ amostras de tecido referentes ao

câncer da laringe e q_{ijk} representa a verdadeira concentração de fragmentos para cada i na amostra j no grupo k ($k = \text{T}, \text{NAT}$).

Os miRNAs e mRNAs significativos foram identificados por meio dos pacotes `DESeq2` (LOVE; HUBER; ANDERS, 2014) e `edgeR` (ROBINSON; MCCARTHY; SMYTH, 2010) incorporado no pacote `GDCRNATools` (LI *et al.*, 2018) do `BiocManager` (GENTLEMAN *et al.*, 2004; HUBER *et al.*, 2015) e implementado no *software* `R` (R Core Team, 2022) que calcula automaticamente os fatores de normalização e estima a dependência entre a variância e a média com base na distribuição Binomial Negativa da equação (5.1).

As estimativas dos parâmetros do modelo relacionados à cavidade oral e laringe foram obtidas usando a função de ligação log-linear:

$$\log_2(q_{ijk}) = \beta_0 + \beta_1 x_{ijk}, \tag{5.2}$$

em que $\hat{\beta}_1$ é relativo para o $\log_2(FC)$ entre a média dos grupos do tumor primário e tecido normal adjacente, ou seja, $\hat{\beta}_1 = \log_2(FC) = \log_2\left(\frac{\text{média T}}{\text{média NAT}}\right)$. Assim, os DEmiRNAs e DEmRNAs na comparação entre os grupos foram identificados com o critério de FDR-BH $< 0,05$ de Benjamini e Hochberg (BH) e sob a condição:

$$\begin{cases} \log_2(FC) > 2 \rightarrow \text{superexpresso ou expressão aumentada}, \\ \log_2(FC) < -2 \rightarrow \text{subexpresso ou expressão diminuída}, \end{cases} \tag{5.3}$$

esse corte é comumente usado para identificar DEmiRNAs com um alto grau de confiança, pois sugere que a mudança na expressão é significativa e biologicamente relevante. No entanto, é importante ressaltar que o valor de corte para o $\log_2(FC)$ pode variar dependendo do contexto biológico e do objetivo do estudo. Em alguns casos, um $\log_2(FC)$ mais alto pode ser necessário para identificar miRNAs com mudanças mais robustas na expressão, enquanto em outros casos um $\log_2(FC)$ mais baixo pode ser suficiente para identificar miRNAs com mudanças mais sutis, mas ainda significativas.

Em seguida, os valores foram representados graficamente utilizando os gráficos de MA (*MA-plot*) e o vulcão (*volcano-plot*) com o auxílio do pacote `ggplot2` (WICKHAM; CHANG; WICKHAM, 2016).

Para interpretar os parâmetros do modelo Binomial Negativo em termos de risco relativo (RR), é necessário calcular a base 2 dos coeficientes da regressão. A fórmula para calcular o RR e o $\text{IC}_{95\%}$ de β_1 é:

$$\text{RR} = 2^{\hat{\beta}_1} \tag{5.4}$$

$$\text{IC}_{95\%} = \hat{\beta}_1 \pm 1,96 \times \widehat{EP}(\hat{\beta}_1) \tag{5.5}$$

se o RR for maior que 1, significa que a expressão do miRNA está aumentada nas amostras do T em comparação com NAT, o que indica que a expressão de miRNAs está associada

a um aumento no risco de câncer. Se o RR for menor que 1 a expressão de miRNAs está associada a uma diminuição no risco de desenvolver câncer. Se o RR for igual a 1, não há efeito do miRNA para o câncer. É importante lembrar que a interpretação do risco relativo depende do grupo de referência escolhido e que o risco relativo é apenas uma medida de associação, não de causalidade. Além disso, outras medidas, como o intervalo de confiança, devem ser consideradas para avaliar a significância estatística dos resultados.

b) **O Teste de Wilcoxon-Mann-Whitney** foi aplicado para comparar o T com o NAT para identificar possíveis diferenças significativas. Para isso, foram utilizados os critérios de Li *et al.* (2022a), em que os valores de expressão foram normalizados pelo TMM utilizando a normalização `edgeR` do pacote de Robinson, McCarthy e Smyth (2010), sendo conhecido como normalização `edgeR`. A função `wilcox_test` do pacote `coin` de Hothorn *et al.* (2017) foi empregada para obter DEmiRNAs e DEmRNAs, levando em consideração o $\log_2(FC)$, FDR-BH $< 0,05$ e o mesmo critério da equação (5.3). Além disso, este teste foi empregado para comparar a expressão dos DEmiRNAs nos estadiamentos iniciais e avançados associadas aos sítios anatômicos da cavidade oral e laringe.

   Neste estudo, depois de reduzir a dimensionalidade dos dados com os três métodos, optou-se por concentrar-se unicamente nos DEmiRNAs e DEmRNAs identificados através do DESeq2, devido à sua ampla aceitação e documentação robusta, além de adotar uma abordagem mais conservadora em comparação com outros métodos de análise diferencial de expressão gênica.

Para análises mais detalhadas, as expressões *count per million* (CPM) dos DEmiRNAs e DEmRNAs foram transformadas em logaritmo de base 2, somado a uma pseudo-contagem ($\log_2(CPM + 1)$), a fim de serem empregadas ao longo de toda a extensão do estudo. Essa conversão foi realizada devido à presença de valores extremos à direita e à heterocedasticidade presente nas contagens de expressão do miRNA-Seq e mRNA-Seq.

5.3 Análise de componentes principais e *Clustering*

Após a seleção de atributos DEmiRNAs e DEmRNAs utilizando o DESEq2, procedeu-se à análise de a identificar as componentes para garantir a retenção da separação de amostras e a clusterização para identificar padrões entre miRNAs e caraterísticas clinicopatológicas. Nesse processo, foi essencial padronizar a expressão ($\log_2(CPM + 1)$) de cada miRNA e mRNA (linhas) em relação às amostras (colunas) utilizando a função `scale` do *software* R Core Team (2022).

Dessa forma, para cada tipo de câncer mencionado anteriormente, a PCA foi realizada utilizando o pacote `FactoMineR` (HUSSON *et al.*, 2016) e `PCAtools` (BLIGHE *et al.*, 2019). Para estimar o número ótimo de grupos prévios ao *clustering*, foi utilizado o pacote `NbClust`

(CHARRAD *et al.*, 2014), considerando a distância Euclidiana e o método de ligação completa (exceto para DEmRNAs, nos quais foi utilizado o método de ligação Ward.D2) na identificação dos *clusters*. A definição dessa distância encontra-se definida na equação 4.29, em que $d(x_i, y_j)$ é a distância entre dois miRNAs ou mRNAs e n é o número de amostras.

Após a definição do número de grupos, foi utilizado o *clustering* hierárquico para agrupar as amostras, DEmiRNAs e DEmRNAs. Foram construídos mapas de calor (do inglês, *heatmap*) com a função `Heatmap` do pacote `complexheatmap` (GU, 2022), a fim de identificar os padrões de expressão dos genes entre as amostras. Além disso, foram construídos dendrogramas circulares para os DEmiRNAs e DEmRNAs usando a função `hclust` e o pacote `dendextend` (GALILI, 2015) com a finalidade de identificar a correlação existente entre eles.

Em seguida, foram realizadas as análises de correlação linear de Spearman (ρ) usando o pacote `corrplot` (WEI *et al.*, 2017) para avaliar a relação entre a expressão de diferentes miRNAs que apresentam diferenças no câncer da cavidade oral e laringe. De acordo com Al-Hameed (2022), o ρ é calculado por meio da fórmula:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d^2}{n(n^2-1)}, \ -1 < \rho < 1, \tag{5.6}$$

em que d é a diferença de classificação entre pares de observações correspondentes nas duas variáveis e n é número total de pares de observações. Quando ρ é igual a 1, há uma correlação linear positiva perfeita entre os miRNAs, o que indica que os miRNAs poderiam estar envolvidos em processos metabólicos semelhantes. Por outro lado, quando ρ se aproxima de -1, surge uma relação inversamente proporcional em sua expressão e quando ρ se aproxima de 0, isso indica uma correlação linear fraca ou ausência de relação aparente entre os miRNAs (LI *et al.*, 2018).

5.4 Validação da expressão dos miRNAs com fontes externas

Com a união dos DEmiRNAs candidatos resultantes da análise DESeq2, edgeR e teste de Wilcoxon-Mann-Whitney para o câncer da cavidade oral e laringe foram avaliados utilizando as ferramentas *web* dbDMEC (v. 3.0) (*A Database of Differentially Expressed miRNAs in Human Cancers*, disponível em: <https://www.biosino.org/dbDEMC/index>), desenvolvida por Xu *et al.* (2022), e oncomiR (WONG *et al.*, 2018) (disponível em: <https://oncomir.org/>). O dbDMEC é uma ferramenta abrangente que aborda os DEmiRNAs associados a diversos tipos de cânceres, incluindo os tipos de CCP, com base em dados de *microarrays* ou miRNA-seq provenientes de repositórios públicos como o GEO, *Sequence Read Archive (SRA)*, *ArrayExpress* e o TCGA. Já o oncomiR é uma plataforma *online* que utiliza dados integrados de miRNA-seq, RNA-seq e informações clínicas do TCGA para identificar miRNAs desregulados associados ao desenvolvimento e progressão de tumores na maioria dos principais tipos de câncer.

Na Tabela 8, são apresentados os conjuntos de dados contidos no dbDMEC (v. 3.0) que foram utilizados na validação dos DEmiRNAs para o câncer da cavidade oral e laringe.

Tabela 8 – Conjuntos de dados destinados à validação dos DEmiRNAs presentes no banco de dados dbDMEC (versão 3.0) (XU *et al.*, 2022).

Experiment ID	GEO ID	*Instrument*	*Cancer type*	*Sample size*	Nº of miRNAs	*Software*
EXP00281	GSE45238	Illumina Human v2 MicroRNA expression beadchip	Oral squamous cell carcinoma (OSCC)	40 OSCC 40 normal tissue	846	Limma
EXP00177	GSE28100	Agilent-021827 Human miRNA Microarray (V3) (miRBase release 12.0 miRNA ID version)	Oral squamous cell carcinoma (OSCC)	17 OSCC 3 healthy control tongue	851	Limma
EXP00491	GSE98463	[miRNA-4] Affymetrix Multispecies miRNA-4 Array [ProbeSet ID version]	Oral squamous cell carcinoma (OSCC)	8 OSCC 8 normal oral mucosa	831	Limma
EXP00560	GSE137865	[miRNA-1] Affymetrix Multispecies miRNA-1 Array	Oral squamous cell carcinoma (OSCC)	6 OSCC 12 tissues adjacent to cancer and normal tissues	180	Limma
EXP00558	GSE124678	Agilent-031181 Unrestricted_Human_miRNA_V16.0_Microarray (miRBase release 16.0 miRNA ID version)	Laryngeal squamous cell carcinoma (LSCC)	32 LSCC 5 normal laryngeal tissue	1205	Limma
EXP00578	GSE116994	[miRNA-4] Affymetrix Multispecies miRNA-4 Array	Larynx cancer	5 laryngeal carcinoma tissue 5 normal tissue	831	Limma

Além disso, foram consultadas outras fontes de referência, como o estudo realizado por Falzone *et al.* (2019) sobre câncer oral, no qual foram empregados conjuntos de dados GEO (GSE45238 e GSE31277) para a identificação de DEmiRNAs. Esse processo foi conduzido por meio do *software* GEO2R, e também utilizando o TCGA para identificar DEmiRNAs por meio do teste t de Student. No contexto do câncer de laringe, recorremos ao trabalho de Luo *et al.* (2022), que utilizou o conjunto de dados GEO (GSE133632) para identificar DEmiRNAs. Este estudo empregou 57 amostras de carcinoma espinocelular de laringe (LSCC) e 57 amostras não tumorais, utilizando o pacote `limma` desenvolvido por Ritchie *et al.* (2015).

5.5 Análise de enriquecimento e anotação funcional de miRNAs

O DIANA-miRPath (v.4.0) (TASTSOGLOU *et al.*, 2023) foi usado para identificar as principais vias de sinalização KEGG e anotação GO que foram afetadas pelos DEmiRNAs, considerando um valor estatisticamente significativo de *FDR-BH* $< 0,05$.

5.6 Predição de alvos, redes de interação miRNA-mRNA e análises de enriquecimento GO e KEGG de mRNAs

Os genes alvo dos DEmiRNAs foram previamente identificados por meio de abordagens integrativas, utilizando duas ferramentas principais: o *microRNA Data Integration Portal* (mirDIP, disponível em <https://ophid.utoronto.ca/mirDIP/>), com ênfase na classe de pontuação *Very High*, e a base de dados de interações miRNA-alvo validadas experimentalmente, conhecida

como *The experimentally validated miRNA-target interactions database* (mirTarBase, acessível em <https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/index.php>).

Seguidamente, esses genes alvo foram sobrepostos com os DEmRNAs obtidos da análise de mRNA-Seq por meio de um diagrama de Venn, utilizando o pacote `ggVennDiagram` (GAO; YU; CAI, 2021). A seleção final dos genes alvo candidatos foi baseada na condição de subexpressão quando o miRNA estava superexpresso e superexpressão quando os miRNAs estavam subexpressos. Os genes escolhidos foram então utilizados para construir redes de interação regulatória entre miRNA e mRNA utilizando o *software* `Cytoscape` (versão 3.10.1) (OTASEK *et al.*, 2019). Na visualização das interações, os valores de $\log_2(FC)$ foram considerados, destacando que o aumento da expressão do miRNA pode estar associado à redução da expressão do gene alvo, e vice-versa.

Posteriormente, a fim de investigar as potenciais funções dos DEmRNAs associados à progressão do câncer na cavidade oral e laringe, conduziu-se uma análise de enriquecimento funcional. Inicialmente, empregou-se o pacote de conversão `org.Hs.eg.db` (CARLSON *et al.*, 2019) para a anotação dos genes alvo candidatos. A análise de enriquecimento GO foi desmembrada em três categorias: Processo Biológico (BP), Função Molecular (MF) e Componente Celular (CC), sendo avaliada por meio da função `richGO` do pacote `clusterProfiler` (WU *et al.*, 2021). Para a investigação do enriquecimento das vias KEGG, recorreu-se à função `richKEGG`, enquanto o pacote `DOSE`, em conjunto com a função `richDO` do `clusterProfiler`, foi empregado para analisar o enriquecimento das ontologias de doenças com base no banco de dados *Disease Ontology* (DO). Em todas as análises, adotou-se o critério de *FDR-BH* $< 0,05$ para determinar a significância estatística, e os resultados foram representados graficamente por meio do pacote `ggplot2` (WICKHAM; CHANG; WICKHAM, 2016).

5.7 Rede de interação proteína-proteína, identificação de *hub* genes, rede regulatória miRNA-*hub* e análise de enriquecimento de vias KEGG

Para avaliar as interações entre proteínas (PPI) de genes alvo candidatos no nível proteico, utilizou-se o banco de dados *online* `STRING` (*Search Tool for the Retrieval of Interacting Genes/Proteins*, disponível em <http://www.string-db.org/>). Foi considerado um *score* de interação $\geq$ 0,4 como critério para a seleção das interações PPI relevantes. A visualização dessas interações foi realizada por meio do *software* `Cytoscape (v.3.10.1)` (OTASEK *et al.*, 2019). Posteriormente, o `CytoHubba` (CHIN *et al.*, 2014), um *plug-in* do `Cytoscape` (v.3.10.1) (OTASEK *et al.*, 2019), foi empregado para identificar genes altamente conectados na rede, conhecidos como *hubs*. Essa identificação baseou-se na *Maximal Clique Centrality* (MCC), considerando como significativos os genes com os 10 maiores valores de MCC e um

FDR-BH $< 0,05$. A seguir, a rede miRNA-*hub* foi construída utilizando os 10 genes *hub* para ambos os casos. Realizou-se a análise das vias KEGG com os *hub* genes utilizando a plataforma *online* `ShinyGO (v.0.77)` (GE; JUNG; YAO, 2020).

5.8 Algoritmos de classificação baseadas em árvores de decisão e florestas aleatórias

A capacidade discriminatória dos tecidos para cada DEmiRNA presente nas redes de interação miRNA-mRNA associadas ao câncer da cavidade oral e laringe através da construção da curva ROC e AUC, juntamente com os intervalos de confiança de 95%, sensibilidade, especificidade e identificação do ponto de corte ideal. Considerou-se como biomarcadores de diagnóstico os miRNAs que demonstraram AUC, sensibilidade e especificidade superiores a 0,6. Em seguida, foram gerados mapas de calor com base em árvores de decisão para fornecer uma visualização abrangente dos padrões de expressão desses miRNAs na classificação das amostras desbalanceadas de tumor primário e tecido normal adjacente. Esse procedimento foi realizado utilizando o pacote `treeheatr`, desenvolvido por Le e Moore (2021).

Seguidamente, realizou-se a classificação supervisionada utilizando árvores de decisão e o RFT (sem a incorporação de características clinicopatológicas) apenas com os miRNAs e mRNAs presentes na rede de interação regulatórias. O objetivo foi investigar se as expressões combinadas dos miRNAs poderiam funcionar como um biomarcador mais eficaz no diagnóstico do câncer da cavidade oral e laringe, abrangendo todos os estadiamentos patológicos (I - IV), iniciais (I - II) e avançados (III - IV). Para atingir esse propósito, os dados foram inicialmente balanceados por meio de quatro algoritmos: *undersampling* utilizando a função `downSample`, *upsampling* utilizando a função `upSample`, ambos implementados no pacote `caret` (KUHN *et al.*, 2020), SMOTE no pacote `DMwR` (TORGO; TORGO, 2013) e ROSE utilizando o pacote `ROSE` (LUNARDON; MENARDI; TORELLI, 2014) devido à predominância de amostras de tumor primário em comparação com amostras de tecido normal adjacente, presente em ambos os conjuntos de dados.

Posteriormente, para cada conjunto de dados balanceados, aplicou-se o método *Holdout* por meio da função `createDataPartition` do pacote `caret` (KUHN *et al.*, 2020), alocando 75% da amostra para treinamento e 25% para teste, com o intuito de avaliar o desempenho do modelo. Para cada conjunto de treinamento, os algoritmos `CTree` e RFT com geração de 500 árvores, foi aplicado utilizando validação cruzada de *5-fold cross validation*, repetida 10 vezes para validar os modelos e buscar os hiperparâmetros ideais. A execução desse procedimento foi realizada por meio da função `train` do pacote `caret` (KUHN *et al.*, 2020).

Os resultados da árvore de decisão, obtidos através do algoritmo CTree, foram visualizados utilizando os pacotes `partykit` (HOTHORN; ZEILEIS; HOTHORN, 2023) e `ggparty` (BORKOVEC; MADIN, 2019). Adicionalmente, as informações sobre a importância dos miRNAs

na discriminação dos tecidos, obtidas pelo algoritmo RFT, foram visualizadas utilizando o pacote `ggplot2` (WICKHAM; CHANG; WICKHAM, 2016). A avaliação dos modelos foi conduzida por meio da matriz de confusão, utilizando métricas como AUC, especificidade, sensibilidade, acurácia, precisão, pontuação F1 e Capa de Cohen.

5.9 Análise de sobrevivência

Finalmente, a análise de sobrevivência foi conduzida considerando as características clinicopatológicas dos pacientes, focalizando exclusivamente nos miRNAs presentes na rede de interação miRNA-mRNA associada ao câncer da cavidade oral e laringe. Para isso, foram empregados métodos como Kaplan-Meier, teste de Log-rank, modelo de riscos proporcionais de Cox com a verificação da suposição de proporcionalidade dos riscos, árvores de sobrevivência e RSF.

No início desta análise, procedeu-se à comparação das curvas de sobrevivência Kaplan-Meier entre os pacientes diagnosticados com câncer na cavidade oral e laringe. Em seguida, o conjunto de dados referente ao câncer na cavidade oral foi aleatoriamente subdividido em conjuntos de treinamento (80%, 70% e 80%) e teste (20%, 30% e 20%) para todos os estadiamentos patológicos (I - IV), iniciais (I - II) e avançados (III - IV), respectivamente. Em relação ao câncer da laringe, a divisão aleatória foi realizada em conjuntos de treinamento (85% e 80%) e teste (15% e 20%) para todos os estadiamentos patológicos (I - IV) e avançados (III - IV), respectivamente, no contexto das árvores de sobrevivência e RSF é relevante destacar que, para o câncer da laringe, o modelo de Cox foi avaliado considerando todos os estadiamentos, sem a divisão do banco de dados, devido à sua propensão em manter a suposição de riscos proporcionais ao considerar um nível de significância de 5%.

A categorização dos níveis de expressão de cada miRNA foi realizada com base na mediana dos níveis de expressão desses miRNAs nas amostras. Aqueles acima da mediana foram considerados como expressão alta ($\geq$ mediana), enquanto aqueles abaixo foram considerados como expressão baixa ($<$ mediana). Essa categorização foi avaliada exclusivamente por meio da análise de regressão de riscos proporcionais de Cox univariado e multivariado no conjunto de treinamento. Na regressão multivariada de riscos proporcionais de Cox, a função `stepAIC` do pacote `MASS` (RIPLEY *et al.*, 2013) foi empregada para a seleção dos miRNAs a serem incluídos no modelo. Além disso, as características clinicopatológicas foram avaliadas apenas por meio de análises de regressão de riscos proporcionais de Cox univariado e multivariado no conjunto de treinamento, considerando valores de p inferiores a 0,05 como significativos.

Os gráficos de Kaplan-Meier e o ajuste do modelo de riscos proporcionais de Cox foram realizados utilizando os pacotes `survival` (THERNEAU; LUMLEY, 2015) e `survminer` (KASSAMBARA *et al.*, 2017), respectivamente. As árvores de sobrevivência podadas foram construídas e visualizadas utilizando o algoritmo CART© por meio dos pacotes `rpart`

(THERNEAU *et al.*, 2022), `partykit` (HOTHORN; ZEILEIS; HOTHORN, 2023), e a função `fancyRpartPlot` do pacote `rattle` (WILLIAMS, 2011). A floresta aleatória foi conduzido utilizando o pacote `randomforestsrc` (ISHWARAN; KOGALUR; KOGALUR, 2023) e gerando 30000 árvores para estabilizar a taxa de erro, e a importância dos miRNAs foi visualizada com o pacote `ggplot2` (WICKHAM; CHANG; WICKHAM, 2016). A avaliação do desempenho dos três modelos foi realizada por meio do C-index utilizando os conjuntos de teste.

6 Resultados

Neste capítulo, são apresentadas os resultados das análises descritivas das variáveis clinicopatológicas e expressão de miRNAs e mRNAs, bem como as análises inferenciais por meio da Bioinformática envolvendo métodos clássicos e algoritmos de aprendizagem de máquina, baseadas em árvores de decisão e análise de sobrevivência.

6.1 Análise preliminar dos dados

6.1.1 Características clinicopatológicas

Na Tabela 9, encontram-se as estatísticas descritivas das características clinicopatológicas. Nota-se que o CCP consistiu em 488 amostras, sendo a cavidade oral o local anatômico mais prevalente (321 casos, representando 65,78% do total), seguido pela laringe (112 casos, 22,95%), orofaringe (47 casos, 9,63%) e hipofaringe (8 casos, 1,64%). Além disso, as amostras NAT do TCGA são numericamente inferiores às dos tumores primários (Tabela 9 e Figura 11 A), conforme evidenciado em vários estudos, provavelmente porque o projeto do TCGA tinha como foco principal o estudo do câncer, e, portanto, coletava e analisava principalmente amostras tumorais.

En relação ao sítio anatômico da cavidade oral, no momento do diagnóstico, 293 dos indivíduos (91%) apresentavam amostras de tumor primário, 48% foram a óbito, 66% eram do sexo masculino e mais de 86% brancos. Além disso, 62,93% eram etilistas e 86,29% foram diagnosticados com CEC, 73,84% encontravam-se em estadiamentos patológicos avançados (III e IV) e foram tratados com radioterapia (53,58%). Já na laringe, observam-se 100 amostras de tumor primário, 15,18% eram do sexo feminino, 61% eram alcoólatras e apenas 15% foram diagnosticados em estadiamentos iniciais da doença (I e II).

Na Figura 11 (B), pode-se observar que o câncer de cavidade oral pode ter um impacto em várias áreas da boca, a língua é uma das regiões mais afetadas pelo câncer da cavidade oral, representando 40,27% dos casos. Em seguida, a lesão sobreposta dos lábios é a segunda área mais afetada, com 22,18%, seguida pelo assoalho da boca, com 7,17%. Outras regiões da boca, como gengivas, palato e mucosa da bochecha, têm uma menor frequência de incidência.

Tabela 9 – Frequência absoluta (f_a) e percentual (f_p) das características clinicopatológicas dos pacientes com câncer de cabeça e pescoço (CCP) em cada sítio anatômico da doença.

Características	Categorias	CCP		Sítio anatômico							
				Cavidade oral		Orofaringe		Hipofaringe		Laringe	
		$n = 488$		$n = 321$		$n = 47$		$n = 8$		$n = 112$	
		f_a	f_p (%)	f_a	f_p (%)	f_a	f_p (%)	f_a	f_p (%)	f_a	f_p (%)
Idade (anos)	Média (DP)	61,12 (12,15)		61,77 (13,23)		54,81 (9,39)		61,88 (5,33)		61,88 (9,25)	
Tempo (dias)	Média (DP)	923,20 (889,68)		900,00 (885,91)		798,10 (570,14)		588,60 (663,29)		1066,20 (1005,41)	
Tecido	NAT	42	8,61	28	8,72	2	4,26	0	0,00	12	10,71
	T	446	91,39	293	91,28	45	95,74	8	100,00	100	89,29
Estado vital	Vivo	268	54,92	166	51,71	32	68,09	5	62,50	65	58,04
	Óbito	220	45,08	155	48,29	15	31,91	3	37,50	47	41,96
Sexo	Feminino	138	28,28	110	34,27	8	17,02	3	37,50	17	15,18
	Masculino	350	71,72	211	65,73	39	82,98	5	62,50	95	84,82
Raça/Cor	Branco	418	85,66	278	86,60	44	93,62	7	87,50	89	79,46
	Preto	42	8,61	23	7,16	3	6,38	1	12,50	15	13,39
	Outros*	13	2,66	10	3,12	0	0,00	0	0,00	3	2,68
	Não definido	15	3,07	10	3,12	0	0,00	0	0,00	5	4,46
Alcoolismo	Sim	309	63,32	202	62,93	34	72,34	5	62,50	68	60,71
	Não	171	35,04	115	35,83	11	3,40	3	37,50	2	1,79
	Não definido	8	1,64	4	1,25	2	4,26	0	0,00	42	37,50
Malignidade anterior	Sim	19	3,89	14	4,36	0	0,00	0	0,00	5	4,46
	Não	469	96,11	307	95,64	47	100,00	8	100,00	107	95,54
Histologia	CCE	415	85,04	277	86,29	33	70,21	7	7,50	98	87,50
	CCEQ	54	11,07	41	12,77	2	4,26	0	0,00	11	9,82
	CCENQ	9	1,84	0	0,00	7	14,89	1	12,50	1	0,89
	CCEF	1	0,20	0	0,00	1	2,13	0	0,00	0	0,00
	CCEB	9	1,84	3	0,93	4	8,51	0	0,00	2	1,79
Estadiamento patológico	I	29	5,94	23	7,17	4	8,51	0	0,00	2	1,79
	II	88	18,03	61	19,00	11	23,40	0	0,00	16	14,29
	III	87	17,83	63	19,63	9	19,15	1	12,50	14	12,50
	IVA/IVB/IVC	284	58,20	174	54,21	23	48,94	7	87,50	80	71,43
Estadiamento patológico T	T1	45	9,22	32	9,97	8	17,02	0	0,00	5	4,46
	T2	151	30,94	106	33,02	25	53,19	0	0,00	20	17,86
	T3	107	21,93	65	20,25	11	23,40	3	37,50	28	25,00
	T4	185	37,91	118	36,76	3	6,38	5	62,50	59	52,68
Estadiamento patológico N	NX	44	9,02	28	8,72	4	8,51	1	12,50	11	9,82
	N0	188	38,52	128	39,88	14	29,79	1	12,50	45	40,18
	N1	73	14,96	54	16,82	6	12,77	1	12,50	12	10,71
	N2	175	35,86	9	2,80	2	4,26	0	0,00	0	0,00
	N3	8	1,64	28	8,72	1	2,13	4	50,00	11	9,82
Estadiamento patológico M	MX	61	12,50	37	11,53	15	31,91	1	12,50	8	7,14
	M0	183	37,50	119	37,07	19	40,43	6	75,00	39	34,82
	M1	1	0,20	0	0,00	0	0,00	0	0,00	1	0,89
	Não definido	243	49,80	165	51,40	13	27,66	1	12,50	64	57,14

Nota: Outros* - Índio americano ou nativo do Alasca e asiático.

Na Figura 12 (A) observa-se que a maioria dos pacientes foi diagnosticada com CCP em estadiamento patológico IVA/IVB, com idade média de 60 anos, independentemente do local do tumor. Na Figura 12 (B), os resultados indicam que a maioria dos pacientes do sexo feminino e masculino está concentrada na faixa etária de 50 a 80 anos.

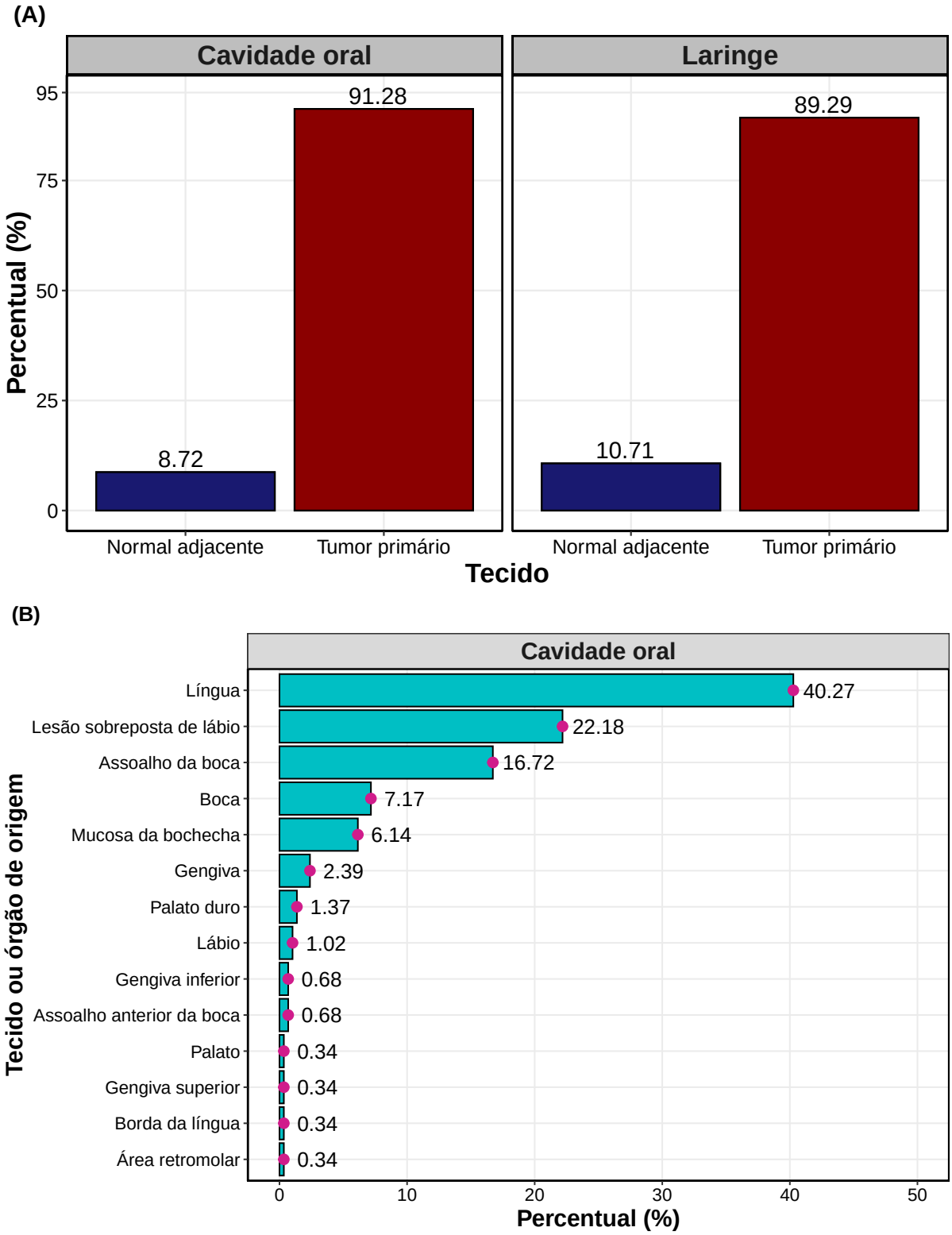


Figura 11 – (A) Gráfico de barras referentes ao tecido normal adjacente e tumor primário nos sítios anatômicos do CCP. (B) Gráfico de barras referentes ao órgão de origem do câncer da cavidade oral.

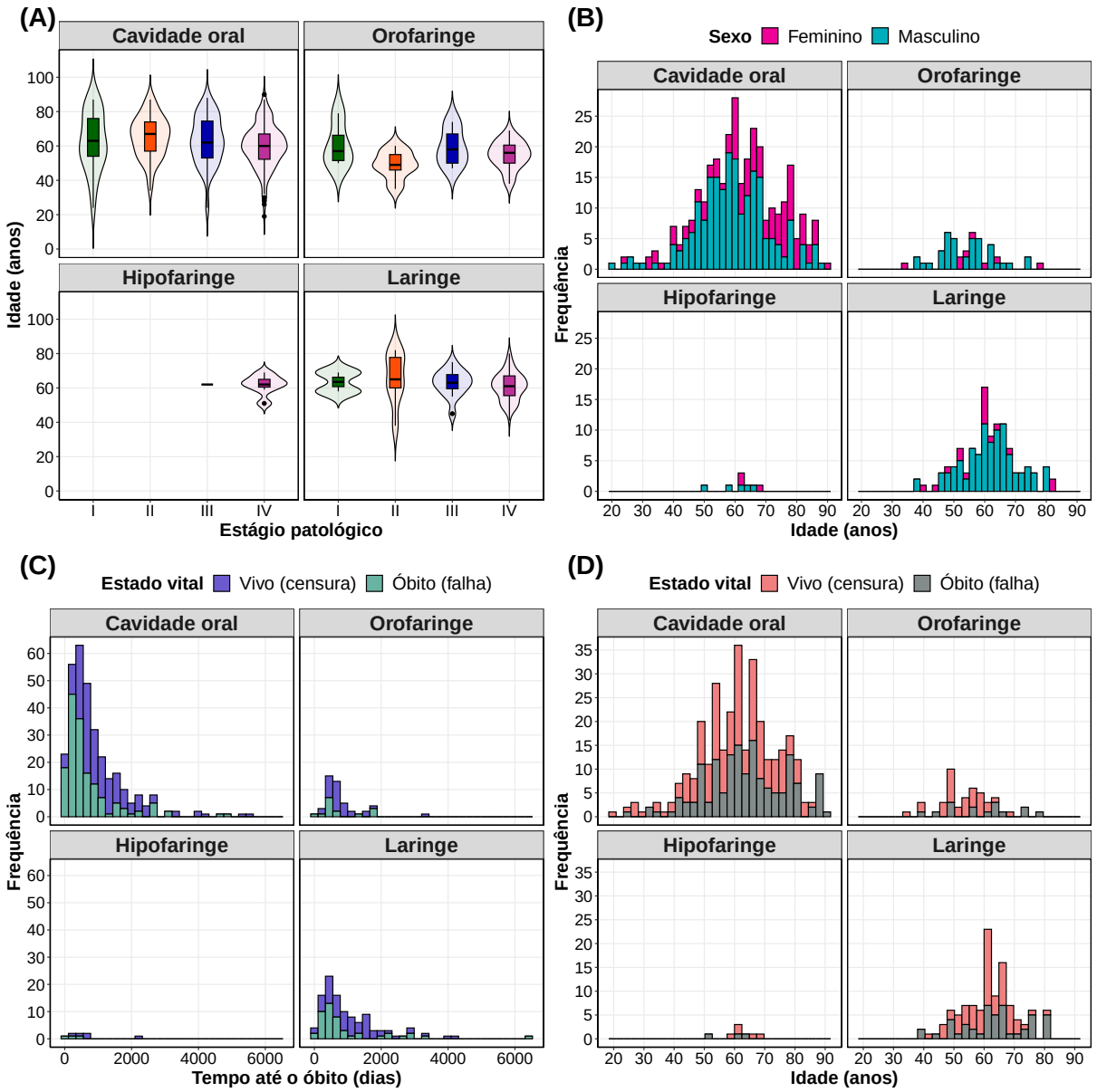


Figura 12 – Gráficos bivariados das características clinicopatológicas. (A) *Boxplots* das idades (anos) dos pacientes por estadiamento patológico. (B) Histograma das idades dos pacientes estratificado por sexo. (C) Histograma dos tempos de censura e falha. (D) Histograma das idades (anos) por estado vital dos pacientes nos sítios anatômicos do câncer de cabeça e pescoço.

Na Figura 12 (C), a assimetria positiva é observada no tempo até o óbito, com a maioria dos eventos de falha ocorrendo em até aproximadamente 3000 dias e entre 100 e 500 dias há o maior número de pacientes que não atingiram o evento de interesse (óbito). Na Figura 12 (D), a maioria dos pacientes ainda estavam vivos no final do estudo, concentrando-se na faixa etária de 50 a 80 anos nos locais do tumor da cavidade oral, orofaringe e laringe.

Além disso, com o teste de Qui-Quadrado apenas foi observada uma associação estatisticamente significativa entre o tecido tumoral na cavidade oral e o estado vital dos pacientes.

No caso do câncer de laringe, foram identificadas associações com o estadiamento patológico, T e N (Tabela 16 do Apêndice A.1).

6.1.2 Perfil de expressão de miRNAs e mRNAs

Nas Figuras 13 (A) e (B), os histogramas dos níveis de expressão de vários genes (miRNAs e mRNAs) associados à cavidade oral e à laringe são exibidos, respectivamente. Nota-se uma assimetria à direita e uma alta variabilidade nos níveis de expressão dos genes.

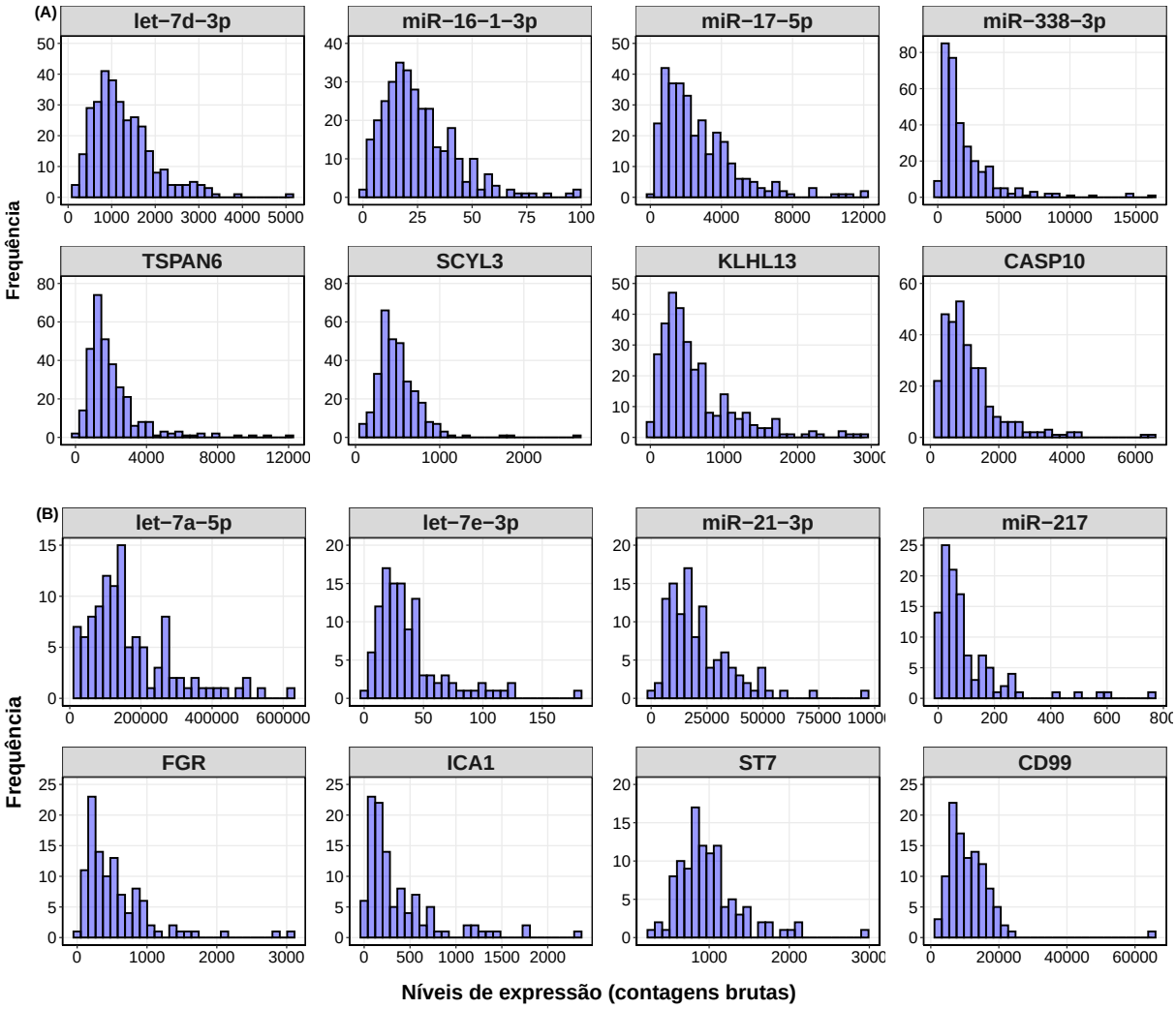
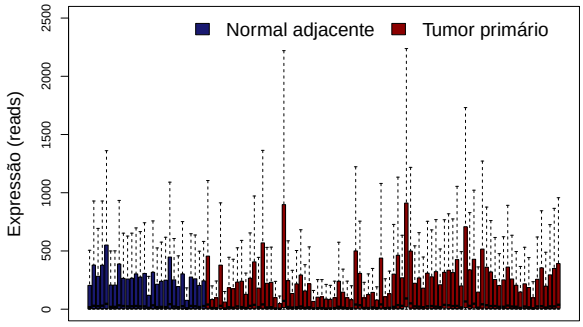
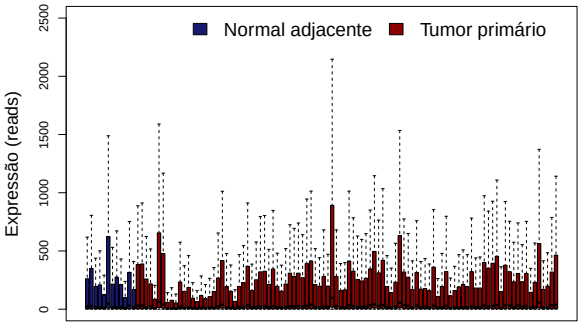


Figura 13 – Distribuição da expressão de oito miRNAs e oito mRNAs nos sítios anatômicos da cavidade oral (A) e laringe (B), respetivamente.

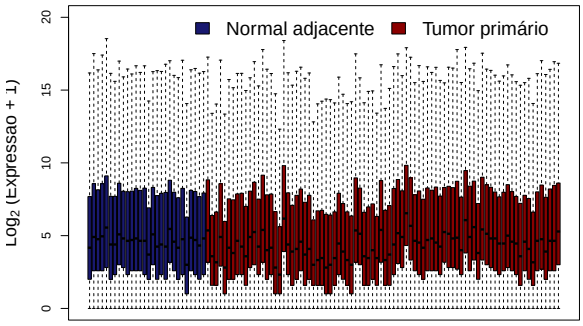
Adicionalmente, nas Figuras 14 (a) e (b), são apresentados os *boxplots* para cada amostra de tecido normal adjacente e tumor primário, abrangendo a expressão dos miRNAs relacionados à cavidade oral e à laringe. As Figuras 14 (c) e (d) ilustram os mesmos *boxplots*, mas com uma transformação logarítmica ($\log_2$), enquanto as Figuras 14 (e) e (f) os apresenta após a normalização das bibliotecas pelo DESeq2, demonstrando a estabilização da variância.



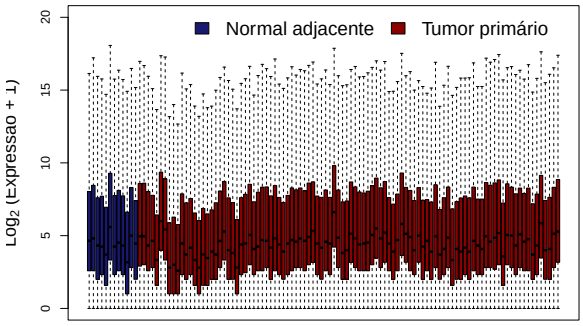
(a) Expressão dos miRNAs nas amostras da cavidade oral.



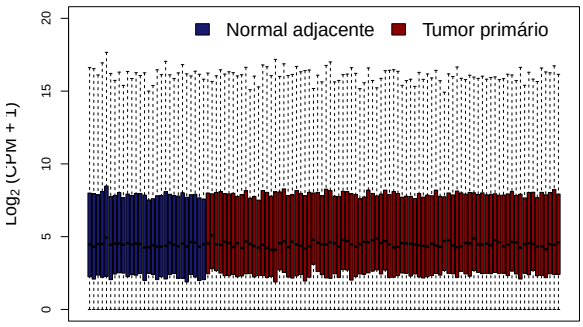
(b) Expressão dos miRNAs nas amostras da laringe.



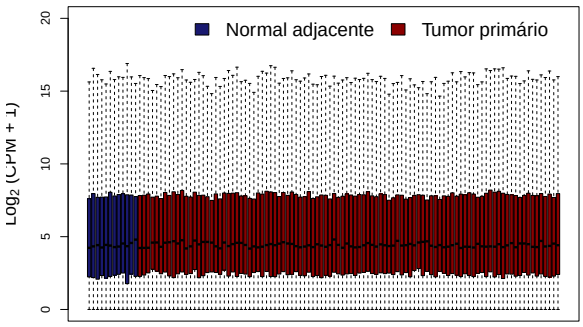
(c) Expressão dos miRNAs nas amostras da cavidade oral, sem normalização das bibliotecas utilizando o método da mediana das razões em log_2 por meio do DESeq2.



(d) Expressão dos miRNAs nas amostras da cavidade oral, sem normalização das bibliotecas utilizando o método da mediana das razões em log_2 por meio do DESeq2.



(e) Expressão dos miRNAs nas amostras da cavidade oral, com normalização das bibliotecas utilizando o método da mediana das razões em log_2 por meio do DESeq2.



(f) Expressão dos miRNAs nas amostras da laringe, com normalização das bibliotecas utilizando o método da mediana das razões em log_2 por meio do DESeq2.

Figura 14 – *Boxplots* da expressão dos miRNAs para os primeiros 112 amostras da cavidade oral (a, c, e) e laringe (b, d, f), respectivamente.

6.2 Análise da expressão diferencial dos miRNAs

6.2.1 Identificação de DEmiRNAs no câncer da cavidade oral

Utilizando o DESEq2, foi possível identificar um total de 383 DEmiRNAs no câncer da cavidade oral, conforme ilustrado na Figura 15 (A). Após a aplicação do critério ($\hat{\beta}_1 = |\log_2(FC)| > 2$ e FDR-BH $< 0{,}05$), identificaram-se 58 DEmiRNAs, dos quais 27 foram superexpressos ou oncogênicos e 31 foram subexpressos ou supressores de tumor em 293 amostras de tumor primário, comparadas com 28 tecidos normais adjacentes ao tumor (Figura 15 (B) e Tabela 17 do Apêndice A.2).

Dentre os 10 principais DEmiRNAs (valores FDR-BH muito baixos), os miRNAs com regulação positiva mais significativa foram miR-455-3p, miR-196b-5p, miR-196a-5p, miR-503-5p e miR-210-3p, enquanto os miR-101-3p, miR-381-3p, miR-30a-3p, miR-30a-5p e miR-378a-5p apresentaram regulação negativa (Figuras 15 C).

A Figura 15 (D) mostra aos miRNAs com maior atividade oncogênica que incluem miR-105-5p (RR = 46,51), miR-767-5p (RR = 38,33), miR-615-3p (RR = 19,57), miR-1910-5p (RR = 19,48) e miR-1269a (RR = 18,65). Por outro lado, miR-381-3p (RR = 0,18), miR-378d (RR = 0,14), miR-30a-3p (RR = 0,14), miR-1-3p (RR = 0,12) e miR-375 (RR = 0,06) demonstraram maior atividade supressora. Esses resultados indicam que cada miRNA pode contribuir de forma positiva ou negativa para o fenótipo tumoral, com a maioria atuando predominantemente como supressores de tumor.

Pelo edgeR e o teste de Wilcoxon-Mann-Whitney, foram identificadas 58 DEmiRNAs (27 superexpressos e 31 subexpressos, conforme detalhado na Tabela 19 do Apêndice A.3) e 60 DEmiRNAs (28 superexpressos e 32 subexpressos, Tabela 20 do Apêndice A), respectivamente.

Dentre os 10 principais DEmiRNAs, é notável que os cinco miRNAs superexpressos (miR-1910-5p, miR-615-3p, miR-196b-5p, miR-455-3p e miR-503-5p) foram os mais significativos em ambas abordagens. No entanto, o miR-101-3p destaca-se como o DEmiRNA subexpresso mais significativo em ambos os métodos (Figuras 16 A e B).

A sobreposição dos DEmiRNAs obtidos pelos três métodos é apresentada na Figura 16 (C), demonstrando que 54 DEmiRNAs (84,4%) foram detectados por ambos os métodos. Embora o DESeq2 e o edgeR tenham identificado a mesma quantidade total de DEmiRNAs, apenas 55 DEmiRNAs foram comuns entre eles. Notavelmente, todos os DEmiRNAs identificados pelo edgeR foram também detectados tanto pelo DESeq2 quanto pelo teste de Wilcoxon-Mann-Whitney.

No entanto, é importante ressaltar que esses resultados indicaram uma quantidade semelhante de miRNAs quando comparados sob as mesmas condições. Caso contrário, os resultados poderiam ser divergentes tanto para o câncer de cavidade oral e laringe quanto para outros tipos de câncer.

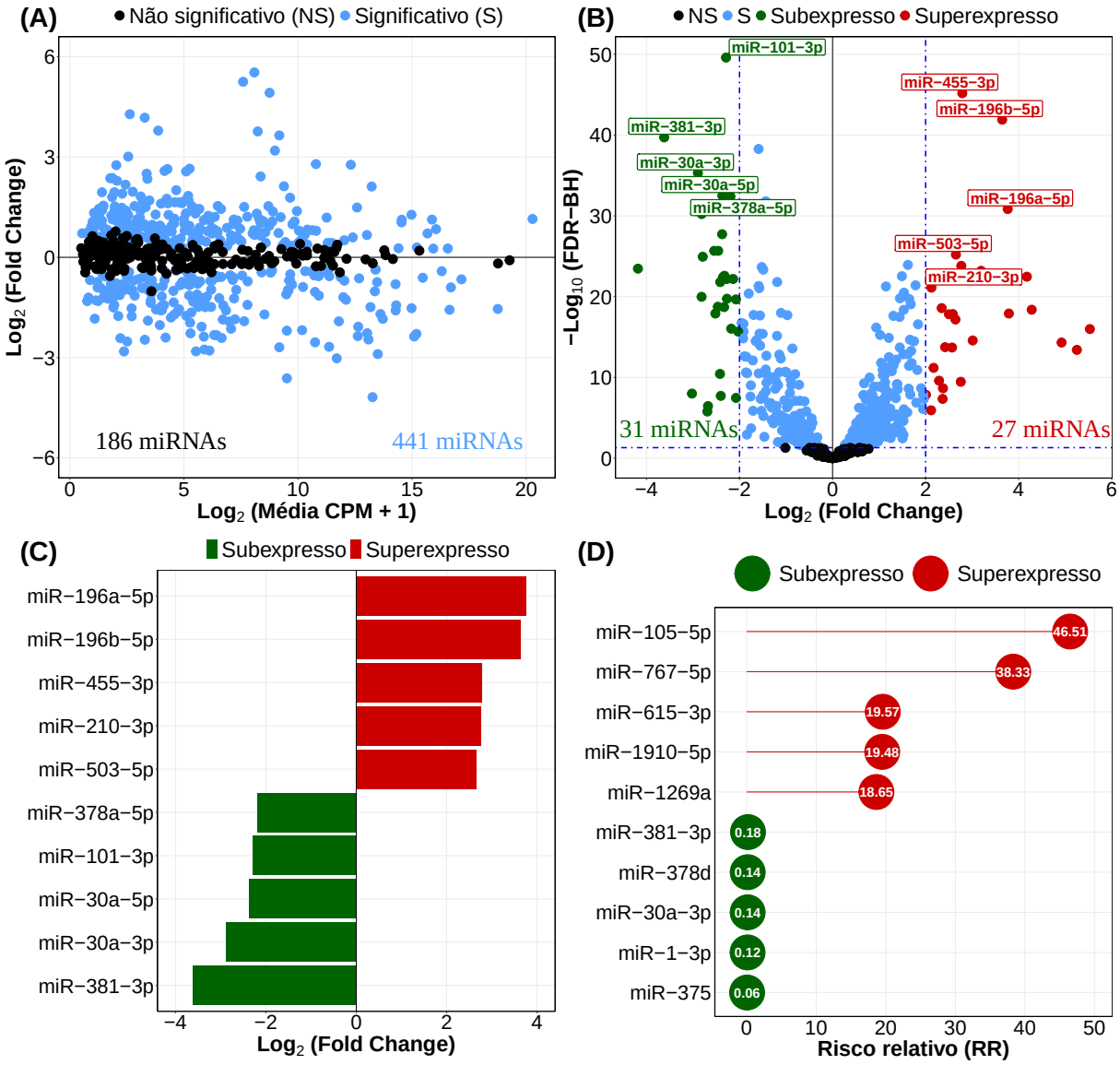


Figura 15 – Expressão Diferencial de miRNAs em comparação entre tecidos tumorais primários e normais adjacentes relacionados ao câncer de cavidade oral. (A) O gráfico MA da razão logarítmica (M) versus a média (A) das diferenças entre os dois grupos para cada miRNA. (B) Gráfico de Vulcão dos DEmiRNAs. (C) Gráfico de Barras com os 10 DEmiRNAs mais desregulados. (D) Gráfico de pirulitos divergentes com os 10 DEmiRNAs em relação ao risco relativo.

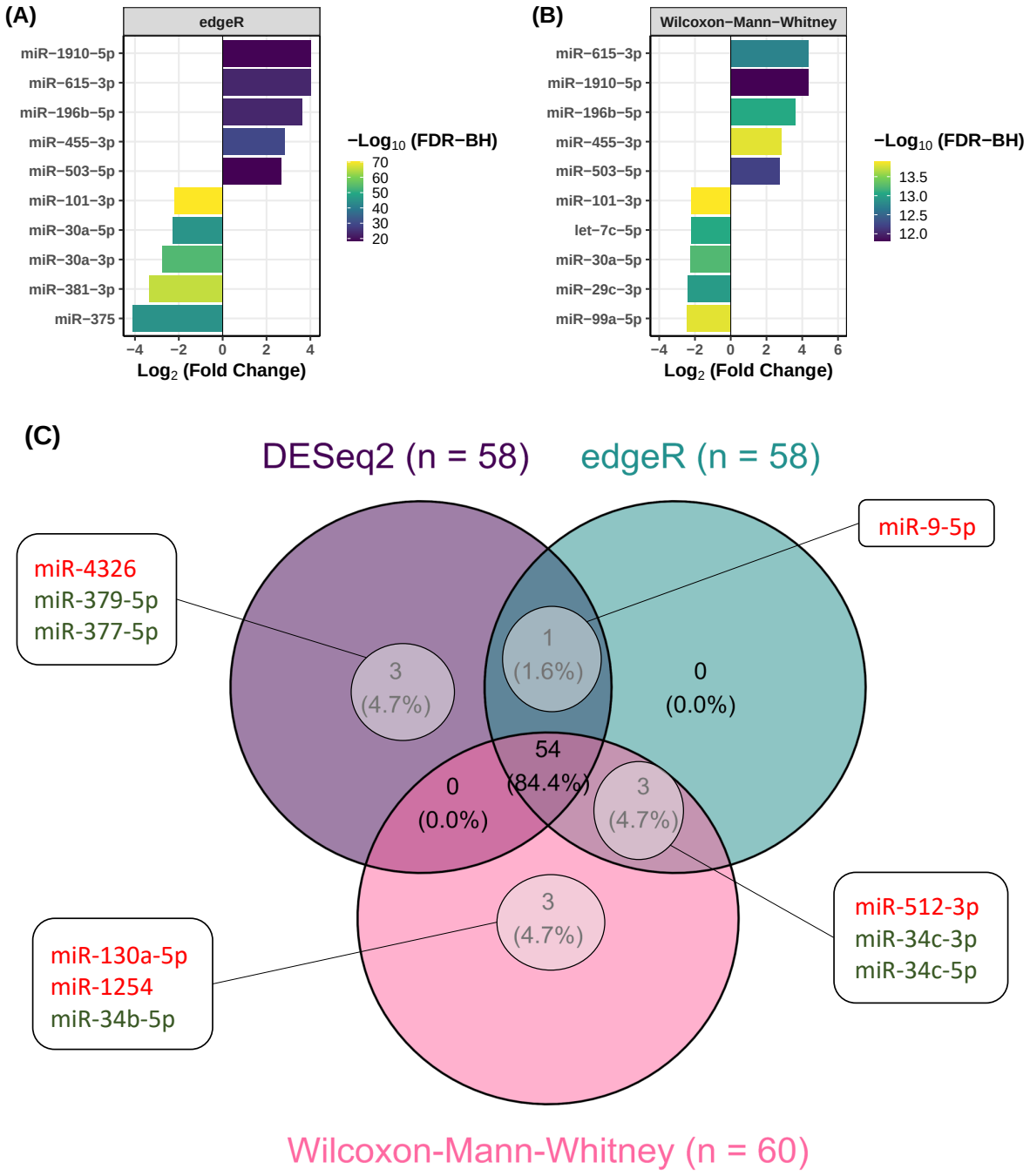


Figura 16 – Gráfico de barras dos 10 DEmiRNAs mais desregulados no câncer da cavidade oral pelo método de edgeR (A) e o teste de Wilcoxon-Mann-Whitney (B), respectivamente. (C) Diagrama de Venn mostrando a sobreposição dos resultados obtidos pelos três métodos (DESeq2, edgeR e o teste de Wilcoxon-Mann-Whitney).

As Figuras 17 (A) e (B) apresentam as distribuições dos dez principais miRNAs (superexpressos e subexpressos), obtidos pelo DESeq2. Esses miRNAs exibiram diferenças significativas na comparação entre o tumor primário e o tecido normal adjacente ao tumor, associados ao câncer de cavidade oral.

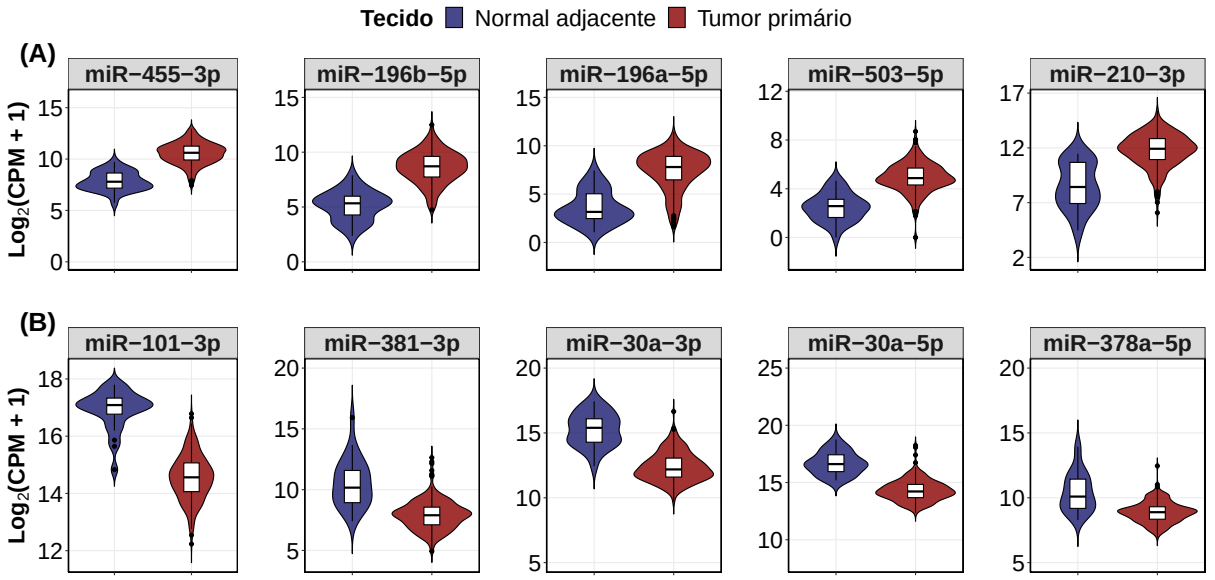


Figura 17 – Gráficos de violino ilustrando a expressão dos cinco principais DEmiRNAs superexpressos (A) e dos cinco DEmiRNAs subexpressos (B) ao comparar o tumor primário com o tecido normal adjacente ao tumor, relacionados ao câncer na cavidade oral.

Em seguida, aplicou-se a PCA com os 58 DEmiRNAs para visualizar seus níveis de expressão entre o tumor primário e o tecido normal adjacente, e outras variáveis clinicopatológicas (Figura 18).

Com base nos critérios de autovalores (Figura 18 A), optou-se por considerar doze componentes que apresentaram autovalores ($\lambda > 1$) responsáveis por 70% da variação nos dados. No entanto, os critérios de Horn's e Elbow sugeriram a escolha de 5 e 6 componentes, respectivamente, abrangendo cerca de 60% da variação dos dados (Figura 18 B).

A Figura 18 (C) mostra que miR-411-5p e miR-381-3p foram os principais contribuintes para a primeira componente, enquanto miR-503-5p, miR-133a-3p, miR-133b e miR-1-3p foram os principais contribuintes para o segundo componente. Esses miRNAs possivelmente desempenham um papel crucial na estratificação dos tecidos, sendo que a linha tracejada em vermelho representa a média esperada de contribuição de cada atributo para a criação das componentes. Na análise de dados genômicos, é frequente que os componentes expliquem apenas uma pequena parte da variação, devido à presença de um grande número de variáveis com efeito fraco a moderado, que apresentam comportamentos bastante heterogêneos.

A Figura 18 (D) exibe os dois primeiros componentes principais, explicando apenas 43,34% da variabilidade total dos dados, embora haja alguma sobreposição das elipses do nível de confiança de 95% em ambos os gráficos, nota-se uma clara tendência de segregação entre o tumor primário e o tecido normal adjacente.

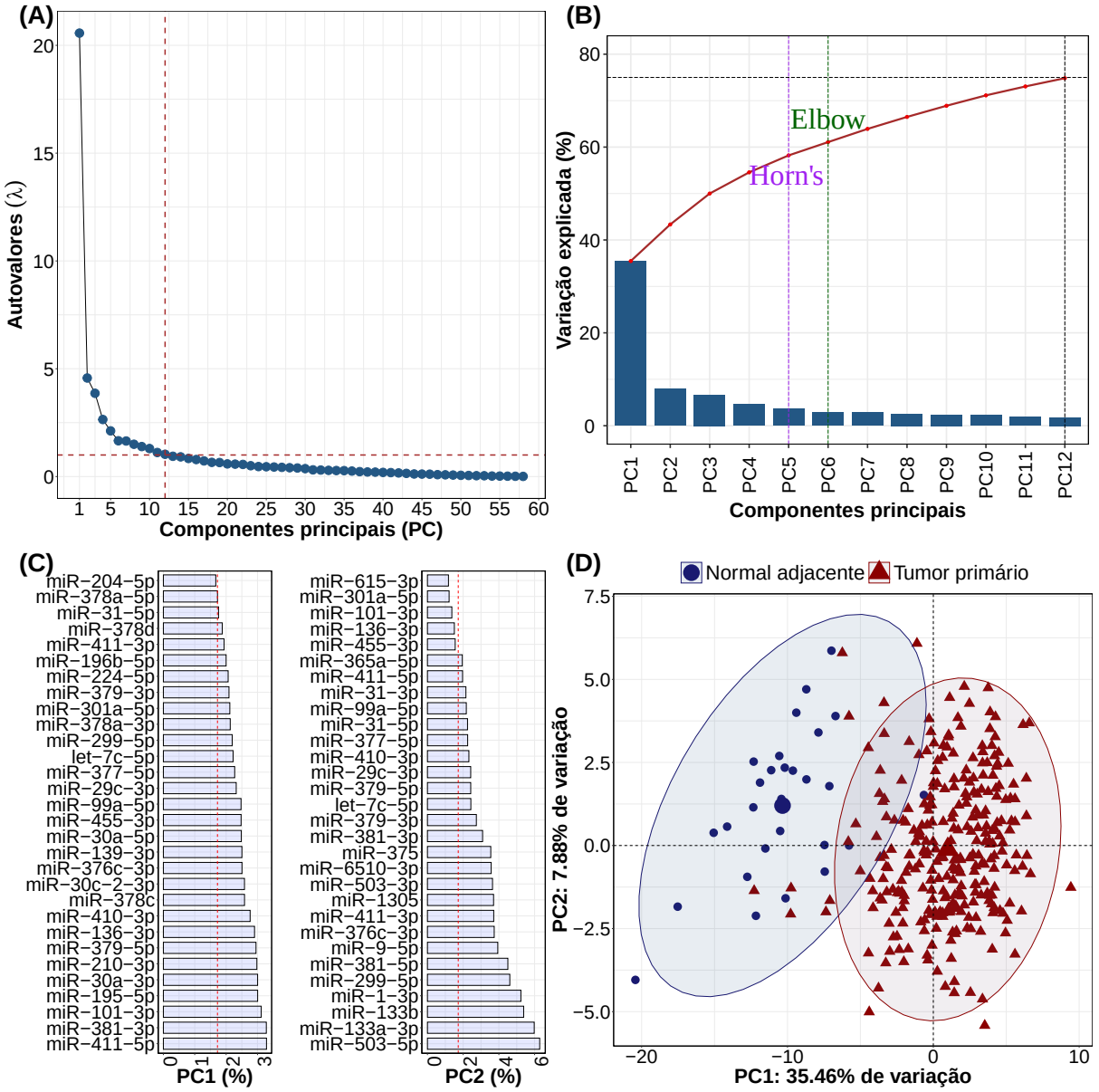


Figura 18 – Análise de componentes principais com base na expressão dos DEmiRNA de cada amostra associados ao câncer da cavidade oral. (A) *Scree plot* dos autovalores para cada componente principal. (B) *Scree plot* da variação explicada dos componentes principais. (C) Gráficos de barras para duas dimensões com base na contribuição dos miRNAs. (D) PCA mostra o padrão de expressão dos miRNAs na estratificação do tumor primário e tecido adjacente.

Por outro lado, na Figura 19, é apresentada o mapa de calor resultante do agrupamento hierárquico dos níveis de expressão de miRNAs ao comparar as características clinicopatológicas, em uma escala de cores, onde o vermelho escuro indica miRNAs superexpressos e o verde escuro indica subexpressos. Esses miRNAs apresentam expressão inversa, ou seja, um grupo de miRNAs é superexpresso nas amostras de tecido tumoral enquanto outros são subexpressos nesses mesmos tecidos. Por exemplo, os miR-378c, miR-378d, miR-378a-3p e miR-378a-5p apresentam expressão semelhante nas amostras de tecido normal adjacente, sugerindo que podem ter funções

semelhantes (co-regulados). Além disso, é possível observar que várias amostras apresentam uma tendência de agrupamento entre si, influenciadas por diversos fatores clinicopatológicos, tais como o estadiamento patológico e o sexo.

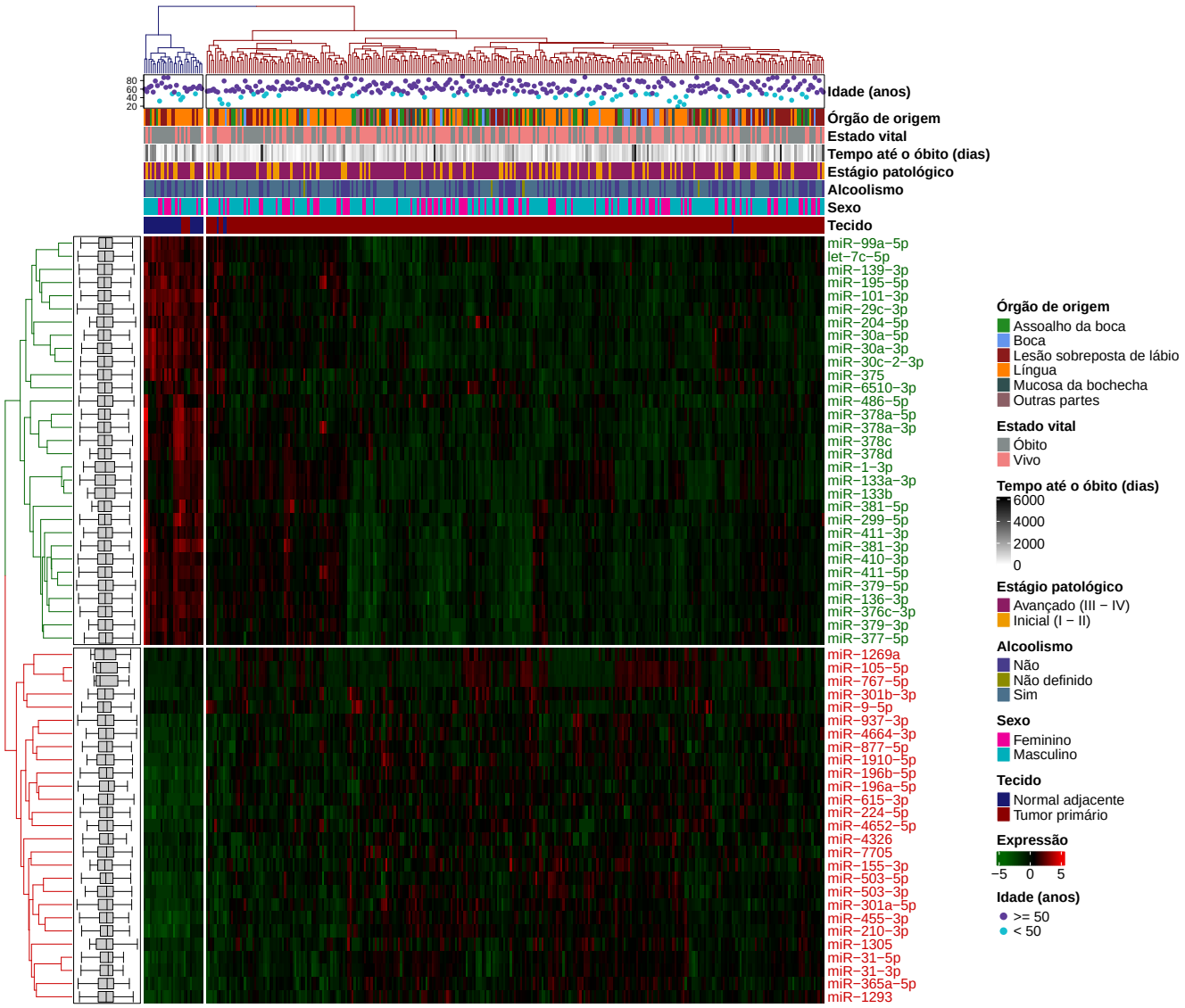


Figura 19 – Mapa de calor de agrupamento hierárquico da expressão de miRNA entre as amostras dos dois grupos (tecido tumoral e normal adjacente ao tumor) no câncer da cavidade oral usando a distância Euclideana e o método de ligação completa.

Por sua vez, o método de agrupamento hierárquico também foi utilizado para agrupar os DEmiRNAs superexpressos e subexpressos. A Figura 20 apresenta uma separação clara entre os dois grupos, destacando que os miRNAs superexpressos tendem a se correlacionar em um nível mais alto e de maneira mais uniforme do que os subexpressos. Observou-se que os miRNAs superexpressos miR-767-5p e miR-105-5p, bem como miR-31-5p e miR-31-3p, apresentam forte correlação entre si. Da mesma forma, os DEmiRNAs subexpressos, let-7c- 5p e miR-99-5p, bem como miR-133b e miR-133a-3p, mostraram indícios de correlação linear forte. Adicionalmente,

os miRNAs com uma maior distância de similaridade apresentam uma correlação linear mais fraca.

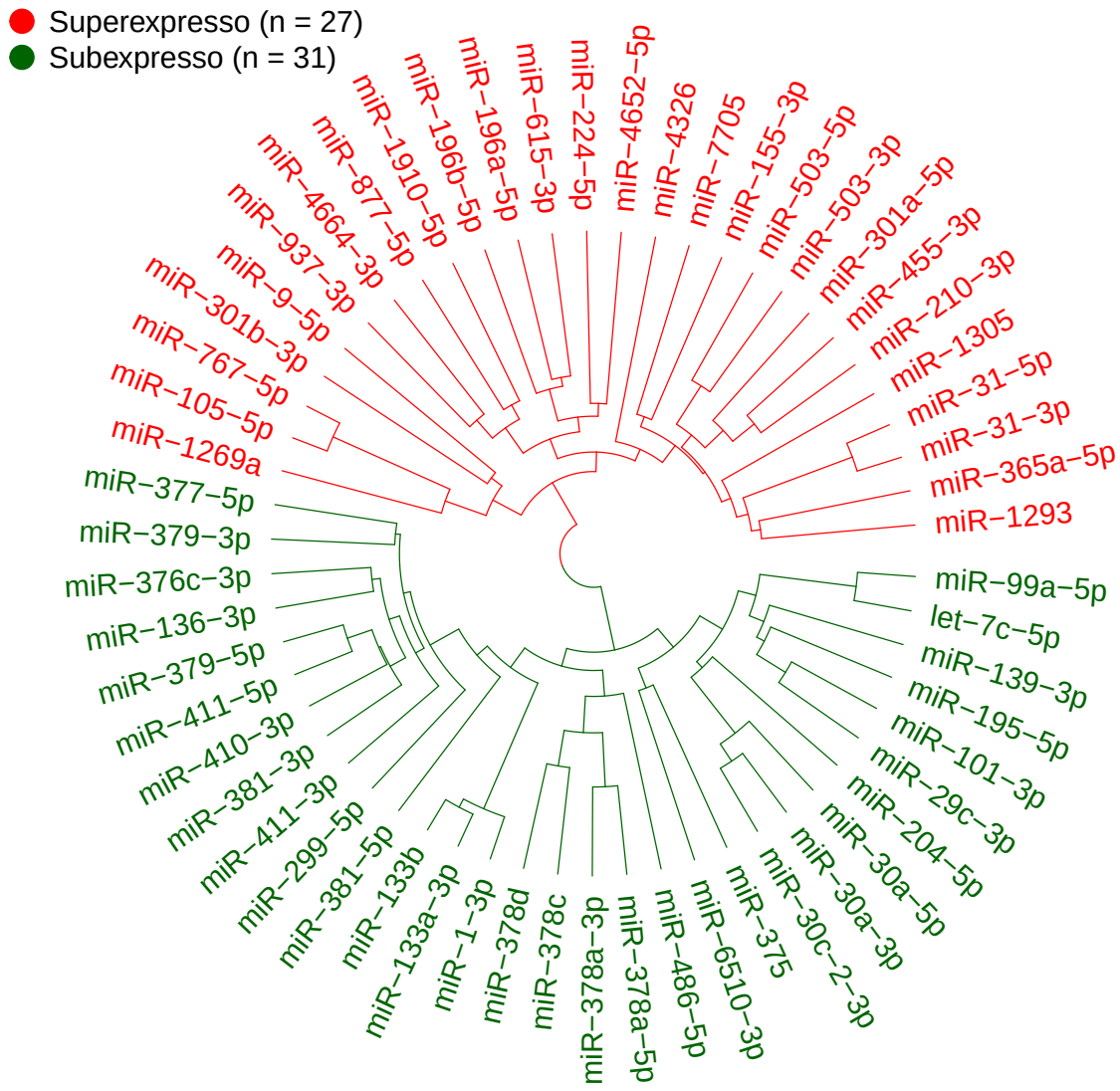


Figura 20 – Dendrograma circular hierárquico da separação entre grupos de DEmiRNAs superexpressos e subexpressos no câncer da cavidade oral usando a distância Euclideana e o método de ligação completa.

Na Figura 21, apresenta-se a correlação de Spearman utilizando o método de ligação completa. Observa-se uma forte correlação linear positiva entre diversos miRNAs, como miR-31-5p e miR-31-3p, bem como miR-767-5p e miR-105-5p, sugerindo possível envolvimento em uma via metabólica comum ou processo biológico similar. Por outro lado, a correlação linear negativa entre miR-210-3p e miR-195-5p sugere que esses miRNAs podem estar associados a processos biológicos opostos ou atuar em vias metabólicas distintas.

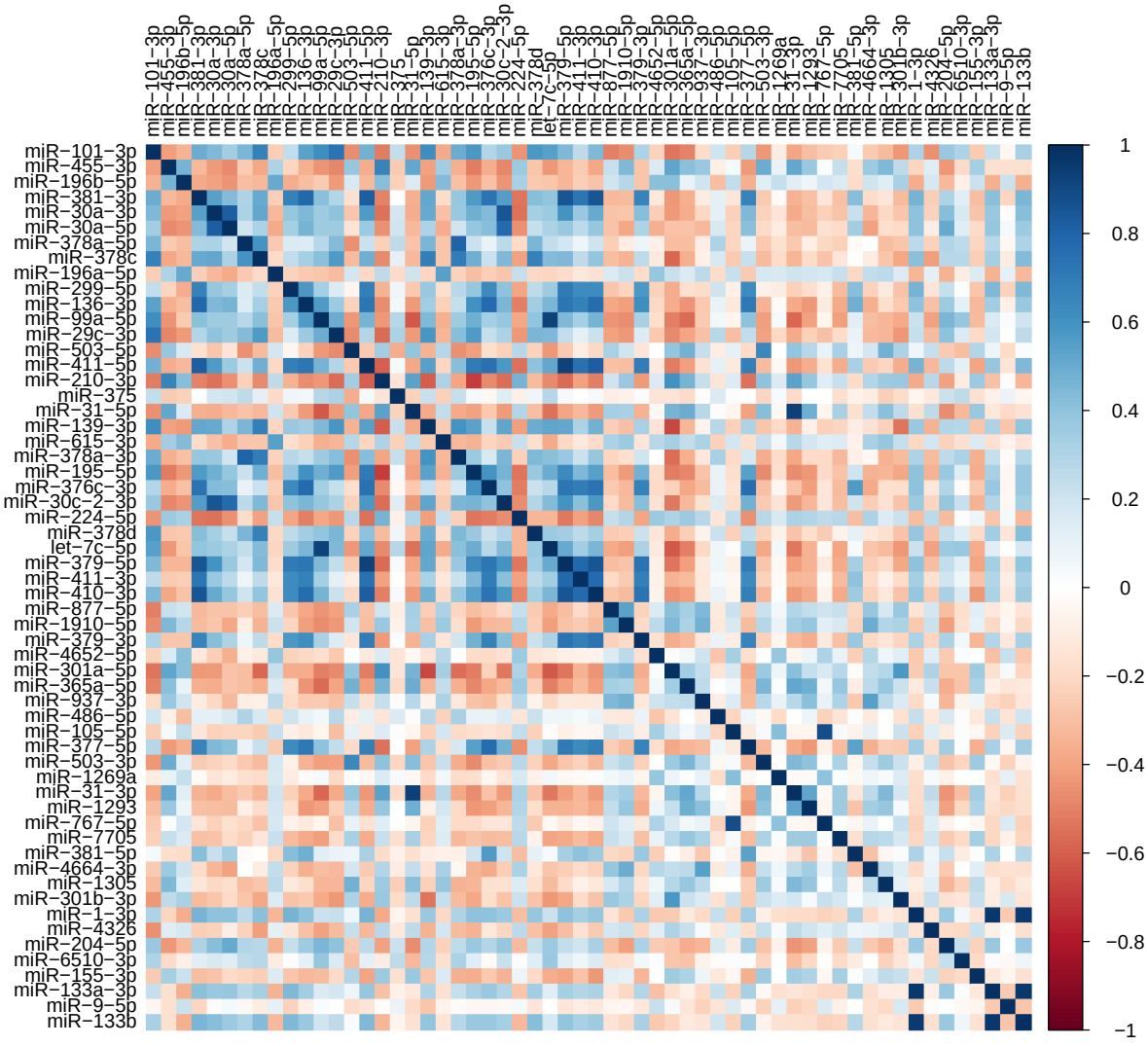


Figura 21 – Matriz de correlação linear de Spearman entre os DEmiRNAs associados ao câncer da cavidade oral.

6.2.2 Identificação de DEmiRNAs no câncer da laringe

A Figura 22 (A) mostra a desregulação de 273 miRNAs no câncer da laringe obtidos pelo DESeq2 baseado no modelo Binomial Negativa. Após aplicar o critério ($\hat{\beta}_1 = |\log_2(FC)| > 2$ e FDR-BH $< 0{,}05$), identificaram-se 44 DEmiRNAs (30 superexpressos ou oncogênicos e 14 subexpressos ou supressores de tumor) em 100 amostras de tecido tumoral primário em comparação com 12 tecidos normais adjacentes, conforme apresentado na Figura 22 (B) e Tabela 18 do Apêndice A.2, respectivamente.

Os miRNAs superexpressos mais significativos foram (miR-196b-5p, miR-196a-5p, miR-615-3p, miR-4652-5p e miR-455-3p), enquanto os subexpressos foram miR-338-5p, miR-139-3p,

miR-378c, miR-378d e miR-99a-5p (Figuras 22 B e C) e a distribuição desses miRNAs encontram-se nas Figuras 23 (A) e (B), respectivamente.

Além disso, os miRNAs associados a um maior risco de desenvolvimento de câncer foram o miR-615-3p (RR = 153,38), miR-105-5p (RR =87,23), miR-767-5p (RR = 77,32), miR-196a-5p (RR = 26,49), miR-514a-3p (RR = 20,98). Por outro lado, os miR-139-3p, miR-375, miR-133a-3p, miR-133b e miR-1-3p demonstraram menor risco de desenvolvimento de câncer, conforme ilustrado na Figura 23 (D).

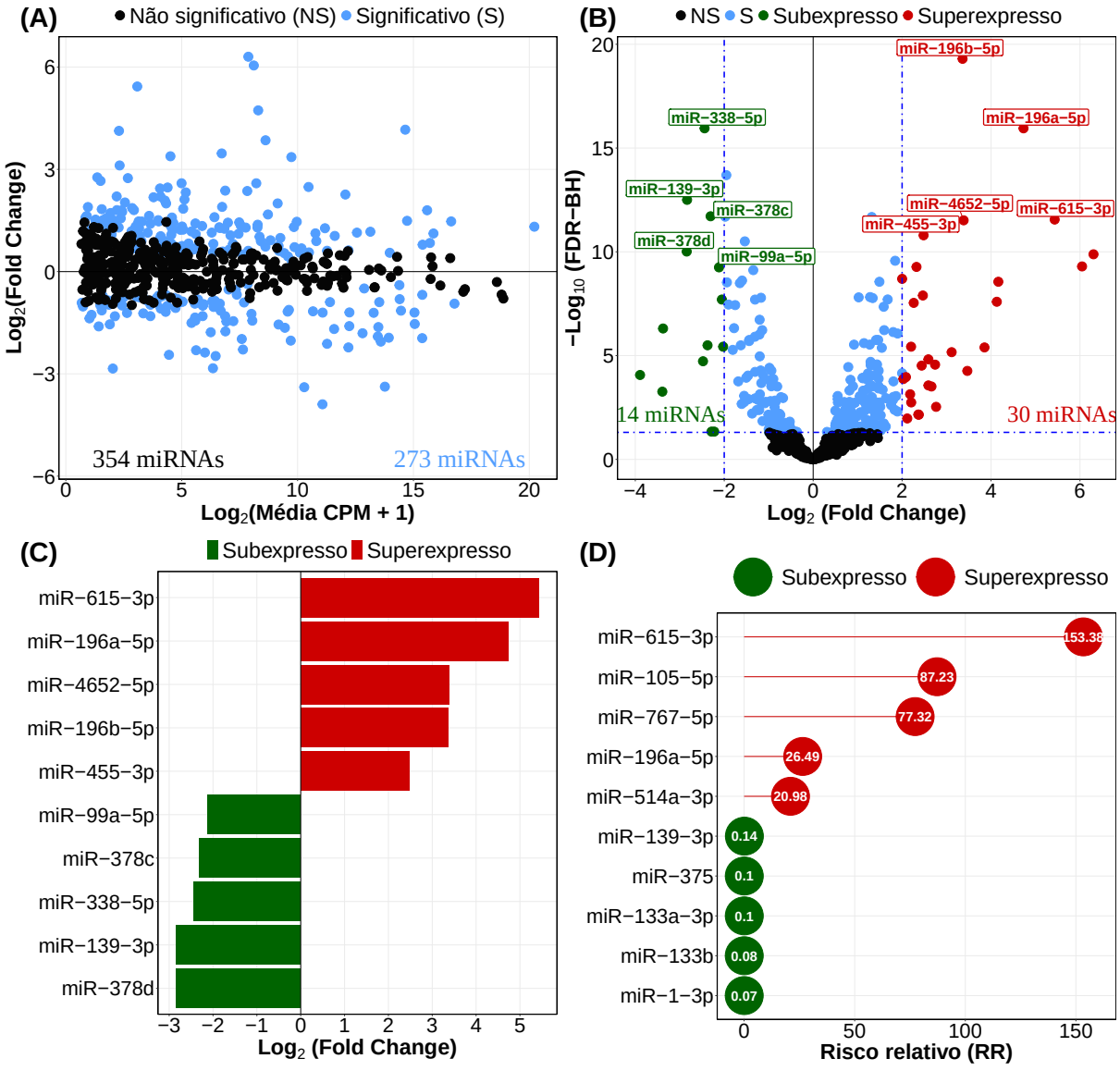


Figura 22 – DEmiRNAs na comparação do tumor primário e normal adjacente relacionados ao câncer da laringe. (A) Gráfico MA da razão logarítmica (M) versus a média (A) das diferenças entre os dois grupos para cada miRNA. (B) Gráfico de Vulcão dos DEmiRNAs. (C) Gráfico de Barras com os 10 DEmiRNAs mais desregulados. (D) Gráfico de pirulitos divergentes com os 10 DEmiRNAs em relação ao risco relativo.

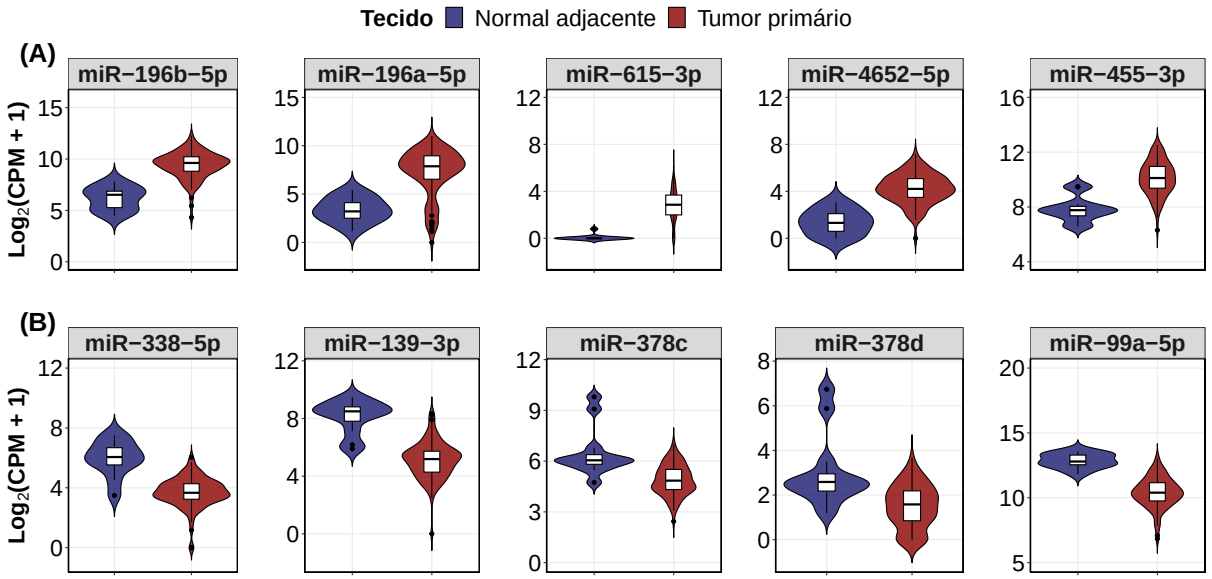


Figura 23 – Gráficos de violino da expressão dos 10 primeiros DEmiRNAs (superexpresso (A) e subexpresso (B)) na comparação do tumor primário com o tecido normal adjacente ao tumor associados ao câncer da laringe.

Os resultados referentes DEmiRNAs obtidos pelos métodos edgeR e teste de Wilcoxon-Mann-Whitney, respectivamente estão apresentados nas Tabelas 19 e 21 do Apêndice A.3 e A.4. Utilizando o edgeR, foram identificados 42 DEmiRNAs no total (30 superexpressos e 12 subexpressos, Tabela 19 do Apêndice A.3). Dentre os 10 principais DEmiRNAs (Figura 24 A), o miR-338-5p foi o mais significativo e apresentou regulação negativa no tumor primário em comparação com o tecido normal adjacente. Além disso, os miR-615-3p e miR-196a-5p demonstraram maior atividade oncogênica, enquanto o miR-375 exibiu maior atividade supressora do tumor.

No teste de Wilcoxon-Mann-Whitney, foram identificados 33 DEmiRNAs no total (23 superexpressos e 10 subexpressos, Tabela 21 do Apêndice A.4). Dentre os 10 principais DEmiRNAs (Figura 24 B), o miR-196b-5p foi o mais significativo (valor de FDR-BH menores a 0,001), sendo regulado positivamente, enquanto o miR-99a-5p foi o mais significativo e regulado negativamente no tumor primário em comparação com o tecido normal adjacente. Adicionalmente, o miR-615-3p exibiu maior atividade oncogênica, e o miR-375 apresentou maior atividade supressora do tumor (Figura 24 B).

A sobreposição desses métodos é ilustrada na Figura 24 (C), revelando que apenas 32 DEmiRNAs (68,1%) foram identificados de maneira comum entre os três métodos. É importante notar que a comparação entre DESeq2 e edgeR com o teste de Wilcoxon-Mann-Whitney mostra uma diferença substancial no conjunto de DEmiRNAs identificados. Essa discrepância pode influenciar interpretações biológicas específicas para este tipo de câncer, destacando a importância de considerar as metodologias utilizadas na análise de dados.

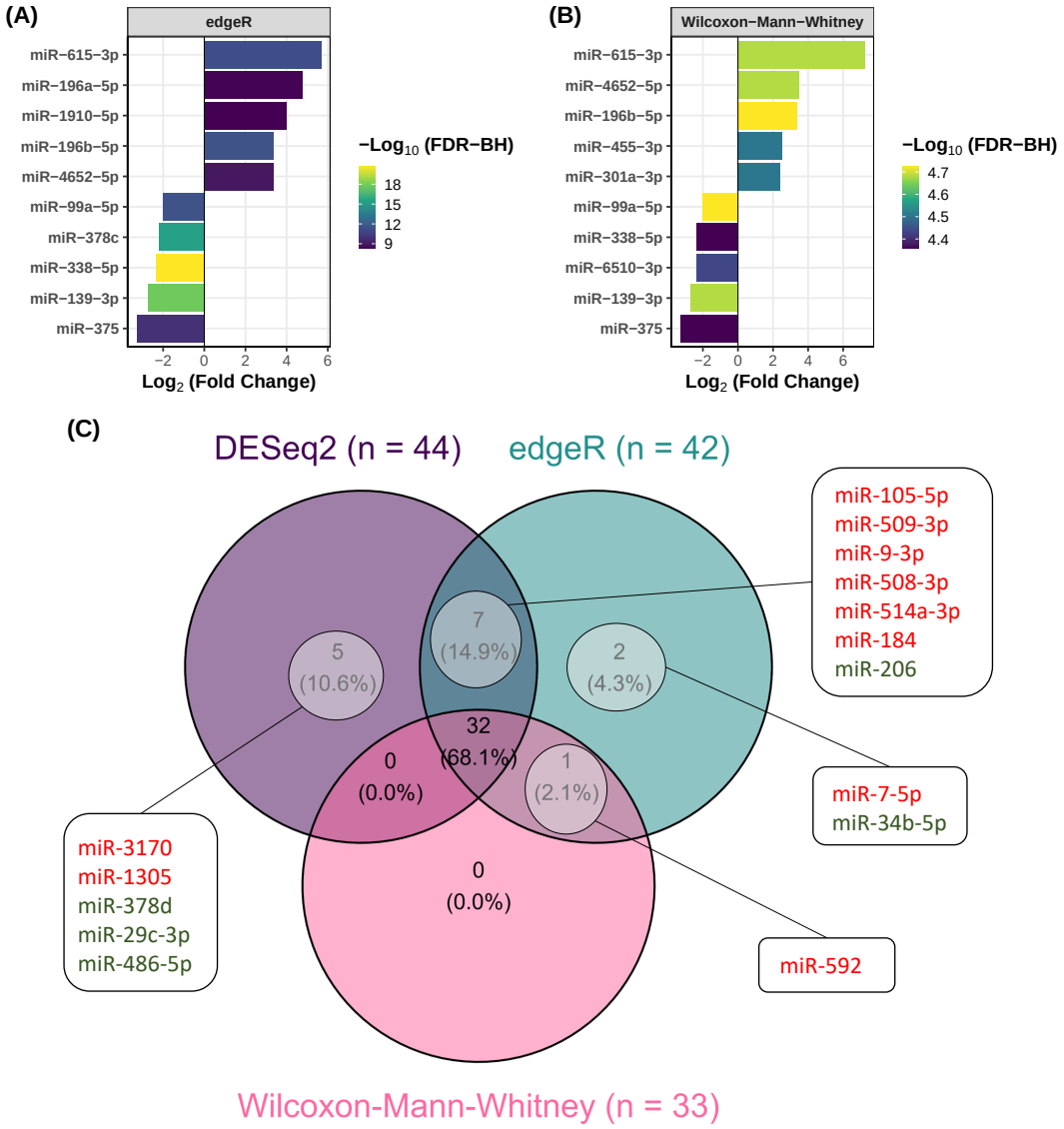


Figura 24 – Gráfico de Barras dos 10 DEmiRNAs mais desregulados no câncer da laringe pelo método de edgeR (A) e Wilcoxon-Mann-Whitney (B), respectivamente. (C) Diagrama de Venn da sobreposição dos resultados obtidos pelos três métodos.

A análise de componentes principais dos 44 DEmiRNAs e das amostras do câncer de laringe são mostradas na Figura 25.

Nas Figuras 25 (A) e (B) é mostrada os 12 componentes principias obtidos pelo critério dos autovalores ($\lambda > 1$) que representa aproximadamente 70% da variação total dos dados, o critério Horn's e Elbow sugere a escolha de apenas 6 e 8 componentes principais, respectivamente, representando aproximadamente 60% da variabilidade total dos dados.

Além disso, é possível visualizar que apenas o primeiro é responsável por 25% da variabilidade total, sendo o miR-139-3p, miR-301a-3p, miR-210-3p e miR-378c os maiores contribuintes na construção do primeiro componente principal e os miR-508-3p, miR-514a-3p, miR-509-3p, miR-9-5p e miR-9-3p os maiores contribuintes na construção do segundo

componente principal (Figura 25 C).

Na Figura 25 (D), observa-se que os dois componentes apenas explicam 35,56 % da variabilidade total dos dados, embora haja alguma sobreposição das elipses do nível de confiança de 95% em ambos os gráficos, nota-se uma clara tendência de segregação entre o tumor primário e o tecido normal adjacente.

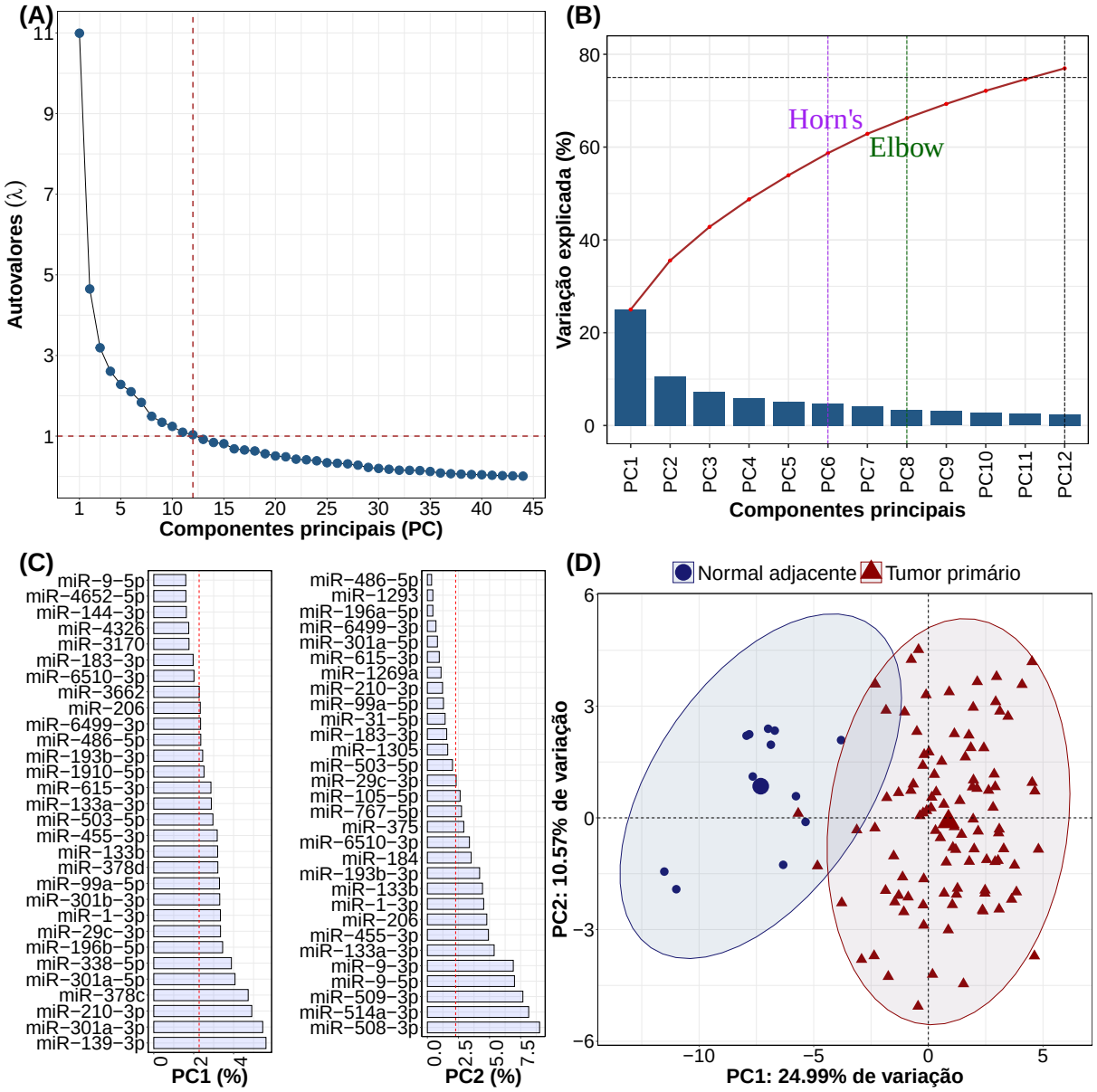


Figura 25 – Análise de componentes principais com base na expressão dos DEmiRNA de cada amostra associados ao câncer da laringe. (A) *Scree-plot* dos autovalores para cada componente principal. (B) *Scree-plot* da variação explicada dos componentes principais. (C) Gráficos de barras para duas dimensões com base na contribuição dos miRNAs. (D) PCA mostra o padrão de expressão dos miRNAs na estratificação do tumor primário e tecido normal adjacente.

Por outro lado, na Figura 26, é apresentado o mapa de calor resultante do agrupamento

hierárquico dos níveis de expressão de miRNAs ao comparar as características clinicopatológicas. A escala de cores utilizada apresenta o vermelho escuro indicando miRNAs superexpressos e o verde escuro indicando subexpressos. É interessante notar que um grupo de miRNAs apresenta superexpressão em tecidos tumorais, enquanto outros apresentam subexpressão nesses mesmos tecidos. Além disso, alguns miRNAs possuem expressão semelhante tanto em amostras normais adjacentes quanto em amostras tumorais primárias. Por exemplo, os miR-133b, miR-133a-3p e miR-1-3p apresentam expressão semelhante em ambos os tipos de amostras, enquanto miR-514a-3p, miR-508- 3p e miR-509-3p apresentam expressão semelhante apenas em amostras tumorais primários.

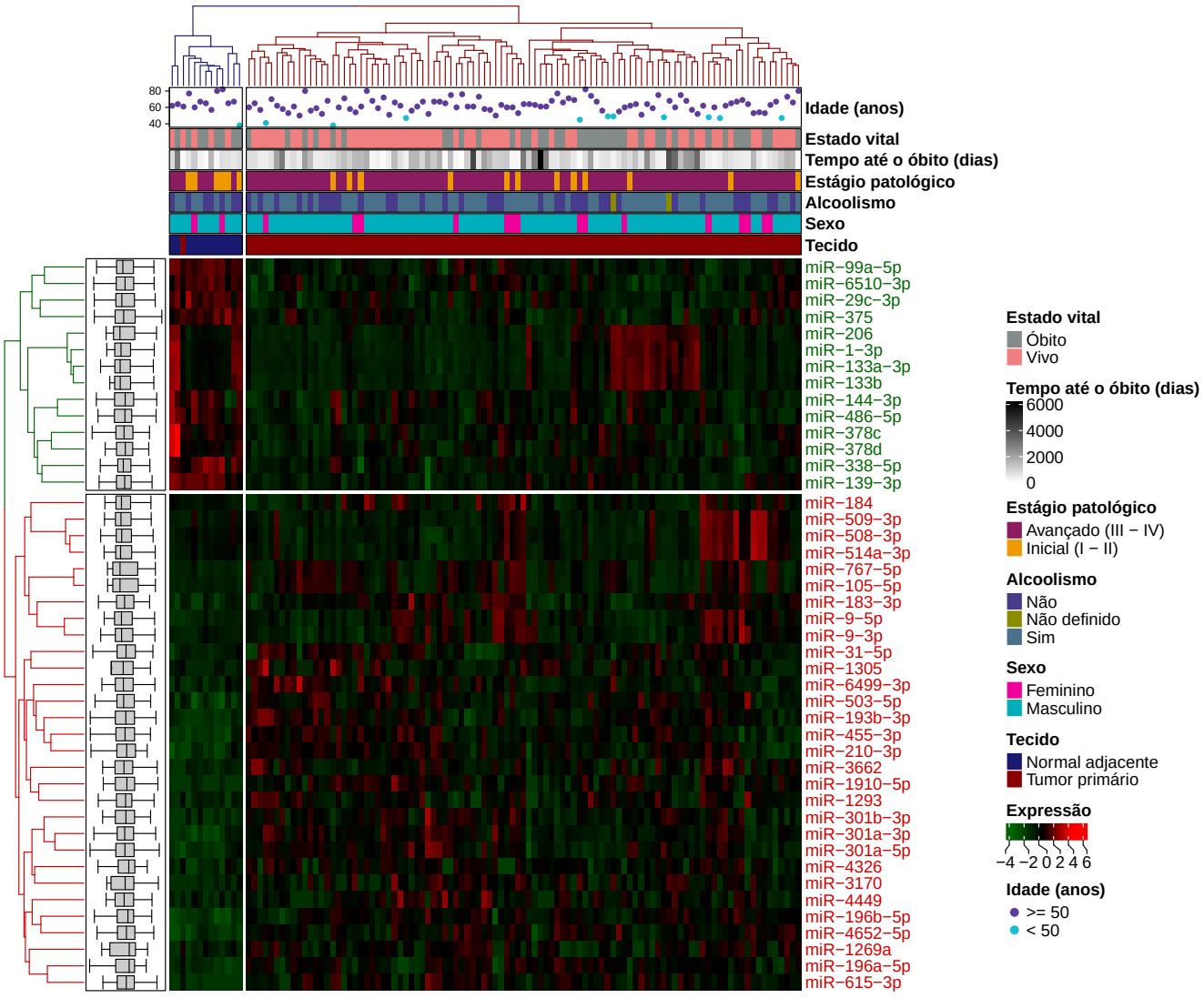


Figura 26 – Mapa de calor de agrupamento hierárquico da expressão de miRNA entre as amostras dos dois grupos (tumor primário e normal adjacente ao tumor) no câncer da laringe oral usando a distância Euclideana e o método de ligação completa.

O dendrograma, na Figura 27, ilustra os DEmiRNAs associados ao câncer de laringe. Observa-se uma forte correlação e formação de grupos entre os miRNAs superexpressos, como

miR-105-5p, miR-767-5p, miR-9-3p e miR-9-5p. Em contraste, os miRNAs subexpressos, incluindo miR-133b, miR-133a-3p, miR-1-3p e miR-206, também exibem uma alta correlação, formando um grupo separado. Adicionalmente, um agrupamento mais fraco é observado entre os miRNAs subexpressos, como miR-99a-5p, miR-6510-3p, miR-29c-3p, miR-375, miR-1305-5p e miR-31-5p.

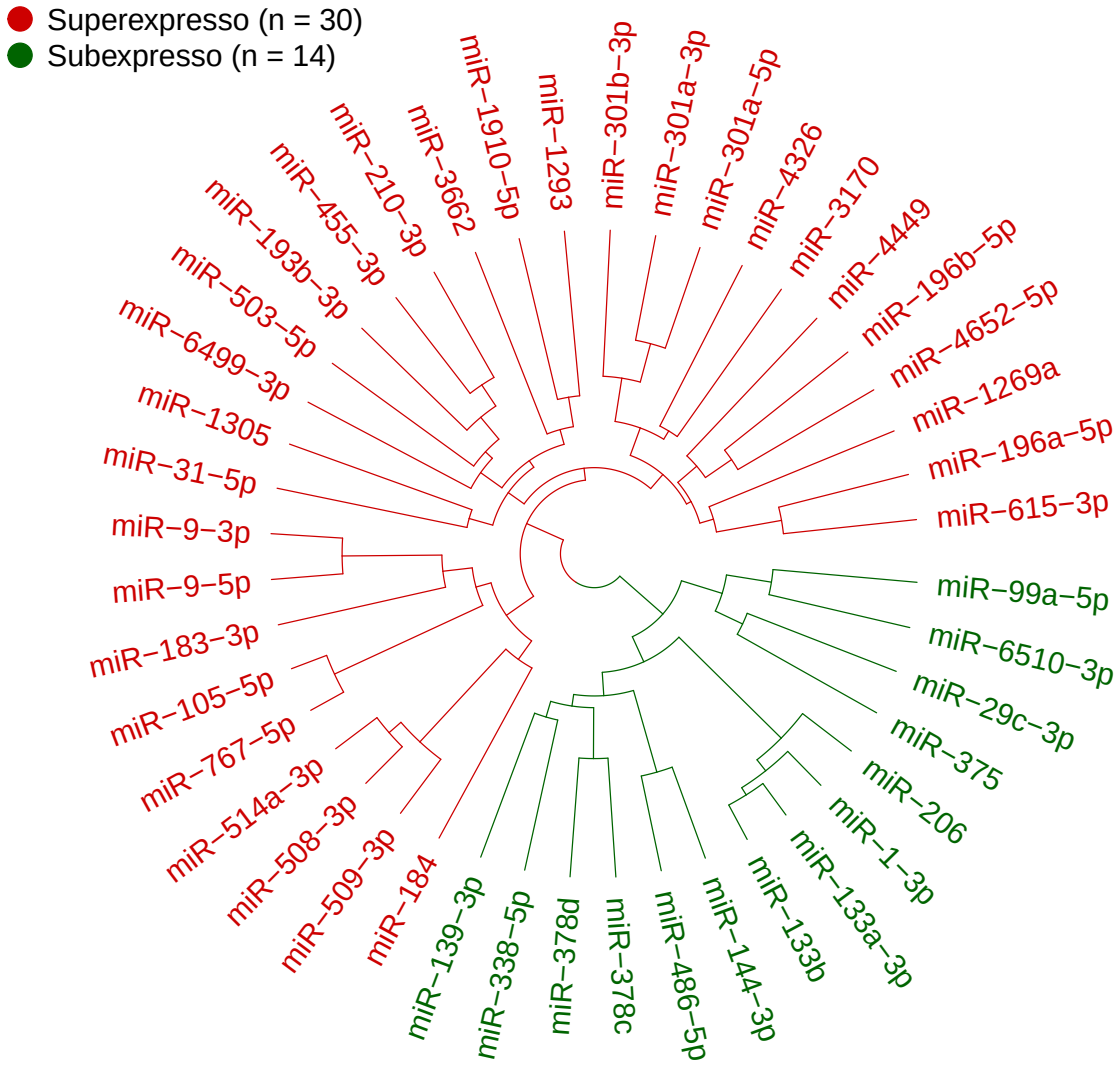


Figura 27 – Dendrograma circular hierárquico da separação entre grupos de DEmiRNAs superexpressos e subexpressos no câncer da laringe usando a distância Euclideana e o método de ligação completa.

A Figura 28 exibe a matriz de correlação linear de Spearman para cada um dos DEmiRNAs. Destaca-se uma correlação positiva significativa entre diversos pares, tais como, miR-133b e miR-1-3p, miR-133a-3p e miR-133b, miR-9-5p e miR-9-3p, indicando potenciais funções biológicas similares. Por outro lado, as correlações lineares negativas entre miR-301a-

5p e miR-139-3p, miR-301a-3p e miR-139-3p, indicam possível envolvimento em processos biológicos opostos.

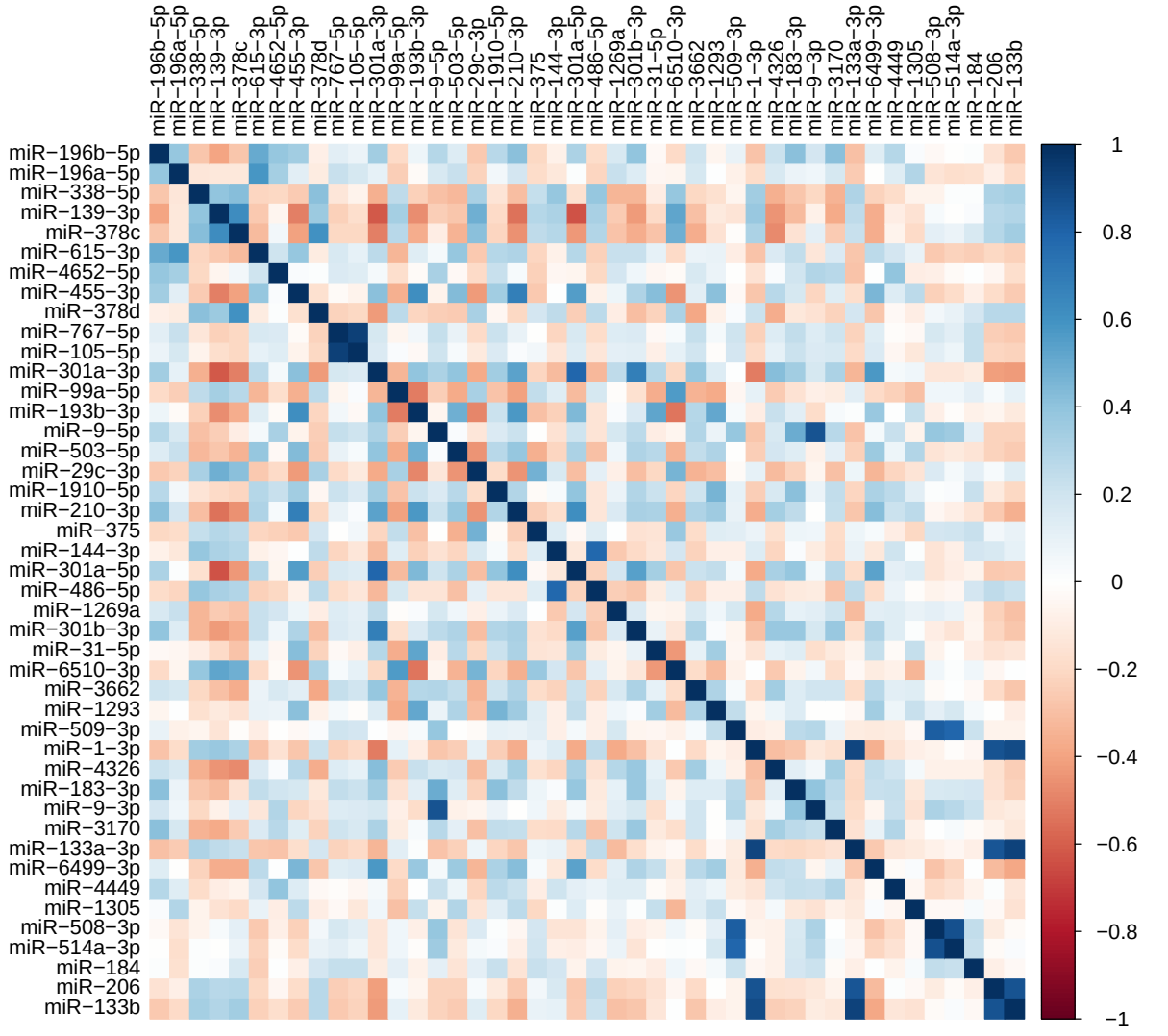


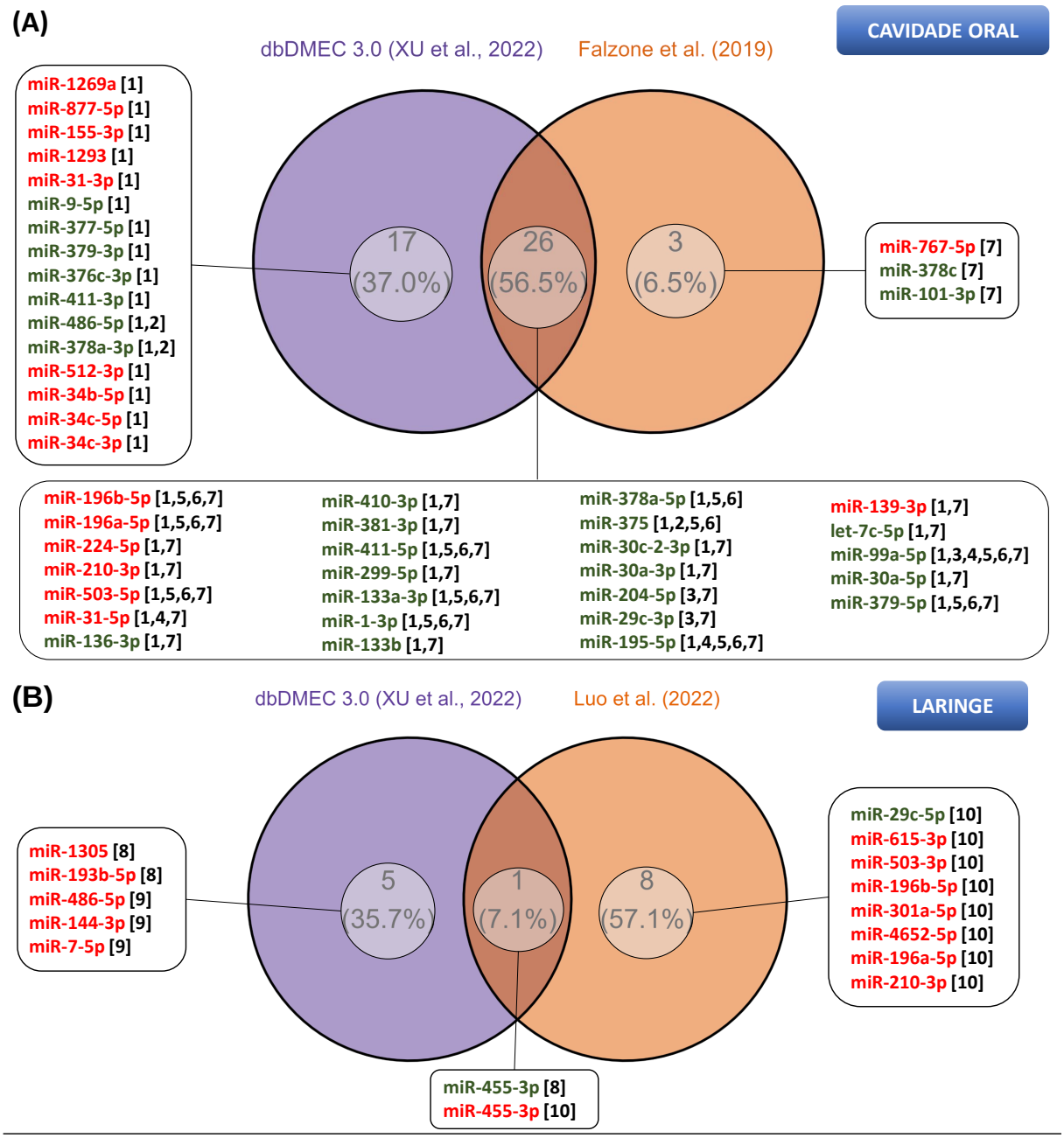
Figura 28 – Matriz de correlação linear de Spearman entre DEmiRNAs no câncer da laringe.

6.2.3 Validação de miRNAs candidatos com outras fontes de dados

Os DEmiRNAs previamente identificados mediante os três métodos para ambos os tipos de câncer foram submetidos à validação por fontes externas. Na Figura 29 (A), referente ao câncer de cavidade oral, identificaram-se 47 miRNAs em diferentes conjuntos de dados deste estudo. Dos 26 miRNAs comuns, apenas miR-196b-5p, miR-196a-5p, miR-503-5p, miR-411-5p, miR-133a-3p, miR-1-3p, miR-375, miR-195-5p, miR-99a-5p e miR-379-5p destacaram-se como os mais validados nos conjuntos de dados analisados. Em relação ao câncer da laringe,

apenas 13 miRNAs foram identificados como superexpressos, sendo que apenas o miR-455-3p foi identificado em comum, mas com expressões distintas (Figura 29 B).

Além disso, os 73 DemiRNAs identificados com o DESEq2 em ambos os tipos de cânceres foram associados a vários tipos de cânceres, como câncer do trato biliar, câncer cerebral, câncer de ovário, adenocarcinoma pulmonar, entre outros (Figura 29 C), enquanto apenas 71 DemiRNAs identificados em ambos os tipos de cânceres foram associados ao CCP (Figura 29 D).



[1] (EXP00281: GSE45238); **[2]** (EXP00177: GSE28100); **[3]** (EXP00491: GSE98463); **[4]** (EXP00560: GSE137865); **[5]** (GSE4538); **[6]** (GSE31277); **[7]** (TCGA); **[8]** (EXP00558: GSE124678); **[9]** (EXP00578: GSE116994); **[10]** (GSE133632).

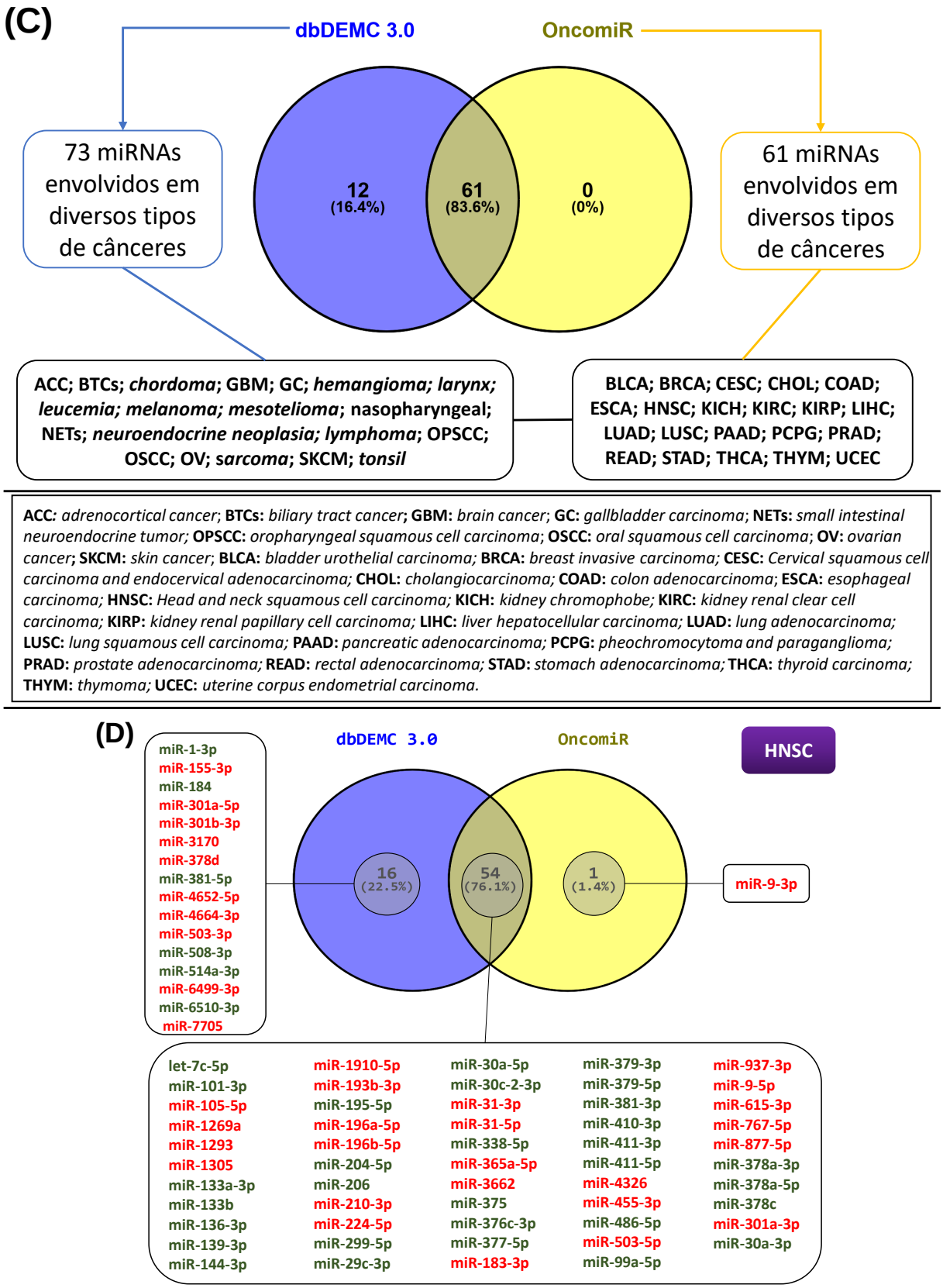


Figura 29 – Diagrama de Venn dos DEmiRNAs validados com outras fontes externas para câncer da cavidade oral (A), laringe (B), tipos de cânceres (C) e câncer de cabeça e pescoço (D), respectivamente.

6.2.4 Análise de enriquecimento e anotação ontológica de miRNAs

A Figura 30 ilustra a sobreposição entre as vias KEGG e as anotações GO nos casos de cânceres da cavidade oral e laringe. Observa-se que esses dois tipos de câncer compartilham um total de 118 vias KEGG e 461 termos GO, abrangendo diversos processos biológicos, componentes celulares e funções moleculares.

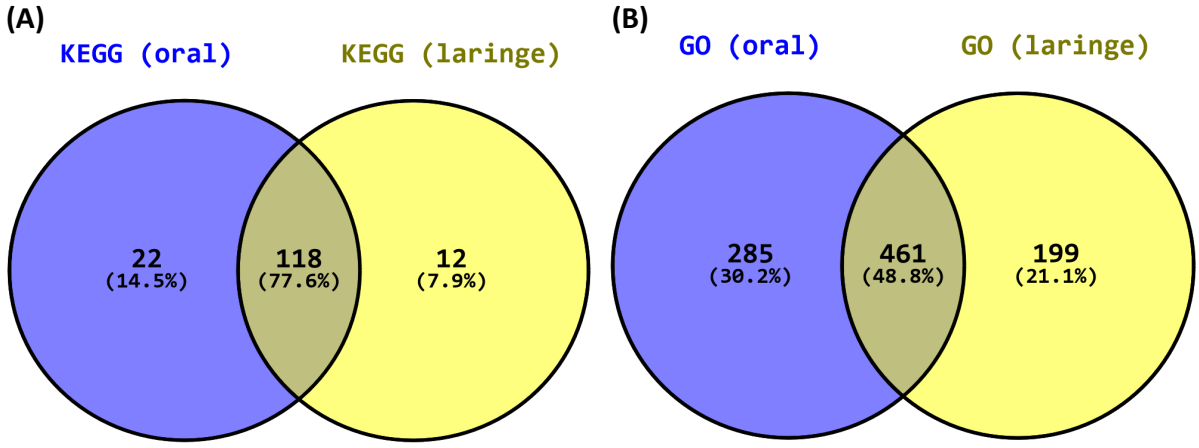


Figura 30 – Diagrama de Venn mostra a sobreposição entre (A) Vias KEGG e (B) Anotação GO em câncer da cavidade oral e laringe, respectivamente.

As 30 principais vias KEGG para esses cânceres são apresentadas nas Figuras 31 (A) e (C), respectivamente. Entre as vias mais destacadas, incluem-se a carcinogênese viral, a infecção por herpesvírus associada ao sarcoma de Kaposi, os caminhos no câncer, os proteoglicanos no câncer, e a shigelose, com menção especial ao ciclo celular e ao câncer de próstata.

De modo semelhante, as Figuras 31 (B) e (D) ilustram as 30 principais anotações Gene Ontology (GO) associadas a processos biológicos, componentes celulares e funções moleculares nos cânceres da cavidade oral e laringe.

No que se refere aos processos biológicos em ambos os tipos de câncer, destacam-se a regulação negativa da transcrição pela RNA polimerase II, a regulação positiva da transcrição pela RNA polimerase II, o processo viral, a regulação positiva da transcrição modelada por DNA, o ciclo celular, o processamento de mRNA e a organização da cromatina. Quanto aos componentes celulares, observa-se destaque para o núcleo, nucleoplasma, citosol, citoplasma, complexo contendo proteínas, adesão focal, mancha nuclear, cromatina nuclear e exossomo extracelular. Já em relação às funções moleculares, destacam-se a ligação proteica, ligação de RNA, ligação ao DNA, ligação de caderina, ligação de nucleotídeos, ligação à proteína ligase da ubiquitina, ligação à cromatina, ligação ao DNA específica da sequência da região reguladora cis da RNA polimerase II e ligação do fator de transcrição.

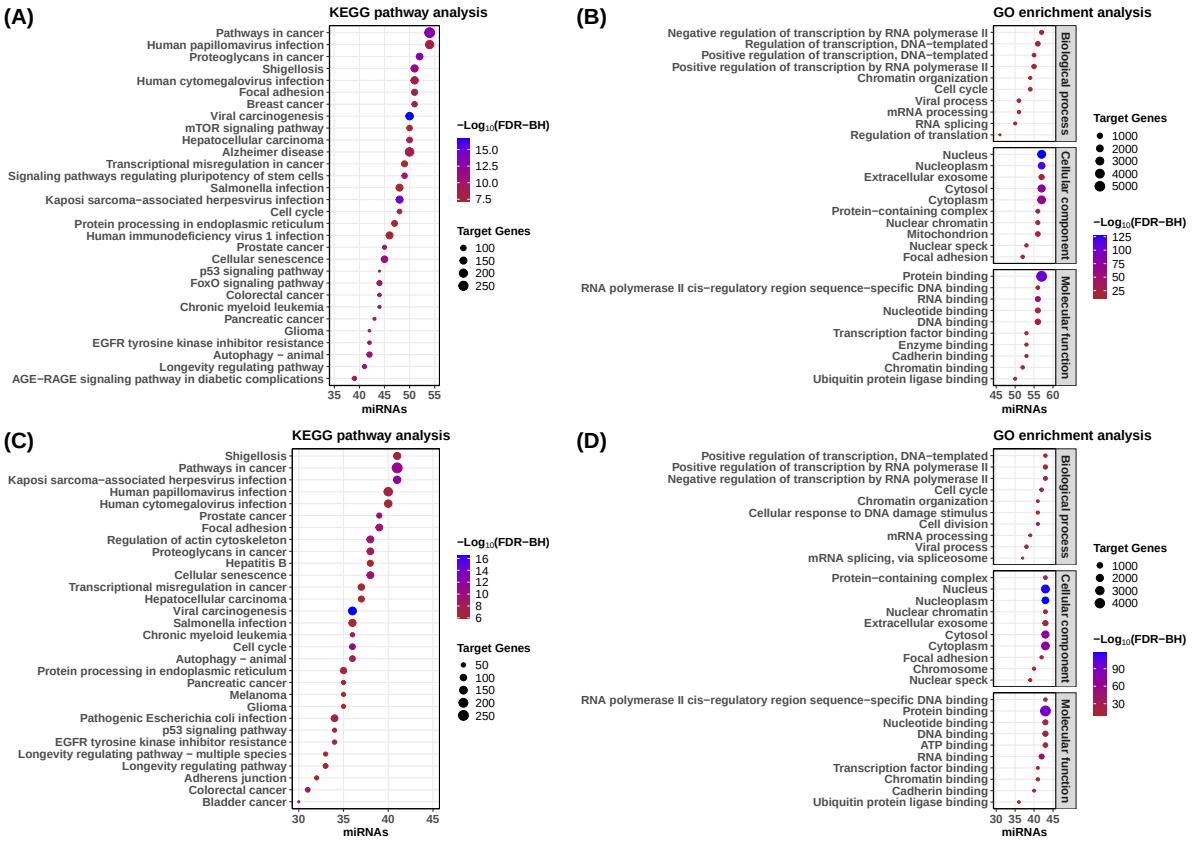


Figura 31 – Gráficos de bolhas mostrando as trinta principais vias KEGG e os termos GO associados aos miRNAs desregulados em pacientes com câncer oral (A, B) e laringe (C, D), respectivamente.

6.2.5 Identificação de DEmiRNAs comumente expressos no câncer da cavidade oral e laringe

As Figuras 32 (A) e (B) apresentam a sobreposição dos DEmiRNAs no câncer da cavidade oral e laringe. Nota-se a presença de 29 (39,7%) DEmiRNAs compartilhados entre esses dois locais anatômicos do CCP, sugerindo seu potencial como biomarcadores para ambos. Por outro lado, os DEmiRNAs que não se encontram na área de interseção são considerados possíveis biomarcadores específicos para seus respectivos sítios anatômicos.

A Figura 33 apresenta os 27 DEmiRNAs e as 17 principais vias KEGG associadas ao câncer da cavidade oral e da laringe. Dentro das 102 vias KEGG identificadas, os 27 DEmiRNAs exibem significância estatística em relação a 17 tipos de câncer, como carcinogênese viral, vias do câncer, câncer de próstata, desregulação transcricional e câncer de bexiga.

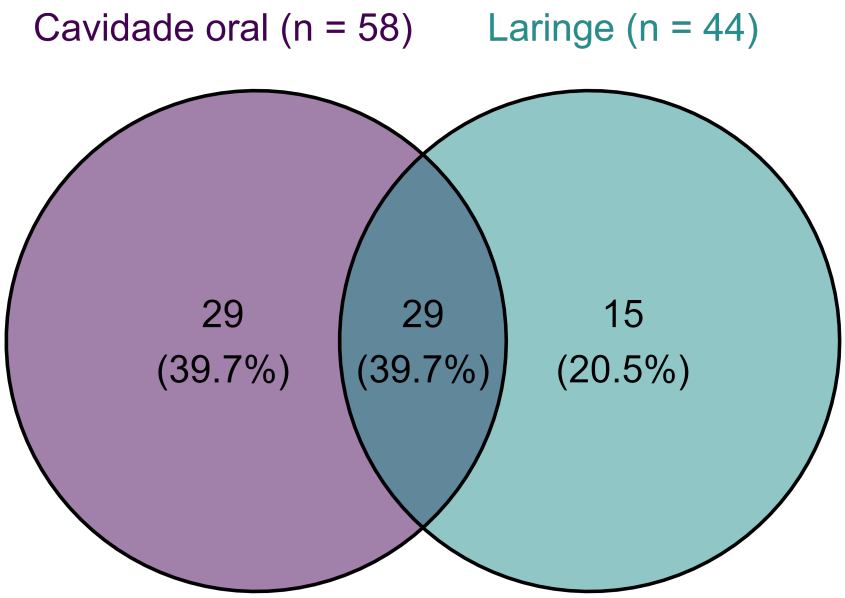


Figura 32 – Diagrama de Venn mostra a sobreposição dos DEmiRNAs entre o câncer da cavidade oral e da laringe.

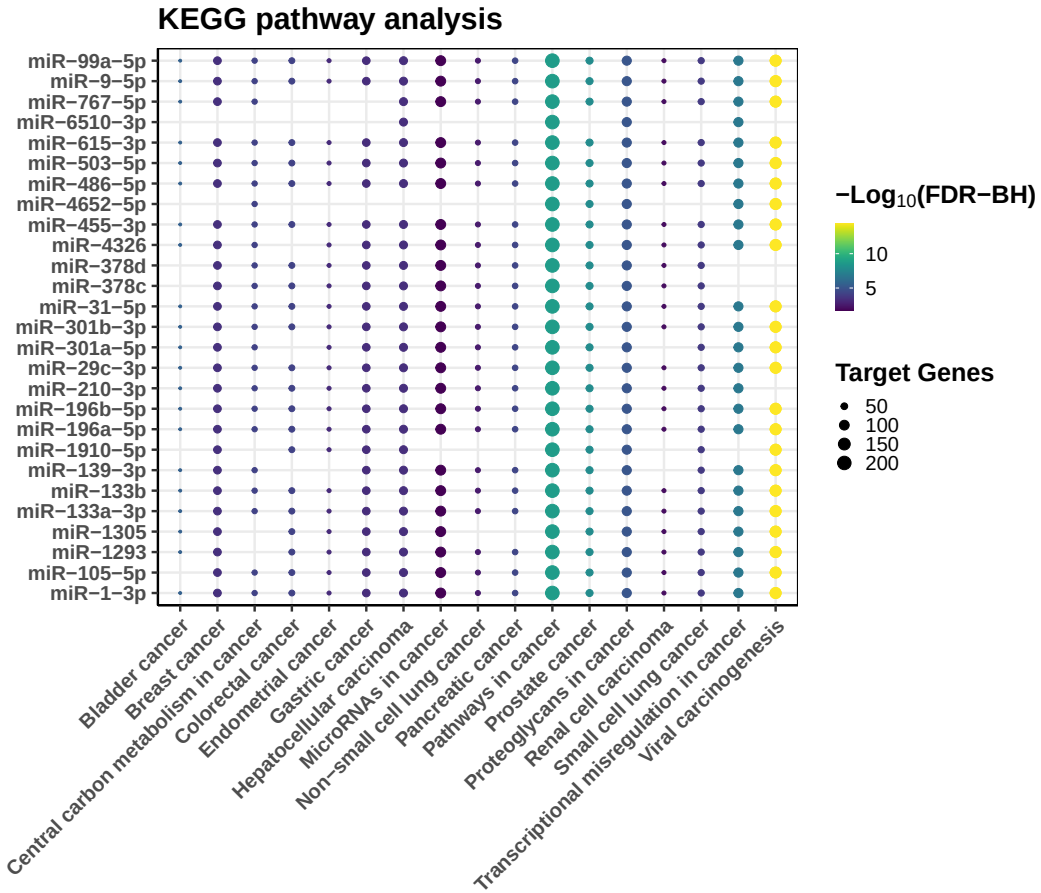


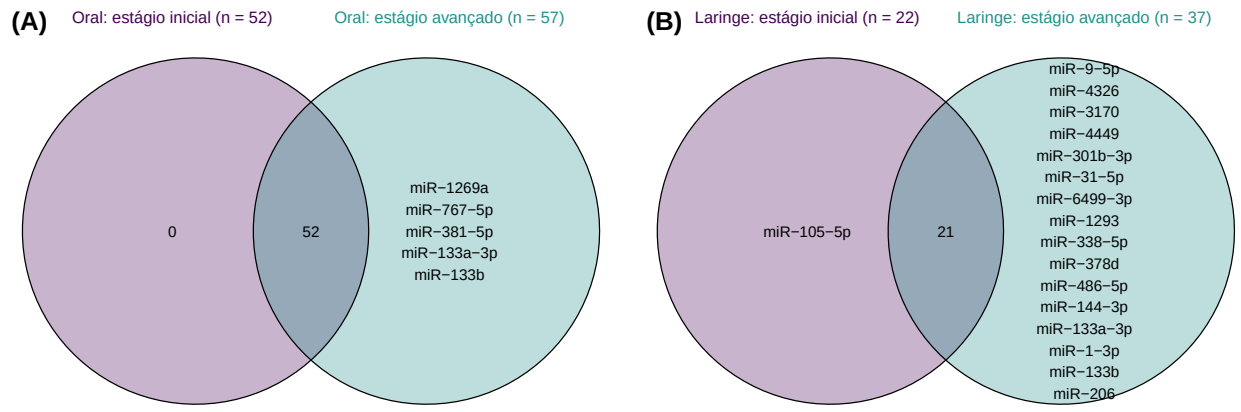
Figura 33 – Análise de Enriquecimento de vias KEGG em DEmiRNAs sobrepostos entre o câncer da cavidade oral e da laringe.

6.2.6 Identificação de DEmiRNAs associados ao estadiamento do câncer da cavidade oral e laringe

A comparação entre os DEmiRNAs em amostras de tecidos nos estadiamento iniciais e avançados de câncer na cavidade oral e laringe foi realizada por meio do teste de Wilcoxon-Mann-Whitney. Os resultados estão detalhados nas Tabelas 23 e 24 do Apêndice A.6, respectivamente.

Nos estadiamentos iniciais do câncer de cavidade oral, foram identificados 52 DEmiRNAs pelo teste de Wilcoxon-Mann-Whitney ao considerar as amostras de tumor primário e normais adjacentes. No estadiamento avançado do câncer da cavidade oral, foram identificados 57 DEmiRNAs e nos estadiamentos iniciais e avançados da laringe, foram identificados 22 e 38 DEmiRNAs, respectivamente (Figuras 34 A e B). Dos DEmiRNAs identificados para o câncer de cavidade oral, 52 estavam presentes em ambos os estadiamentos, enquanto apenas 21 eram compartilhados nos estadiamentos iniciais e avançados do câncer de laringe. Adicionalmente, apenas cinco DEmiRNAs (miR-1269a, miR-767-5p, miR-381-5p, miR-133a-3p e miR-133b) foram identificados como potenciais biomarcadores para o estadiamento avançado do câncer na cavidade oral (Figura 34 A). No caso do câncer de laringe, apenas o miR-105-5p pode ser considerado um biomarcador útil para o diagnóstico precoce. No entanto, no estadiamento avançado, foram identificados 16 DEmiRNAs (miR-9-5p, miR-4326, miR-3170, miR-4449, miR-301b-3p, miR-31-5p, miR-6499-3p, miR-1293, miR-338-5p, miR-378d, miR-486-5p, miR-144-3p, miR-133a-3p, miR-1-3p, miR-133b e miR-206) como potenciais biomarcadores (Figura 34 B).

As distribuições dos DEmiRNAs identificados nos estadiamentos iniciais e avançados do câncer de cavidade oral e laringe, comparando os tumores primários com os tecidos normais adjacentes (Figuras 34 C e D). No estadiamento avançado do câncer oral, observa-se superexpressão dos miR-1269 e miR-767-5p, enquanto os miR-381-5p, miR-133a-3p e miR-133b estão subexpressos (Figuras 34 C). Por outro lado, no estadiamento inicial do câncer da laringe, o miR-105-5p está superexpresso, enquanto, nos estadiamentos avançados, os miR-9-5p, miR-4326, miR-3170, miR-4449, miR-301b-3p, miR-31-5p, miR-6499-3p e miR-1293 apresentam superexpressão (Figuras 34 D).



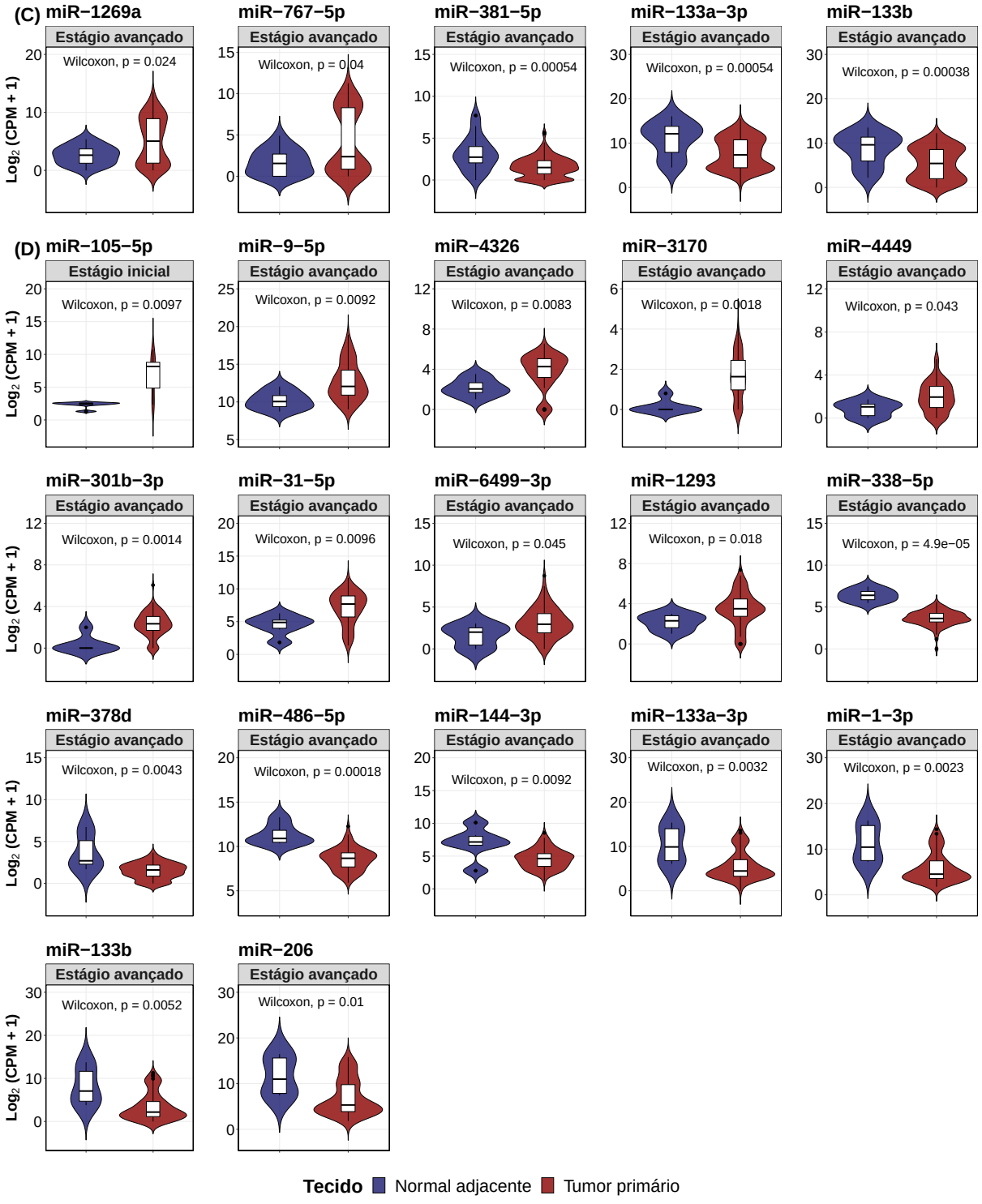


Figura 34 – (A, B) Diagrama de Venn mostrando a sobreposição dos DEmiRNAs entre estadiamento inicial e avançado do câncer da cavidade oral e laringe, respectivamente. (C) Gráficos de violino da expressão dos cinco DEmiRNAs nos estadiamentos avançados do câncer de cavidade oral. (D) Gráfico de violino da expressão dos DEmiRNAs no estadiamento inicial e avançado do câncer de laringe.

6.3 Identificação DEmRNAs, predição de alvos, redes de interação miRNA-mRNA, redes de proteína-proteína e análise funcional

A expressão diferencial dos mRNAs em câncer oral e laringe pelo DESeq2 estão mostrados na Figura 35. Identificou-se 9981 e 7717 mRNAs significativamente desregulados no câncer oral e laringe, respectivamente (Figuras 35 A e C). A partir do critério ($\hat{\beta}_1 = |\log_2(FC)| > 2$ e FDR-BH $< 0{,}05$), foram identificados 1423 DEmRNAs no câncer oral (469 superexpressos ou oncogênicos e 954 subexpressos ou supressores de tumor) e 1507 DEmRNAs no câncer de laringe (947 superexpressos e560 subexpressos) (Figuras 35 B e D) .

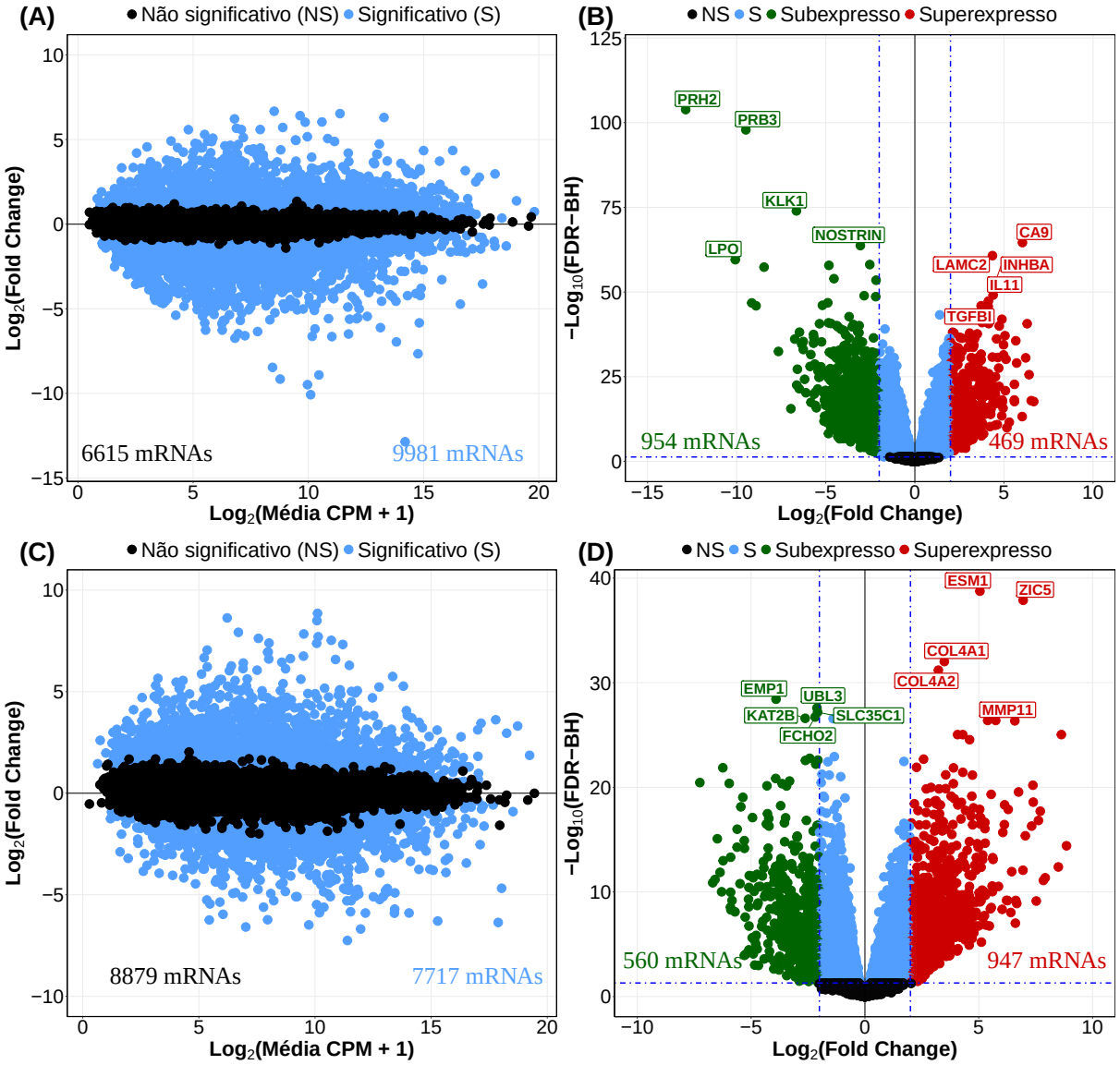


Figura 35 – *MA-Plot* (A) e *volcano-Plot* (B) dos DEmRNAs entre o tumor primário e tecido de normal adjacente relacionados ao câncer da cavidade oral (A, B) e laringe (C, D), respectivamente.

Após a identificação dos DEmRNAs, o próximo passo envolveu a predição dos genes alvo candidatos dos DEmiRNAs associados ao câncer oral e de laringe, respectivamente. Pelo mirDIP e o miRTarBase foram preditos 5042 e 6636 genes alvo dos DEmiRNAs para o câncer da cavidade oral, enquanto, para a laringe, foram preditos 4284 e 5557 genes alvo, respectivamente. Esses genes alvo foram sobrepostos aos DEmRNAs, resultando na identificação de 127 e 85 genes alvo candidatos, tanto para o câncer da cavidade oral quanto para o de laringe, conforme ilustrado nas Figuras 36 (A) e (B).

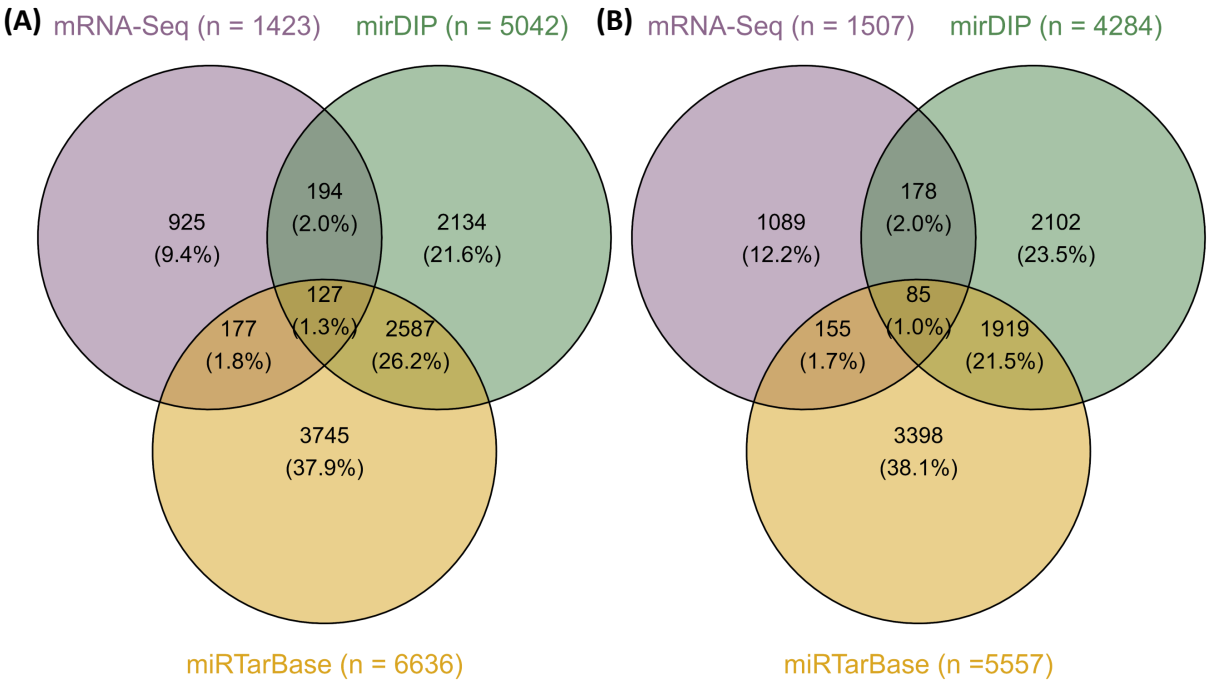


Figura 36 – Diagrama de Venn mostrando a sobreposição entre DEmRNAs e genes alvo dos DEmiRNAs (mirDIP e miRTarBase) associados ao câncer da cavidade oral (A) e da laringe (B), respectivamente.

No entanto, foram escolhidos os genes alvo candidatos cuja expressão é oposta à dos DEmiRNAs (se um miRNA estiver superexpresso, seu gene alvo deve estar subexpresso, e vice-versa), resultando na identificação de 103 (50 superexpressos e 53 subexpressos) e 63 (37 superexpressos e 26 subexpressos) genes alvo associados ao câncer da cavidade oral e laringe, respectivamente. Nas Figuras 37 (A) e (C), mostram-se os 10 principais DEmRNAs mais significativos, para câncer da cavidade oral e laringe, respectivamente.

Além disso, os resultados do PCA para as amostras de câncer da cavidade oral e laringe estão representados nas Figuras 37 (B) e (D). No contexto do câncer da cavidade oral, foram identificados oito componentes principais, os quais representam 72,48% da variabilidade total dos dados, conforme o critério dos autovalores ($\lambda > 1$). Adicionalmente, os critérios de Horn's e Elbow resultaram em quatro e seis componentes, abarcando 62,27% e 68,08% da variabilidade total dos dados, respectivamente. Para o câncer de laringe, a análise de PCA, seguindo o critério dos autovalores ($\lambda > 1$), revelou a presença de oito componentes principais, responsáveis por

74,45% da variabilidade total dos dados. Em contraste, os critérios de Horn's e Elbow resultaram em quatro e seis componentes, abrangendo 61,78% e 68,99% da variabilidade total dos dados, respectivamente.

No entanto, somente os dois primeiros elementos conseguiram explicar 52,99% e 50,56% da variabilidade total dos dados para o câncer oral e de laringe, respectivamente. Embora haja sobreposição nas elipses de nível de confiança de 95% entre as amostras, nota-se uma clara tendência de segregação entre o tumor primário e o tecido normal adjacente no câncer da cavidade oral, sendo essa separação ainda mais evidente no caso do câncer de laringe.

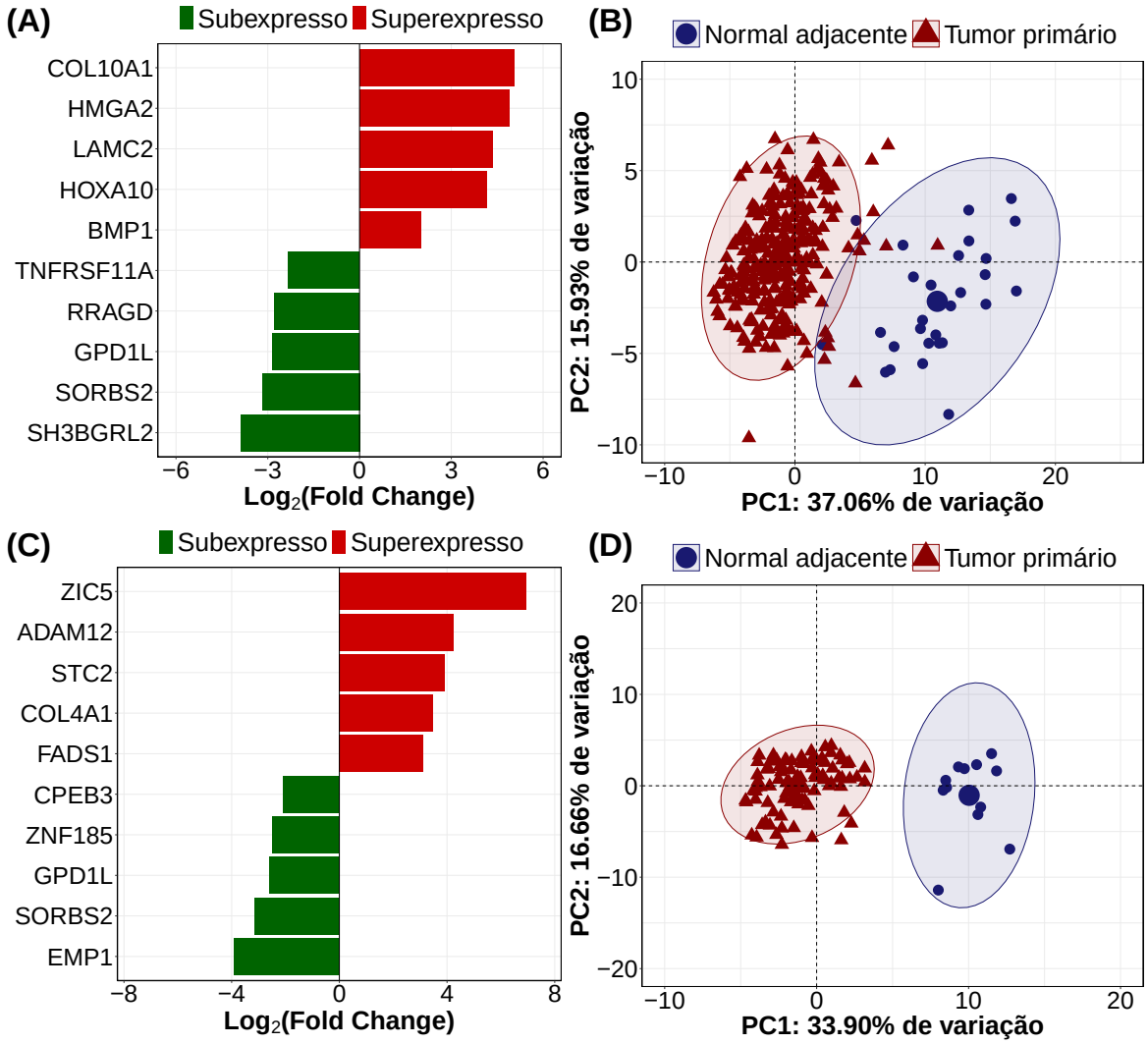


Figura 37 – (A, C) Gráfico de barras dos 10 DEmRNAs mais desregulados (valores FDR muito baixos) e influenciado pelo los valores $\log_2(FC)$ para o câncer da cavidade oral e laringe, respectivamente. (B, C) Análise de componentes principais dos DEmRNAs relacionados ao câncer da cavidade oral e laringe.

Além disso, as Figuras 38 (A) e (C) ilustram os mapas de calor derivados do agrupamento hierárquico, evidenciando que determinados mRNAs apresentam expressão semelhante em amostras de tumor primário, tecido normal adjacente e com as características clinicopatológicas.

Adicionalmente, nas Figuras 38 (B) e (D), são exibidos os dendrogramas circulares dos DEmRNAs superexpressos e subexpressos, relacionados ao câncer de cavidade oral e laringe, respectivamente.

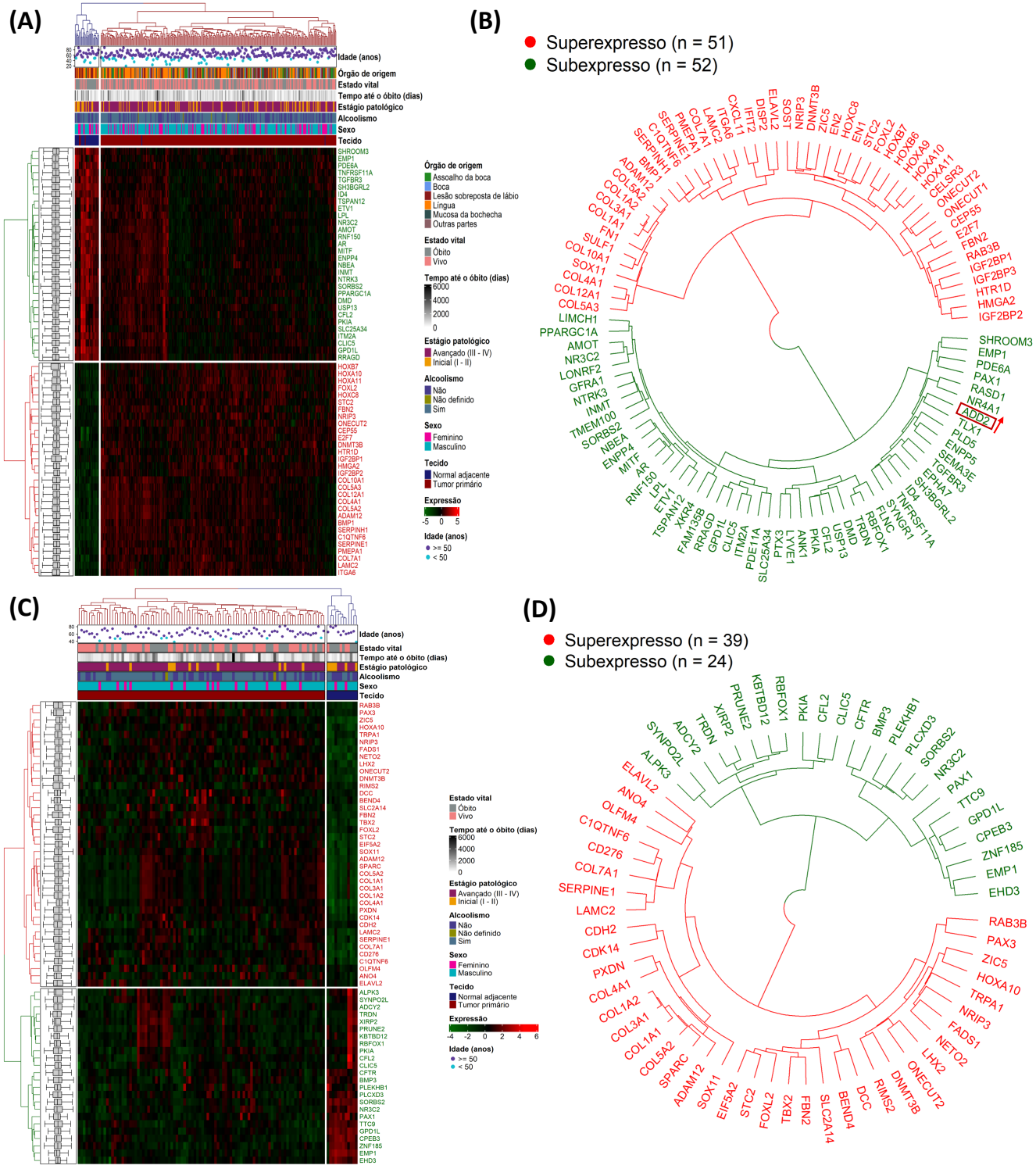


Figura 38 – *Heatmap* e dendrograma circular dos DEmRNAs entre as amostras dos dois grupos (tecido do tumor primário e normal adjacente) no câncer da cavidade oral (A,B) e laringe (C, D) usando a distância Euclideana e o método de ligação completa para as colunas (amostras) e Ward.D2 para as linhas (DEmRNAs). A seta vermelha em (B) significa que o gene ADD2 é superexpresso e que foi agrupado incorretamente.

6.3.1 Rede de interação regulatória miRNA-mRNA

As Figuras 39 (A) e (B) ilustram as redes regulatórias DEmiRNA-DEmRNA associadas ao câncer nas regiões da cavidade oral e laringe, respectivamente.

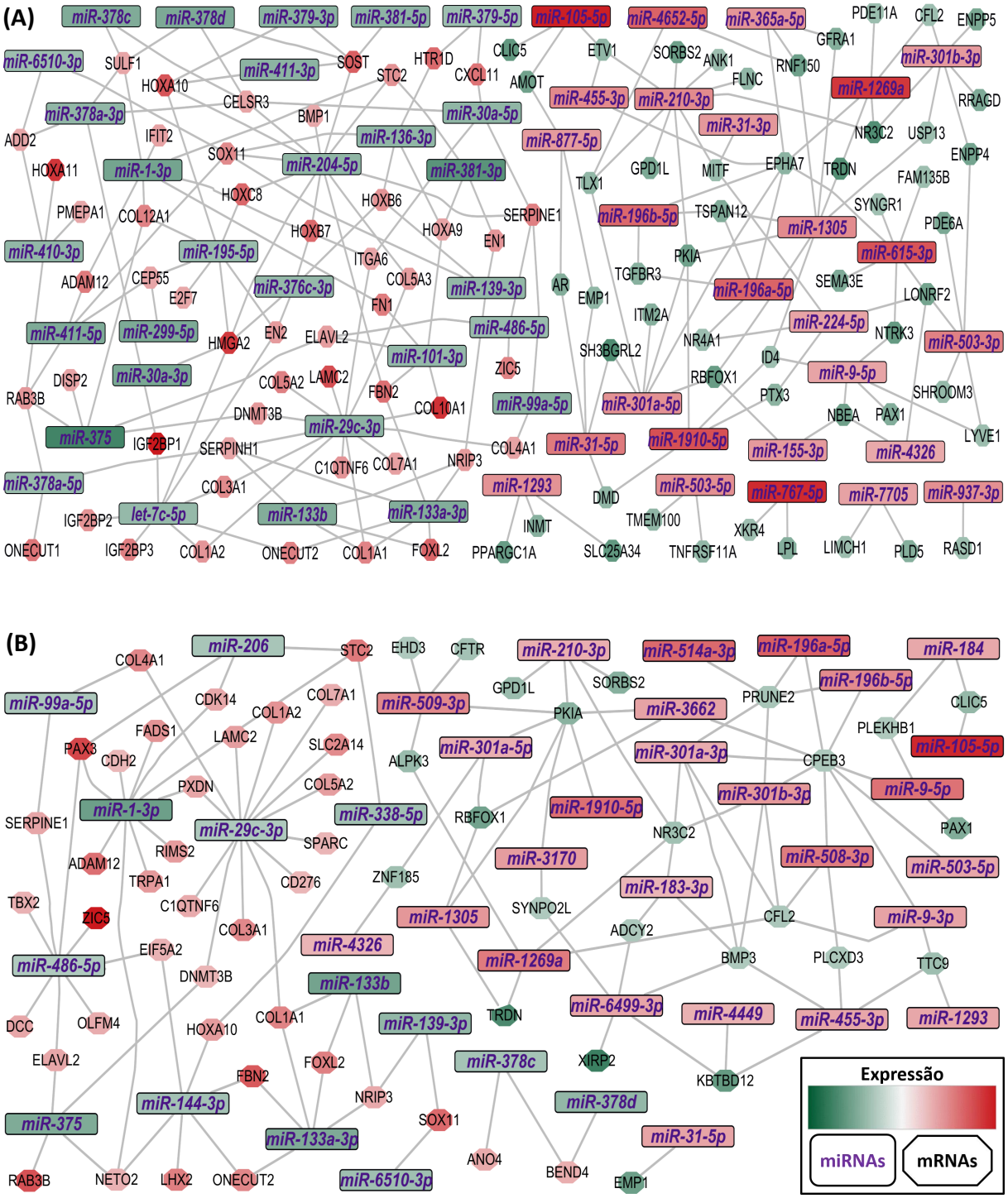


Figura 39 – Rede regulatória entre DEmiRNAs e DEmRNAs associadas ao câncer da cavidade oral (A) e laringe (B), respectivamente.

Por meio da Figura 39 (A), observa-se que na neoplasia da cavidade oral, a regulação

ocorre por meio de duas redes distintas, uma composta por 80 nós (29 DEmiRNAs subexpressos e 51 DEmRNAs superexpressos) e outra por 78 nós (26 DEmiRNAs superexpressos e 52 DEmRNAs subexpressos), respectivamente. Destaca-se que o miR-29c-3p subexpresso como o principal regulador, associado a 13 DEmRNAs superexpressos.

No câncer da laringe, a rede regulatória é composta por 53 nós (14 DEmiRNAs subexpressos e 39 DEmRNAs superexpressos) e 50 nós (26 DEmiRNAs superexpressos e 24 DEmRNAs subexpressos), respectivamente. Notavelmente, os miRNAs subexpressos, tais como miR-29c-3p, miR-1-3p e miR-486-5p, emergem como os reguladores mais proeminentes. Esses miRNAs estão associados a 12, 10 e 8 genes alvo superexpressos, respectivamente, conforme destacado na Figura 39 (B).

Na Figura 40 (A), torna-se evidente que, no estadiamento avançado do câncer de cavidade oral, os cinco DEmiRNAs identificados revelam que o miR-1269a superexpresso desempenha uma regulação específica sobre cinco genes (CFL2, EPHA7, NR3C2, PDE11A e TRDN), enquanto o miR-133a-3p subexpresso regula especificamente seis genes (COL1A1, FBN2, FOXL2, NRIP3, ONECUT2 e SERPIH1). Além disso, o miR-133b regula os genes superexpressos COL1A1, FOXL2, NRIP3 e SERPINH1, ao passo que o miR-381-5p exerce regulação apenas quando o gene SOST está superexpresso, enquanto no estadiamento inicial do câncer de laringe, observa-se que o gene CLIC5, quando subexpresso, é exclusivamente regulado pelo miR-105-5p. Entretanto, no estadiamento avançado desse tipo de câncer, os principais reguladores são os miR-1-3p, miR-486-5p, miR-144-3p e miR-133a-3p, todos subexpressos, como destacado na Figura 40 (B).

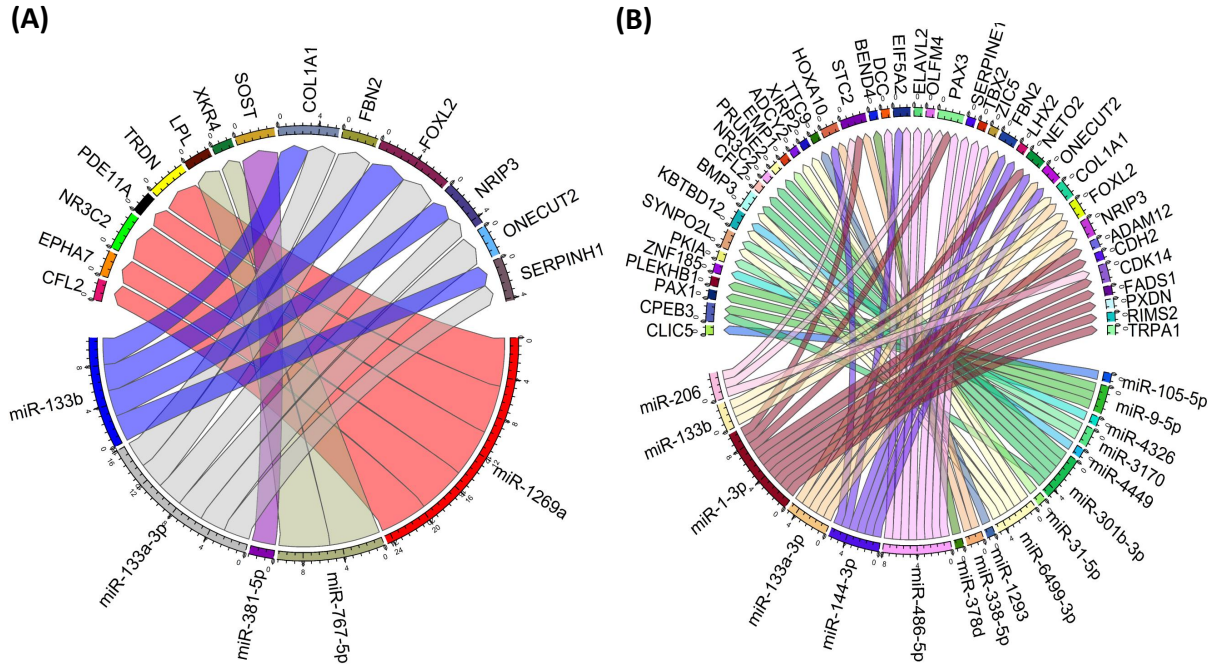


Figura 40 – Rede regulatória entre DEmiRNAs e DEmRNAs do estadiamento avançado do câncer da cavidade oral (A), inicial e avançado do câncer da laringe (B), respectivamente.

6.3.2 Análises de enriquecimento e de vias dos DEmRNAs

A Figura 41 apresenta as análises de enriquecimento de vias KEGG e anotação GO dos DEmRNAs associados ao câncer da cavidade oral e laringe, respectivamente.

Nas Figuras 41 (A) e (B), a análise das vias KEGG revelou a associação de 22 DEmRNAs a sete vias de sinalização significantemente enriquecidas no câncer da cavidade oral. Essas vias estão relacionadas a processos como digestão e absorção de proteínas, adesão focal e desregulação transcricional no câncer. Destaca-se que a desregulação dos genes TLX1, MITF, ETV1, HOXA9, HOXA11, HOXA10 e HMGA2 contribuiu para o enriquecimento da via de desregulação transcricional no câncer. Além disso, os genes LAMC2, ITGA6, FN1 e COL4A1 apresentaram enriquecimento específico no câncer de pulmão de células pequenas.

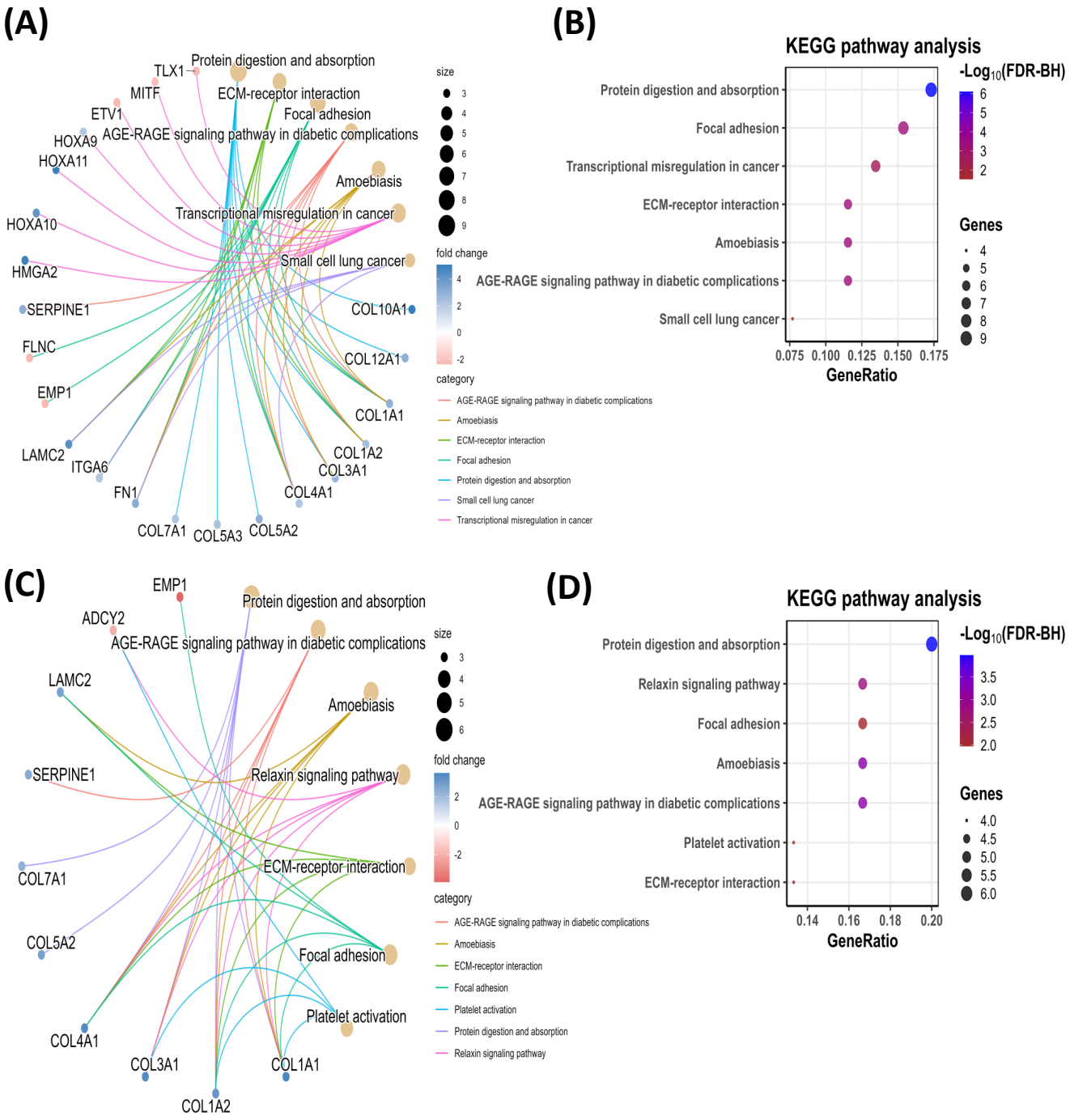
No contexto do câncer de laringe, os 10 DEmRNAs demonstraram significativo enriquecimento em sete vias moleculares distintas como digestão e absorção de proteínas e via de sinalização da relaxina. Ademais, a superexpressão dos genes LAMC2, COL4A1, COL1A2 e COL1A1 foi associada ao enriquecimento na via de interação entre células e a matriz extracelular (ECM), conforme demostrado nas Figuras 41 (C) e (D).

Por outro lado, a análise de enriquecimento ontológica GO associadas aos DEmRNAs do câncer da cavidade oral identificou a participação dos genes em 59 processos biológicos, destacando especialmente aqueles relacionados ao desenvolvimento, morfogênese e adesão à matriz celular. Esses resultados indicam a relevância desses genes na formação e função de vários tecidos, incluindo o sistema esquelético, a próstata e os músculos dos membros. Adicionalmente, a identificação de seis componentes celulares sugere o envolvimento significativo dos DEmRNAs na biossíntese de colágeno e na organização da matriz extracelular (MEC), sugerindo papéis essenciais no suporte estrutural, na adesão celular e nos processos de sinalização. A análise também revelou 12 funções moleculares, indicando que esses genes desempenham papéis diversos em processos, como cicatrização de feridas, angiogênese, inflamação, crescimento celular, diferenciação, proliferação e transdução de sinal (Figura 41 E) .

No contexto do câncer de laringe, foram identificados 42 processos biológicos, 11 componentes celulares e 7 funções moleculares. Os processos biológicos foram relacionados principalmente ao desenvolvimento embrionário, o crescimento e a regulação de vários tecidos, órgãos e sistemas. Os componentes celulares foram relacionados ao citoesqueleto, matriz extracelular e estruturas associadas ao tecido conjuntivo. A função molecular foi principalmente relacionada à composição e organização da matriz extracelular, contribuindo para a resistência à tração da mesma, sugerindo um papel na manutenção da integridade estrutural dos tecidos (Figura 41 F).

Nas Figuras 41 (G) e (H), apresentam-se a análise de enriquecimento da ontologia de doenças para o câncer da cavidade oral e laringe. Observa-se que os DEmRNAs estão relacionados a 22 e 8 doenças específicas, respectivamente. No câncer da cavidade oral, foram

identificadas associações significativas entre os DEmRNAs (COL1A1, COL7A1, ITGA6, LAMC2, SERPINH1) e condições cutâneas, como epidermólise bolhosa, doença de pele vesiculobolhosa e doenças bolhosas da pele. Quanto ao câncer da laringe, observaram-se principalmente associações com o síndrome de Ehlers-Danlos (COL1A1, COL1A2, COL3A1 e COL5A2), defeitos no colágeno (COL1A1, COL1A2, COL3A1, COL4A1 e COL5A2) e a infertilidade (CFTR, CDH2, EIF5A2, FOXL2, HOXA10, SERPINE1 e SOX11).



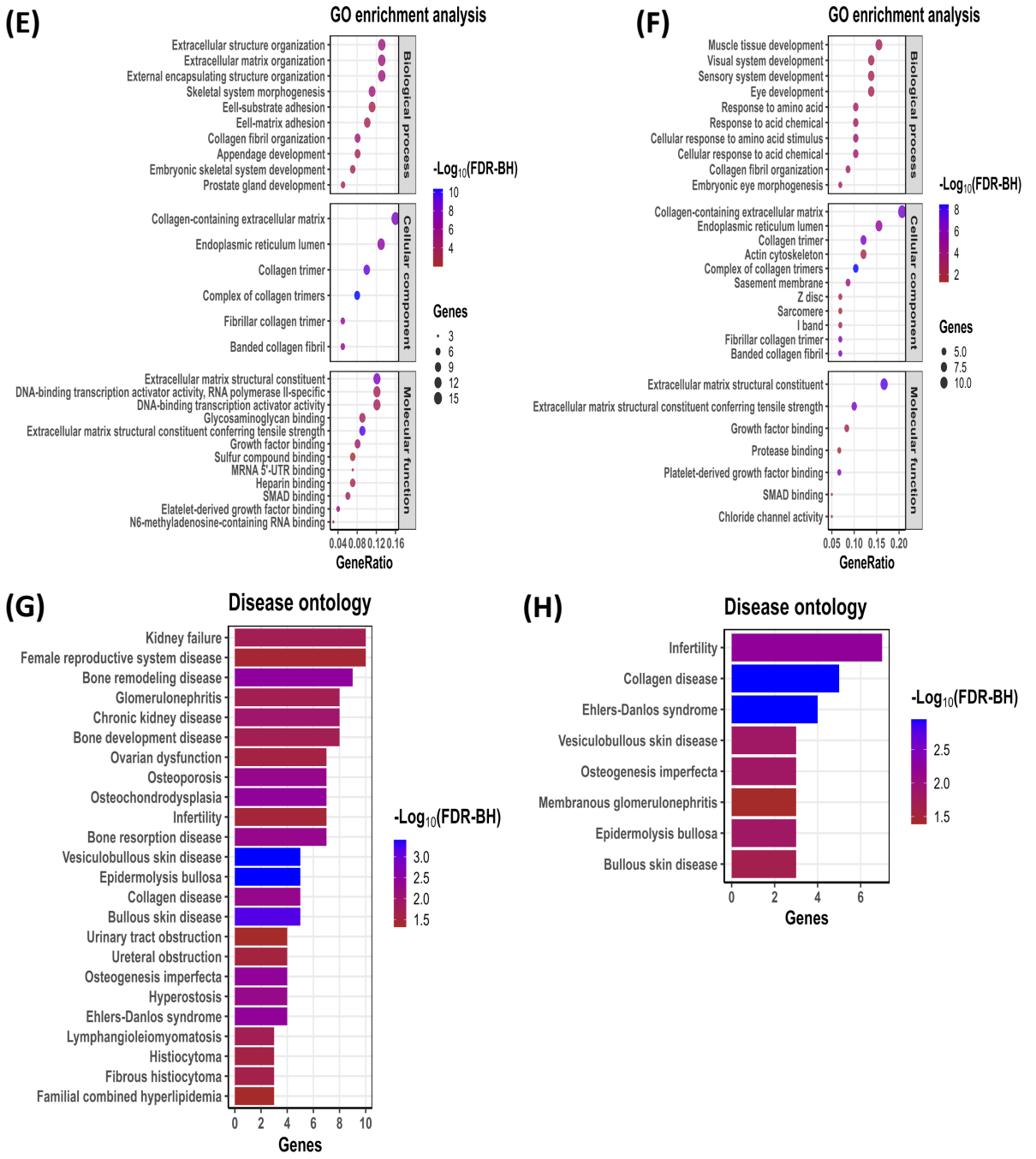


Figura 41 – Anotação funcional e análise de enriquecimento dos DEmRNAs. (A, B, C, D) Rede interativa de análise de enriquecimento e gráfico de bolhas exibindo as vias KEGG dos DEmRNAs no câncer da cavidade oral e laringe, respectivamente.(E, F) Gráfico de bolhas de enriquecimento do termo Ontologia Genética para cada um dos termos de instância (processo biológico, componente celular e função molecular) no câncer da cavidade oral e laringe, respectivamente. (G, H) Gráfico de barras da análise de enriquecimento da Ontologia de Doenças no câncer da cavidade oral e laringe, respectivamente.

6.3.3 Rede de interação proteína-proteína e redes regulatórias miRNA-*hub*

Com base na análise do banco de dados STRING com uma *combined score* $> 0,4$, foi observado que, para o câncer da cavidade oral, a rede PPI possui 103 nós e 175 arestas. O número esperado de arestas era 45, com um grau médio de nós de 3,4. O coeficiente médio de agrupamento local foi de 0,358, indicando uma relevância regulatória significativa dessa proteína. Além disso, o valor de enriquecimento PPI foi inferior a $1,0 \times 10^{-16}$, sugerindo uma conexão biológica significativa entre as proteínas. A interação com a pontuação combinada mais alta ocorreu entre as proteínas COL1A1 e COL3A1 (*combined score* $= 0,999$) (Figura 42 A). Com base nessa análise, identificou-se 10 genes *hub* que foram regulados positivamente, sendo o gene COL1A1 quem apresentou o maior grau de conectividade (*score* $= 7305486$) (Figura 42 B). Esses genes, em conjunto, desempenham um papel na síntese de diferentes tipos de colágeno e outras proteínas da matriz extracelular, sendo essenciais para manter a integridade estrutural e a função de vários tecidos no corpo.

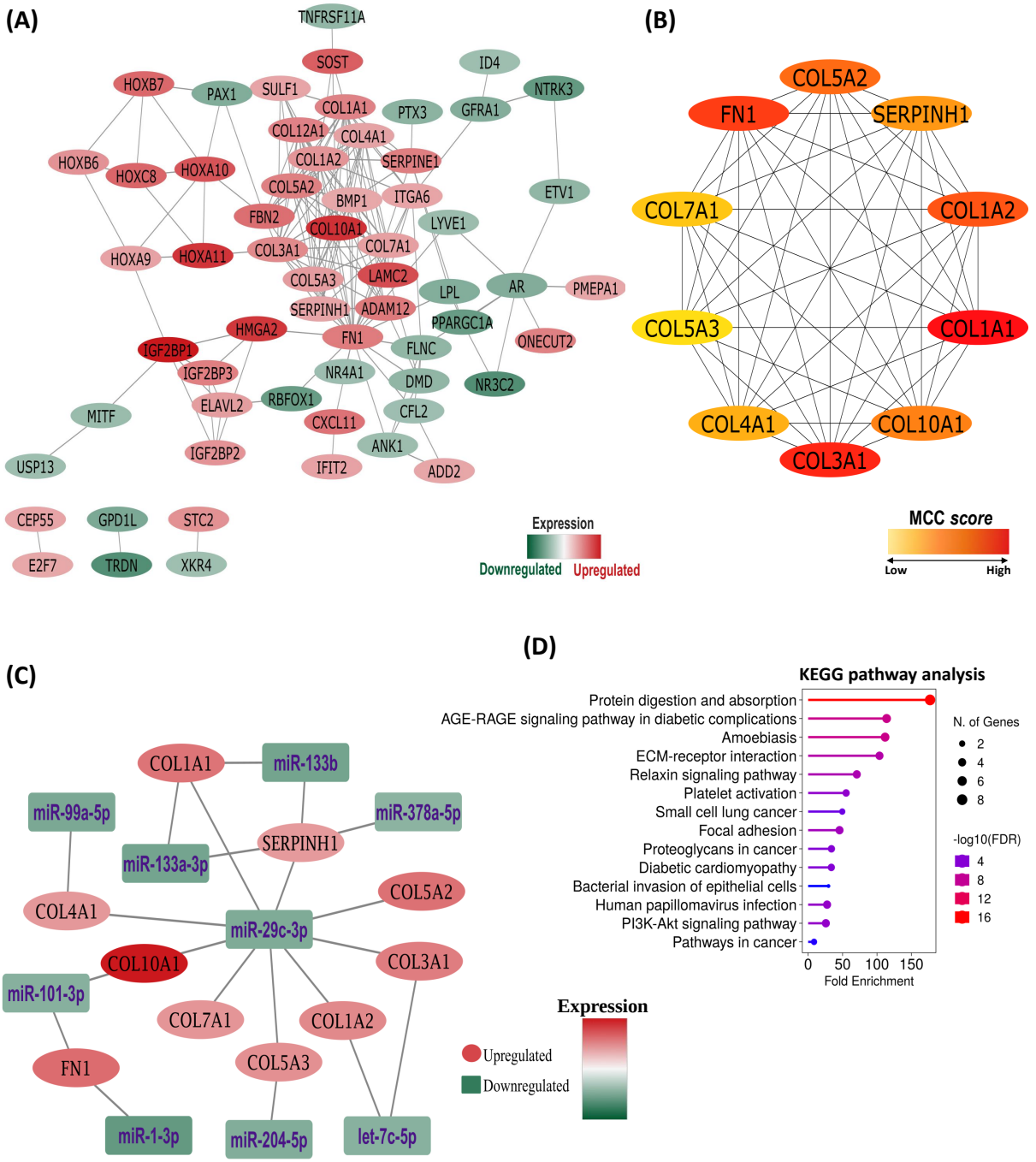
Em seguida, Figura 42 (C) mostra-se a rede regulatórias do miRNA-*hub*. Observa-se que, dos 10 genes *hub*, nove foram identificados como *hubs* potenciais (COL1A1, SERPINH1, COL5A2, COL3A1, COL1A2, COLA3, COL7A1, COL10A1, COL4A1) que poderiam ser modulados negativamente pelo miR-29c-3p. Na análise de enriquecimento dos genes *hub*, foram identificadas 14 vias KEGG, abrangendo processos como digestão e absorção de proteínas, interação ECM-receptor, adesão focal, AGE-RAGE em complicações diabéticas, amebíase, infecção por papilomavírus humano, via de sinalização PI3K-Akt, câncer de pulmão de pequenas células, via de sinalização de relaxina, ativação plaquetária, proteoglicanos no câncer, cardiomiopatia diabética, invasão bacteriana de células epiteliais e vias no câncer (Figura 42 D).

Para o câncer da laringe, a rede PPI consistiu em 63 nós e 65 arestas, com um número esperado de arestas de 17. O grau médio dos nós foi de 2,06, indicando que cada nó tinha, em média, pelo menos 2,06 nós interagindo. O coeficiente médio de agrupamento local foi de 0,38, e o valor de enriquecimento PPI foram significativas, sugerindo novamente uma significativa conexão biológica entre as proteínas neste contexto. A interação com a pontuação combinada mais alta ocorreu entre as proteínas COL1A1 e COL3A1 (*combined score* $= 0,999$) (Figura 42 E). Dessa forma, identificou-se 10 genes *hub* superexpressos, sendo que os genes COL1A1, COL3A1, COL4A1 e COL1A2 apresentaram o maior grau de conectividade (Figura 42 F). Esses genes estão associados à síntese e manutenção de componentes da matriz extracelular, desempenhando papéis cruciais na estrutura e função dos tecidos.

Na, Figura 42 (G) mostra-se a rede regulatória do miRNA-*hub*. Observa-se que, oito genes *hub* potenciais (COL5A2, COL4A1, COL3A1, COL1A1, SPARC, COL1A2, COL7A1 e PXDN) são modulados negativamente pelo miR-29c-3p. A análise de enriquecimento identificou 11 vias KEGG para os genes *hub*, incluindo digestão e absorção de proteínas, via de sinalização AGE-RAGE em complicações diabéticas, amebíase, interação ECM-receptor, via de sinalização de relaxina, ativação plaquetária, adesão focal, cardiomiopatia diabética, proteoglicanos no câncer,

infecção por papilomavírus humano e via de sinalização PI3K-Akt (Figura 42 H).

Ambos os tipos de câncer compartilham vias comuns, como digestão e absorção de proteínas, interação ECM-receptor, adesão focal, AGE-RAGE em complicações diabéticas, amebíase, infecção por papilomavírus humano, via de sinalização PI3K-Akt, ativação plaquetária, proteoglicanos no câncer e cardiomiopatia diabética.



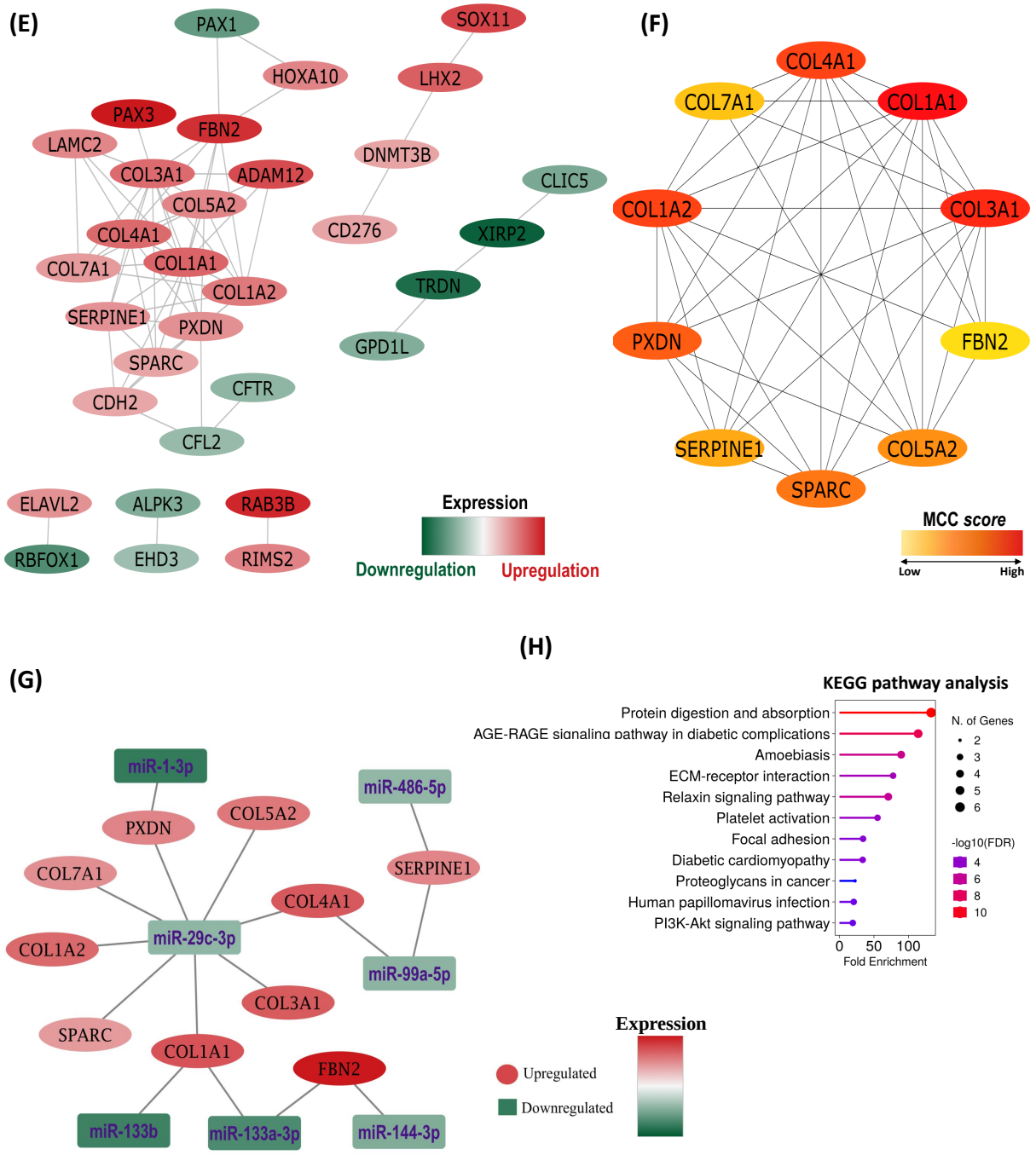


Figura 42 – (A, E) Rede de interação proteína-proteína (PPI) associada ao câncer da cavidade oral e laringe, respectivamente. (B, F) *Hub* genes referentes ao câncer da cavidade oral e laringe, respectivamente. A cor vermelha a amarela representa o grau de conectividade de cima para baixo baseada no algoritmo de *Maximal Clique Centrality* (MCC). (C, G) Rede regulatória entre DEmiRNAs e *hub* genes referentes ao câncer da cavidade oral e laringe, respectivamente. (D, H) Enriquecimento da via KEGG dos *hubs* referentes ao câncer da cavidade oral e laringe, respectivamente.

6.4 Discriminação de tecidos por meio de curvas ROC: avaliando a expressão individual de miRNAs

Com base apenas nos miRNAs da rede de interação DEmiRNAs-DEmRNAs, foi realizada uma análise da curva ROC para avaliar o poder diagnóstico de cada miRNA para diferenciar o tumor primário do tecido normal adjacente. A análise ROC mostrou que a maioria dos miRNAs têm boa capacidade diagnóstica, independentemente do câncer da cavidade oral e laringe, com uma AUC superior a 70% ao distinguir o tumor primário do tecido normal adjacente, como mostrado na Tabela 10.

Tabela 10 – Análise das curvas de ROC na discriminação do tumor primário e do tecido normal adjacente ao tumor no câncer da cavidade oral e laringe.

Cavidade oral					Laringe				
miRNAs	AUC ($IC_{95\%}$)	*Cuttof*	TNR (%)	TPR (%)	miRNAs	AUC ($IC_{95\%}$)	*Cuttof*	TNR (%)	TPR (%)
miR-101-3p↓	98,1 (95,6 - 100,0)	16,2	89,3	98,6	miR-99a-5p↓	97,6 (94,9 - 100,0)	11,8	100,0	88,0
miR-99a-5p↓	97,6 (95,4 - 99,8)	11,5	92,2	95,2	miR-196b-5p↑	96,5 (93,3 - 99,7)	7,8	100,0	89,0
miR-455-3p↓	97,2 (95,3 - 99,2)	9,1	96,4	92,2	miR-139-3p↓	96,4 (92,2 - 100,0)	7,1	83,3	96,0
miR-30a-5p↓	96,7 (94,6 - 98,7)	15,2	100,0	85,0	miR-6510-3p↓	93,6 (87,4 -99,7)	7,4	83,3	93,0
miR-196b-5p↑	96,1 (93,5 - 98,7)	7,0	96,4	86,3	miR-301a-3p↑	93,5 (87,2 - 99,8)	4,0	83,3	83,0
let-7c-5p↓	96,0 (92,9 - 99,1)	13,7	92,9	88,4	miR-29c-3p↓	93,1 (86,3 - 99,8)	14,2	91,7	87,0
miR-29c-3p↓	95,7 (92,9 - 98,6)	13,7	89,3	91,8	miR-455-3p↑	93,0 (86,4 - 99,6)	8,2	83,3	93,0
miR-204-5p↓	95,4 (93,0 - 97,7)	4,7	96,4	89,4	miR-338-5p↓	93,0 (83,0 - 100)	5,3	83,3	97,0
miR-30a-3p↓	95,3 (91,9 - 98,8)	13,9	89,3	89,8	miR-375↓	92,7 (87,4 - 97,9)	14,0	91,7	90,0
miR-195-5p↓	95,2 (92,5 - 98,0)	7,7	85,7	93,5	miR-503-5p↑	92,1 (84,9 - 99,3)	3,1	91,7	82,0
miR-136-3p↓	95,0 (91,7 - 98,3)	6,4	89,3	90,1	miR-1910-5p↑	91,4 (86,6 - 96,3)	1,5	100,0	80,0
miR-615-3p↑	94,6 (91,6 - 97,7)	1,3	89,3	91,1	miR-196a-5p↑	88,7 (82,6 - 94,7)	5,6	100,0	82,0
miR-139-3p↓	94,3 (91,2 - 97,4)	6,5	92,9	83,6	miR-210-3p↑	88,0 (78,1 - 97,9)	10,7	91,7	79,0
miR-503-5p↑	93,4 (89,3 - 97,5)	3,6	85,7	88,4	miR-301b-3p↑	87,1 (79,3 - 94,9)	1,2	83,3	81,0
miR-1910-5p↑	92,7 (89,3 - 96,1)	1,1	92,9	84,6	miR-301a-5p↑	86,5 (75,1 - 96,9)	2,6	83,3	82,0
miR-877-5p↑	92,0 (88,7 - 95,3)	2,6	100	77,1	miR-486-5p↓	86,3 (74,1 - 98,6)	9,5	91,7	77,0
miR-410-3p↓	91,5 (87,0 - 96,1)	6,6	92,9	80,9	miR-378c↓	86,2 (76,9 - 95,5)	5,5	91,7	74,0
miR-196a-5p↑	91,2 (87,0 - 95,4)	6,0	92,9	77,8	miR-3662↑	83,2 (75,0 - 91,5)	1,9	100,0	66,0
miR-411-5p↓	90,7 (85,5 - 95,9)	6,5	78,6	90,8	miR-183-3p↑	83,0 (73,9 - 92,1)	2,2	100,0	62,0
miR-378c↓	90,6 (85,2 - 96,0)	5,9	85,7	82,9	miR-3170↑	82,2 (73,2 - 91,1)	1,2	91,7	67,0
miR-210-3p↑	90,4 (85,6 - 95,2)	11,3	96,4	67,9	miR-4326↑	81,3 (70,4 - 92,2)	2,9	83,3	79,0
miR-379-5p↓	90,1 (84,1 - 96,2)	13,0	85,7	81,9	miR-1293↓	81,0 (72,4 - 89,6)	3,0	100,0	65,0
miR-375↓	90,1 (82,2 - 98,1)	12,9	89,3	89,8	miR-378d↓	80,9 (69,3 - 92,6)	2,1	83,3	72,0
miR-365a-5p↑	89,8 (85,7 - 94,0)	2,0	85,7	85,3	miR-9-5p↑	80,3 (70,6 - 90,1)	11,4	91,7	65,0
miR-224-5p↑	89,6 (85,0 - 94,2)	8,4	96,4	71,3	miR-1-3p↓	79,9 (69,7 - 90,1)	6,0	91,7	67,0
miR-301a-5p↑	89,4 (84,6 -94,3)	2,8	89,3	76,5	miR-144-3p↓	78,5 (58,8 - 98,2)	6,3	83,3	86,0
miR-376c-3p↓	89,4 (82,5 - 96,2)	5,8	82,1	89,8	miR-133a-3p↓	76,8 (64,3 - 89,4)	5,5	91,7	64,0
miR-381-3p↓	89,3 (83,1 - 95,6)	8,9	85,7	80,9	miR-31-5p↑	76,6 (66,7 - 86,4)	6,5	91,7	65,0
miR-503-3p↑	89,0 (84,8 - 93,1)	1,5	92,9	78,8	miR-133b↓	76,3 (64,8 - 87,9)	3,4	91,7	64,0
miR-378a-3p↓	88,2 (82,4 - 93,9)	12,0	96,4	63,8	miR-6499-3p↑	74,0 (59,8 - 88,2)	2,2	66,7	71,0
miR-31-5p↑	87,9 (83,5 - 92,3)	7,1	92,9	79,2					
miR-4652-5p↑	87,3 (82,6 - 92,0)	2,1	96,4	77,8					
miR-937-3p↑	87,1 (81,7 - 92,4)	2,8	89,3	76,5					
miR-1293↑	85,4 (78,0 -92,7)	3,1	82,1	79,9					
miR-7705↑	85,2 (79,8 - 90,6)	1,8	89,3	73,4					
miR-379-3p↓	84,4 (75,6 - 93,3)	2,6	85,7	74,7					
miR-31-3p↑	83,7 (77,3 - 90,1)	6,7	89,3	77,5					
miR-299-5p↓	83,4 (75,1 - 91,6)	5,0	85,7	68,3					
miR-378d↓	82,6 (72,4 - 92,9)	3,4	64,3	92,8					
miR-378a-5p↓	82,4 (73,9 - 90,8)	9,7	64,3	85,3					
miR-1305↑	82,0 (75,9 - 88,2)	1,1	85,7	74,1					
miR-486-5p↓	80,6 (72,3 - 89,0)	10,0	71,4	74,1					
miR-6510-3p↓	78,8 (68,1 - 89,5)	6,2	64,3	87,4					
miR-155-3p↑	78,6 (70,9 - 86,3)	0,9	67,9	78,5					
miR-4326↑	76,8 (70,1 - 83,4)	3,1	78,6	74,7					

Nota: AUC - Área sob a curva; TNR - Especificidade; TPR - Sensibilidade; IC - Intervalo de confiança; ↑ (Superexpresso); ↓ (Subexpresso).

6.4.1 Classificadores baseados em árvore de decisão no diagnóstico de câncer da cavidade oral e laringe usando a expressão de miRNAs

A Figura 43 apresenta um mapa de calor que representa a árvore de classificação dos tecidos (tumor primário e normal adjacente) usando os miRNAs da rede regulatória de DEmiRNAs-DEmRNAs do câncer da cavidade oral sem balanceamento. A árvore é composta por três miRNAs (miR-101-3p, miR-455-3p e miR-375), que demonstraram um desempenho excepcional (AUC = 0,90) e uma acurácia superior a 98%, indicando excelente classificação dos tecidos. A cor vermelho escuro indica a probabilidade de classificação ser do tumor primário, enquanto a cor azul escuro indica a probabilidade de a classificação ser do tecido normal adjacente.

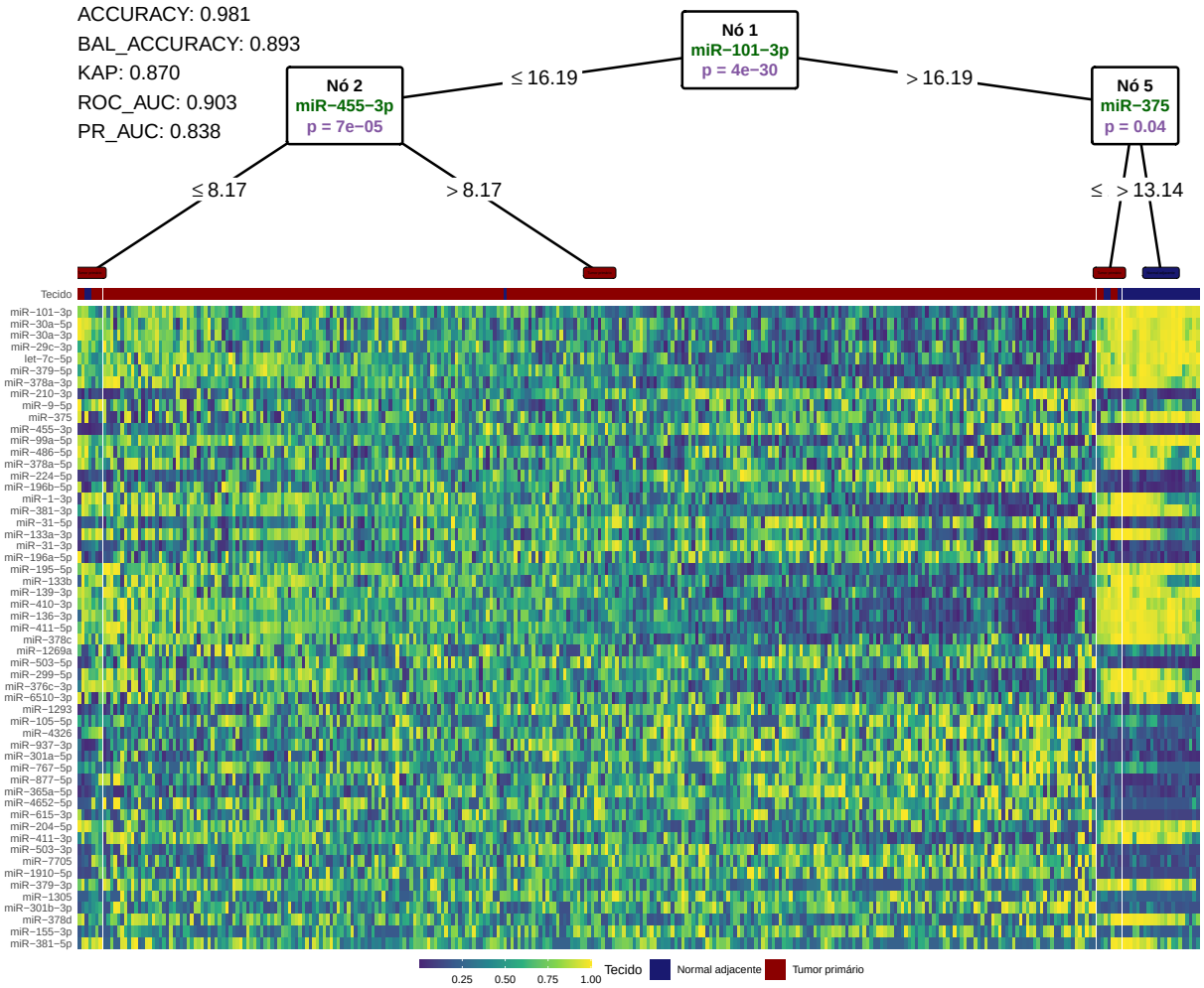


Figura 43 – *Heatmap* baseada na árvore de decisão para classificação de tecidos do câncer da cavidade oral, considerando todos os estadiamentos patológicos. As cores do mapa apresentam o valor relativo da expressão dos miRNAs em cada amostra.

O miR-101-3p é a principal fonte de variabilidade nas amostras e representa o preditor mais significativo nos tecidos tumorais. O nó 5 (miR-375) da árvore classificou o tumor primário e normal adjacente, evidenciando uma diferenciação entre as cores amareladas e azul-escuro em relação aos outros nós. O que sugere um padrão de expressão diferencial dos miRNAs nessas

amostras. Também é possível notar que os pares miR-30a-5p e miR-30a-3p, bem como miR-29c-3p e let-7c-5p, exibem cores similares em várias amostras, sugerindo que provavelmente esses pares de miRNAs podem ser co-expressos em conjunto nas diversas condições biológicas.

Na Tabela 11 são apresentadas a matriz de confusão e o desempenho das árvores de decisão CTree e do RFT, obtidos por meio do conjunto de teste e utilizando diferentes técnicas de balanceamento. O processo foi realizado utilizando *Five-Fold Cross-Validation* com 10 repetições, abrangendo todos os estadiamentos (I - IV), tanto inicial (I - II) quanto avançado (III - IV), no conjunto de treinamento. Destaca-se que a técnica de balanceamento SMOTE demonstrou o melhor desempenho tanto no CTree quanto no RFT em diferentes estadiamento. As métricas do RFT alcançaram valores iguais a 1, indicando um desempenho excelente.

Tabela 11 – Matriz de confusão e avaliação de desempenho dos modelos baseados nas árvores de decisão (CTree e floresta aleatória) utilizando *Five-Fold Cross-Validation* na classificação binária de tumor primário (1) e tecido normal adjacente (0) em relação ao câncer de cavidade oral com diferentes técnicas de balanceamento.

	CTree (única árvore)										floresta aleatória									
Métodos	Matriz de confusão			Métricas[a]							Matriz de confusão			Métricas[a]						
	Predito	Real **0**	Real **1**	AUC	TNR	TPR	ACC	PR	F1	KAP	Predito	Real **0**	Real **1**	AUC	TNR	TPR	ACC	PR	F1	KAP
Todos os estadiamentos (I - IV)																				
Sem balanceamento	**0**	5	2	0,93	1,00	0,97	0,98	1,00	0,99	0,82	**0**	5	2	0,93	1,00	0,97	0,98	1,00	0,99	0,82
	1	0	73								**1**	0	73							
Undersampling	**0**	6	1	0,93	1,00	0,88	0,93	1,00	0,93	0,86	**0**	7	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	0	7								**1**	0	7							
Oversampling	**0**	73	0	0,99	0,90	1,00	0,95	0,89	0,95	0,89	**0**	73	0	1,00	0,99	1,00	0,99	0,99	0,99	0,99
	1	8	65								**1**	1	72							
SMOTE	**0**	68	2	0,98	0,93	0,97	0,95	0,93	0,95	0,90	**0**	70	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	5	68								**1**	0	73							
ROSE	**0**	34	7	0,93	0,89	0,83	0,86	0,90	0,86	0,73	**0**	37	1	1,00	0,97	0,98	0,97	0,98	0,97	0,95
	1	4	35								**1**	1	40							
Estadiamento patológico inicial (I - II)																				
Sem balanceamento	**0**	2	0	0,97	0,67	1,00	0,95	0,94	0,80	0,77	**0**	2	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	1	17								**1**	0	18							
Undersampling	**0**	2	0	0,50	0,50	*	0,50	0,00	*	0,00	**0**	2	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	2	0								**1**	0	2							
Oversampling	**0**	18	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00	**0**	18	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	0	18								**1**	0	18							
SMOTE	**0**	16	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00	**0**	16	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	0	18								**1**	0	18							
ROSE	**0**	10	0	0,98	0,83	1,00	0,90	0,82	0,91	0,81	**0**	10	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	2	9								**1**	0	10							
Estadiamento patológico avançado (III - IV)																				
Sem balanceamento	**0**	3	1	0,99	1,00	0,98	0,98	1,00	0,99	0,85	**0**	2	2	0,99	1,00	0,96	0,97	1,00	0,98	0,65
	1	0	55								**1**	0	55							
Undersampling	**0**	3	1	0,87	1,00	0,80	0,88	1,00	0,89	0,75	**0**	3	1	1,00	1,00	0,80	0,88	1,00	0,89	0,75
	1	0	4								**1**	0	4							
Oversampling	**0**	55	0	1,00	0,93	1,00	0,96	0,93	0,96	0,93	**0**	55	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	4	51								**1**	0	55							
SMOTE	**0**	50	1	1,00	0,98	0,98	0,98	0,98	0,98	0,96	**0**	51	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	1	54								**1**	0	55							
ROSE	**0**	21	7	0,96	0,95	0,81	0,86	0,97	0,88	0,73	**0**	27	1	1,00	0,96	0,97	0,97	0,97	0,97	0,93
	1	1	30								**1**	1	30							

Nota: * Indeterminado

[a] AUC - Área sob a curva, TNR - Especificidade, TPR - Sensibilidade / *Recall*, ACC - Acurácia, PR - precisão, F1- F1-*score*, KAP - Capa de Cohen.

Na Figura 44 os resultados da árvore de decisão gerados pelos algoritmos CTree e RFT que classificam todos os miRNAs de acordo com a sua importância, utilizando o método de balanceamento SMOTE, são apresentados para todos os casos.

De acordo com a árvore de decisão baseada na classificação da Figura 44 (A), os miR-101-3p (Nó 1), miR-455-3p (Nó 2 e Nó 4), miR-615-3p (Nó 7) e let-7c-5p (Nó 8) são biomarcadores relevantes na diferenciação dos tecidos no câncer de cavidade oral, considerando todos os estadiamentos patológicos (I - IV). Especificamente, quando os níveis de expressão do miR-101-3p ultrapassam 15,91 e os do miR-615-3p excedem 1,63, há uma probabilidade de 75% do tecido ser considerado normal adjacente e 25% de ser classificado como tumoral primário. No entanto, se a expressão for igual ou inferior a 1,63 para miR-615-3p e superior a 14,07 para let-7c-5p, o tecido seria classificado como normal adjacente. A Figura 44 (B) mostra os 15 primeiros miRNAs obtidos pelo algoritmo RFT. De acordo com esta classificação o miR-101-3p na discriminação dos tecidos.

A Figura 44 (C) mostra que, no estadiamento inicial da doença (I - II), apenas o miR-101-3p é capaz de classificar ambos os tecidos. Das 105 amostras avaliadas, aquelas com níveis de expressão menores ou iguais a 16,02 foram classificadas como tumor primário, enquanto as demais foram consideradas amostras de tecido normal adjacente. A Figura 44 (D) mostra que os miR-99a-5p e miR-101-3p foram os mais importantes na discriminação dos tecidos nesse estadiamento.

A Figura 44 (E) mostra que a árvore no estadiamento avançado (III - IV) do câncer da cavidade oral é composta por cinco miRNAs (miR-101-3p, miR-455-3p, miR-937-3p, miR-615-3p e miR-1305). As amostras de câncer da cavidade oral foram classificadas como tumor primário se a expressão de miR-101-3p e miR-455-3p fosse menor ou igual a 16,19 e maior que 8,82, respectivamente. Caso contrário, se a expressão de miR-937-3p fosse menor ou igual a 8,82 e menor ou igual a 1,14, as amostras foram classificadas como tecido normal adjacente. A Figura 44 (F) apresenta os 15 miRNAs mais importantes, sendo o miR-30a-5p o mais importante na diferenciação dos tecidos.

Nota-se que, o miR-101-3p subexpresso e o miR-455-3p superexpresso foram identificados como os mais significativos usando os três métodos de expressão diferencial (DEseq2, edgeR e o teste de Wilcoxon-Mann-Whitney). Além disso, a capacidade de discriminação individual dos miR-101-3p, hsa-miR-455-3p, miR-615-3p e let-7c-5p foram superiores a 90%, com exceção dos miR-937-3p e miR-1305 que apresentaram AUC de 87,1% e 82%, respectivamente, conforme mostrado na Tabela 10.

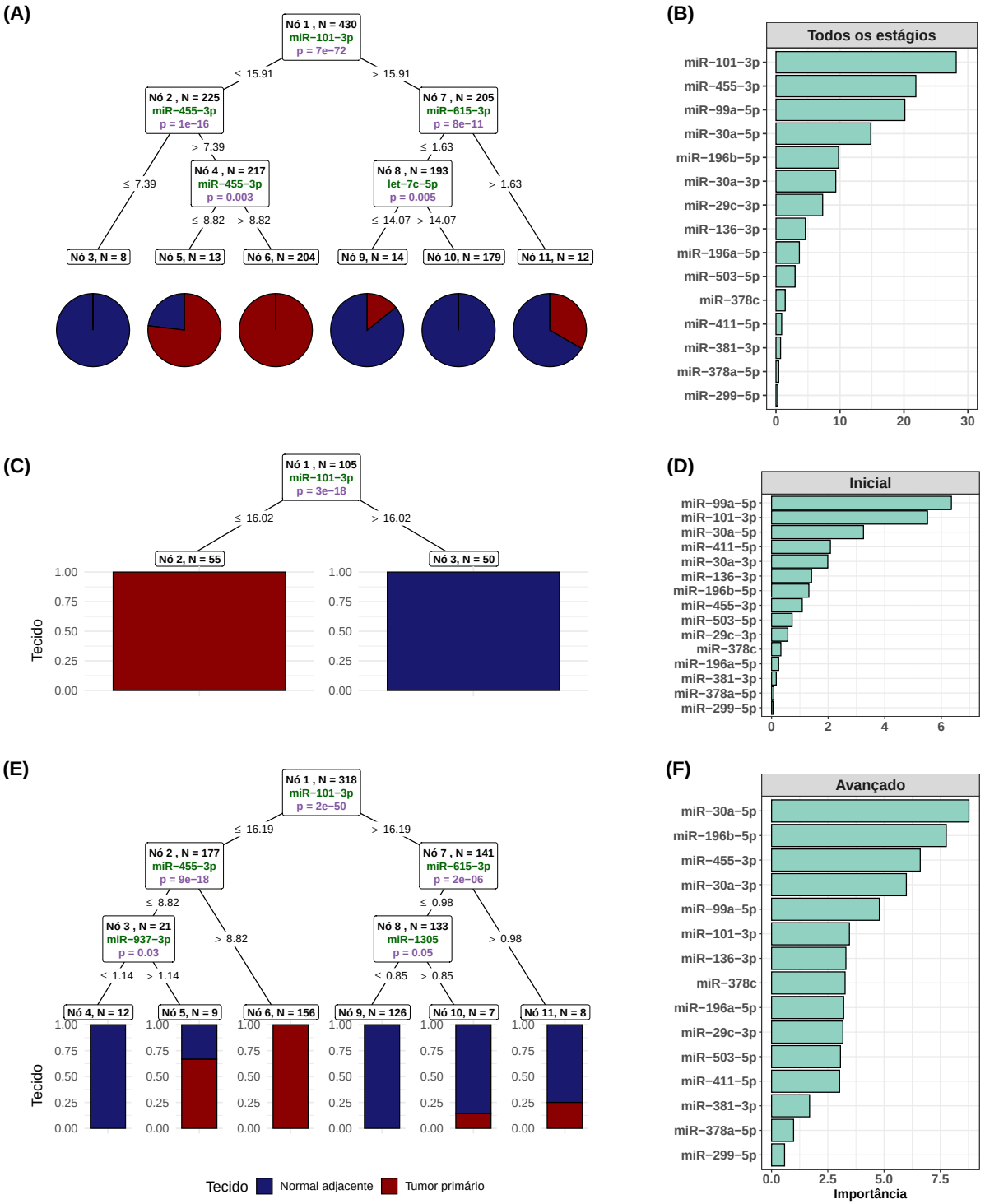


Figura 44 – Classificação dos tecidos utilizando CTree e floresta aleatória, abrangendo todos os estadiamentos patológicos (A, B), iniciais (C, D) e avançados (E, F) do câncer da cavidade oral.

A Figura 45 exibe o *heatmap* representando a árvore de classificação para todos os estadiamento do câncer de laringe. Esta árvore revela que apenas dois miRNAs desempenham um papel na classificação dos tecidos, alcançando uma acurácia de 94,6% e uma classificação aceitável (AUC = 79,4%). Além disso, nota-se um padrão distintivo nos miRNAs, muitos dos

quais são coexpressos em diversas amostras.

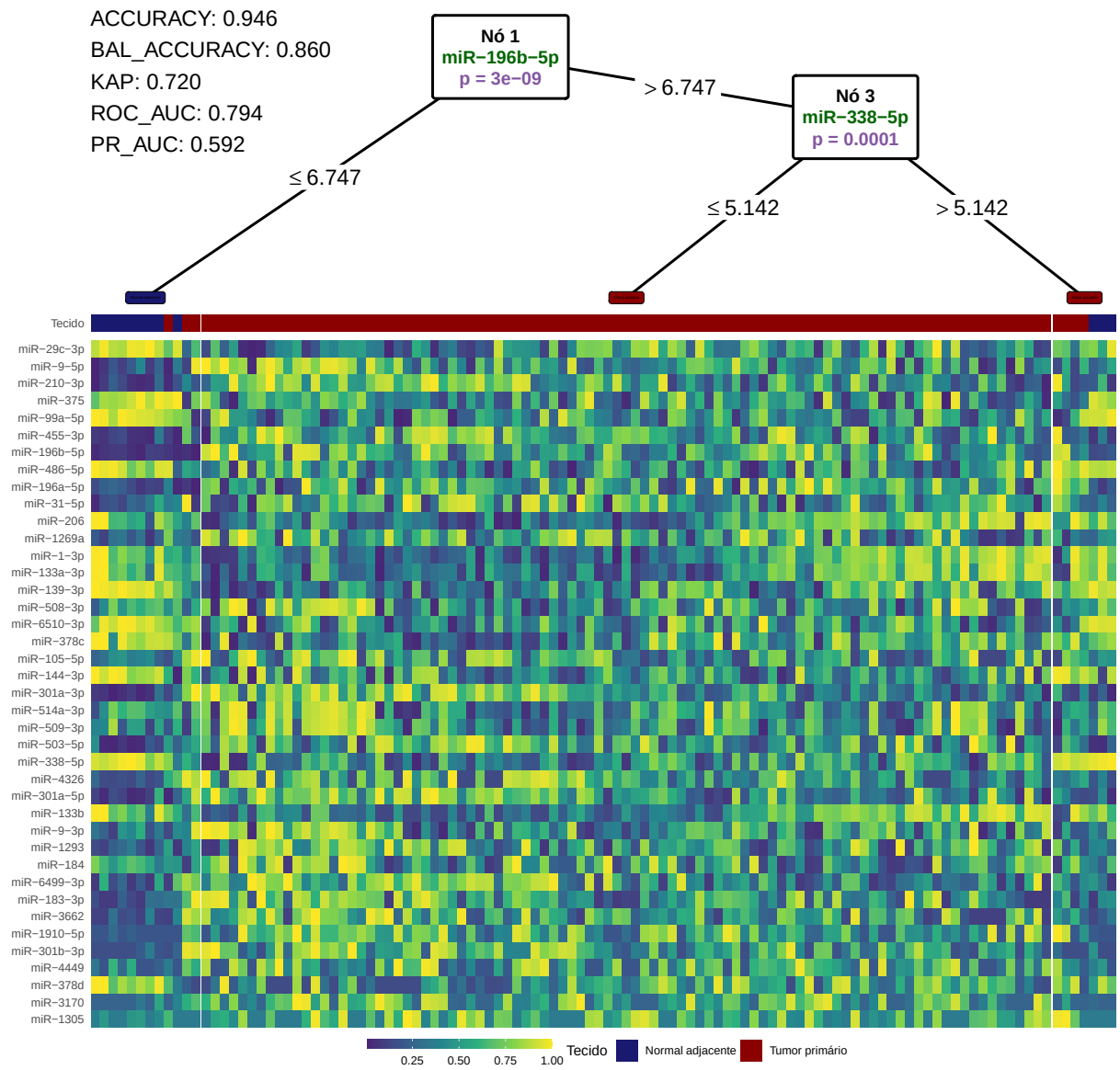


Figura 45 – *Heatmap* baseada na árvore de decisão para classificação de tecidos do câncer da laringe, considerando todos os estadiamentos patológicos. As cores do mapa apresentam o valor relativo da expressão dos miRNAs em cada amostra.

Na Tabela 12 são apresentadas a matriz de confusão e o desempenho das árvores de decisão CTree e do RFT, obtidos por meio do conjunto de teste e utilizando diferentes técnicas de balanceamento no câncer de laringe. O processo foi realizado utilizando *Five-Fold Cross-Validation* com 10 repetições, para todos os estadiamentos, tanto inicial quanto avançado, no conjunto de treinamento.

Destaca-se que a técnica de balanceamento SMOTE demonstrou o melhor desempenho tanto para o CTree quanto o RFT em todos os casos. As métricas do RFT alcançaram valores iguais a 1, indicando um desempenho excepcional.

Tabela 12 – Matriz de confusão e avaliação de desempenho dos modelos baseados nas árvores de decisão (CTree e floresta aleatória) utilizando *Five-Fold Cross-Validation* na classificação binária de tumor primário (1) e tecido normal adjacente (0) em relação ao câncer da laringe com diferentes técnicas de balanceamento.

Métodos	CTree (única árvore)										floresta aleatória									
	Matriz de confusão			Métricas[a]							Matriz de confusão			Métricas[a]						
	Predito	Observado		AUC	TNR	TPR	ACC	PR	F1	KAP	Predito	Observado		AUC	TNR	TPR	ACC	PR	F1	KAP
		0	**1**									**0**	**1**							
Todos os estadiamentos (I - IV)																				
Sem balanceamento	**0**	2	1	0,96	0,67	0,96	0,93	0,96	0,79	0,63	**0**	3	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	1	24								**1**	0	25							
Undersampling	**0**	3	0	0,50	0,50	*	0,50	0,00	*	0,00	**0**	3	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	3	0								**1**	0	3							
Oversampling	**0**	25	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00	**0**	25	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	0	25								**1**	0	25							
SMOTE	**0**	22	2	0,99	1,00	0,93	0,96	1,00	0,96	0,92	**0**	24	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	0	25								**1**	0	25							
ROSE	**0**	11	2	0,88	0,79	0,85	0,81	0,79	0,81	0,63	**0**	13	0	1,00	0,93	1,00	0,96	0,93	0,96	0,93
	1	3	11								**1**	1	13							
Estadiamento patológico inicial (I - II)																				
Sem balanceamento	**0**	0	1	0,50	*	0,75	0,75	1,00	*	0,00	**0**	1	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	0	3								**1**	0	3							
Undersampling	**0**	1	0	0,50	0,50	*	0,50	0,00	*	0,00	**0**	1	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	1	0								**1**	0	1							
Oversampling	**0**	3	0	0,50	0,50	*	0,50	0,00	*	0,00	**0**	3	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	3	0								**1**	0	3							
SMOTE	**0**	3	0	0,50	0,50	*	0,50	0,00	*	0,00	**0**	3	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	3	0								**1**	0	3							
ROSE	**0**	2	0	0,50	0,50	*	0,50	0,00	*	0,00	**0**	2	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	2	0								**1**	0	2							
Estadiamento patológico avançado (III- IV)																				
Sem balanceamento	**0**	1	0	0,93	0,25	1,00	0,87	0,86	0,40	0,36	**0**	1	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	3	19								**1**	0	22							
Undersampling	**0**	1	0	0,50	0,04	*	0,04	0,00	*	0,00	**0**	1	0	1,00	0,50	1,00	0,96	0,95	0,67	0,65
	1	22	0								**1**	1	21							
Oversampling	**0**	22	0	0,98	0,96	1,00	0,98	0,95	0,98	0,95	**0**	22	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	1	21								**1**	0	22							
SMOTE	**0**	21	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00	**0**	21	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	0	22								**1**	0	22							
ROSE	**0**	8	3	0,97	1,00	0,80	0,87	1,00	0,89	0,74	**0**	11	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	1	0	12								**1**	0	12							

Nota: * Indeterminado

[a] AUC - Área sob a curva, TNR - Especificidade, TPR - Sensibilidade / *Recall*, ACC - Acurácia, PR - precisão, F1- F1-*score*, KAP - Capa de Cohen.

Na Figura 46 são exibidos os resultados da árvore de decisão gerada pelos algoritmos CTree e RFT, utilizando o método de balanceamento SMOTE, abrangendo todos os casos.

Na Figura 46 (A) é apresentada a árvore de decisão que se baseia na classificação entre o tumor primário e o tecido normal adjacente, considerando todos os estadiamento patológicos (I - IV). Observa-se que somente os miR-196b-5p e miR-99a-5p demonstraram capacidade de discriminar entre os tecidos. As amostras com expressão do miR-196b-5p acima de 7,79 são classificadas como tumor primário. Adicionalmente, na Figura 46 (B), gerada pelo algoritmo RFT, são destacados os 15 miRNAs mais relevantes, com o miR-99a-5p emergindo como o mais crucial na diferenciação dos tecidos, em conjunto com os miR-139-3p e miR-196-5p.

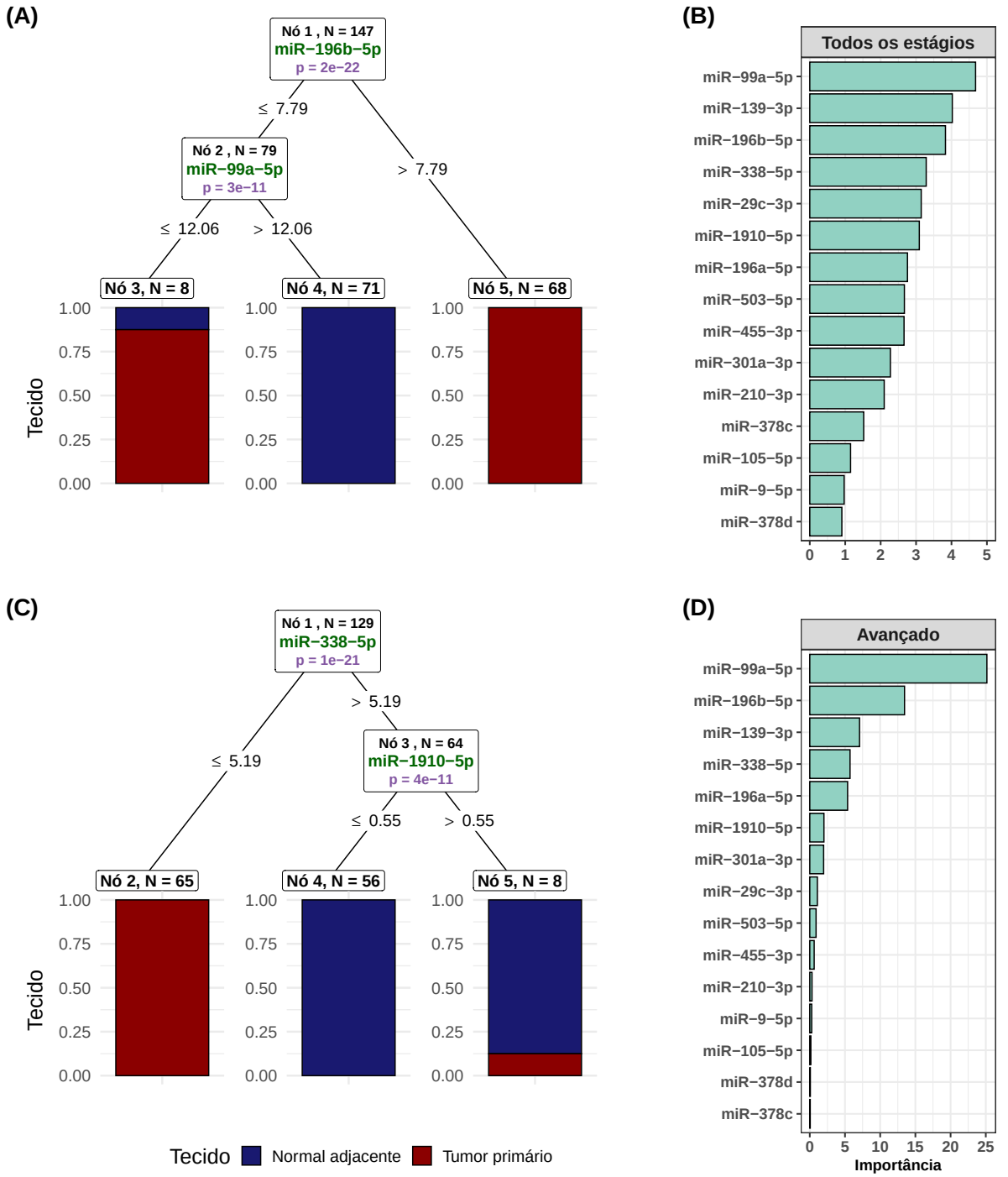


Figura 46 – Classificação dos tecidos utilizando CTree e floresta aleatória, abrangendo todos os estadiamentos patológicos (A, B) e avançados (C, D) do câncer da laringe.

Na fase inicial, a construção da árvore não foi viável devido a questões de convergência do modelo, possivelmente decorrentes da insuficiência de amostras de tecido. Para o estadiamento patológico avançado (III - IV), apenas dois miRNAs importantes, com diferentes pontos de corte de expressão, conseguiram diferenciar os tecidos: miR-338-5p e miR-1910-5p. O miR-338-5p desempenha um papel crucial na primeira separação e variação das amostras, onde uma expressão menor ou igual a 5,19 classifica as amostras como tumor primário (Figura 46 C). No entanto, na

Figura 46 (D), observa-se que o miR-99a-5p é o miRNA mais significativo na separação dos tecidos.

6.5 MiRNAs como biomarcadores de sobrevivência para câncer da cavidade oral e laringe

As estimativas de Kaplan-Meier (KM) da função de sobrevivência para o câncer da cavidade oral e laringe são mostradas na Figura 47. Pode-se observar uma queda acentuada na sobrevivência dos pacientes de até cerca de 800 dias (2 anos) para ambos os tipos de câncer, seguida por uma diminuição menos abrupta e, posteriormente, uma tendência à estabilização a partir de cerca de 4760 dias (13 anos) para o câncer oral e de cerca de 3314 dias (9 anos) para o câncer de laringe. Em termos de probabilidade, um paciente com câncer oral tem aproximadamente 22% de chance de sobreviver 9 anos, enquanto que um paciente com câncer de laringe tem cerca de 11% de chance de sobreviver 13 anos. Ademais, a probabilidade de sobrevivência tende a ser menor para pacientes com diagnóstico de câncer oral em comparação com aqueles diagnosticados com câncer de laringe, porém não há diferença estatisticamente significativa na sobrevivência entre os dois tipos de câncer ($p = 0{,}076$).

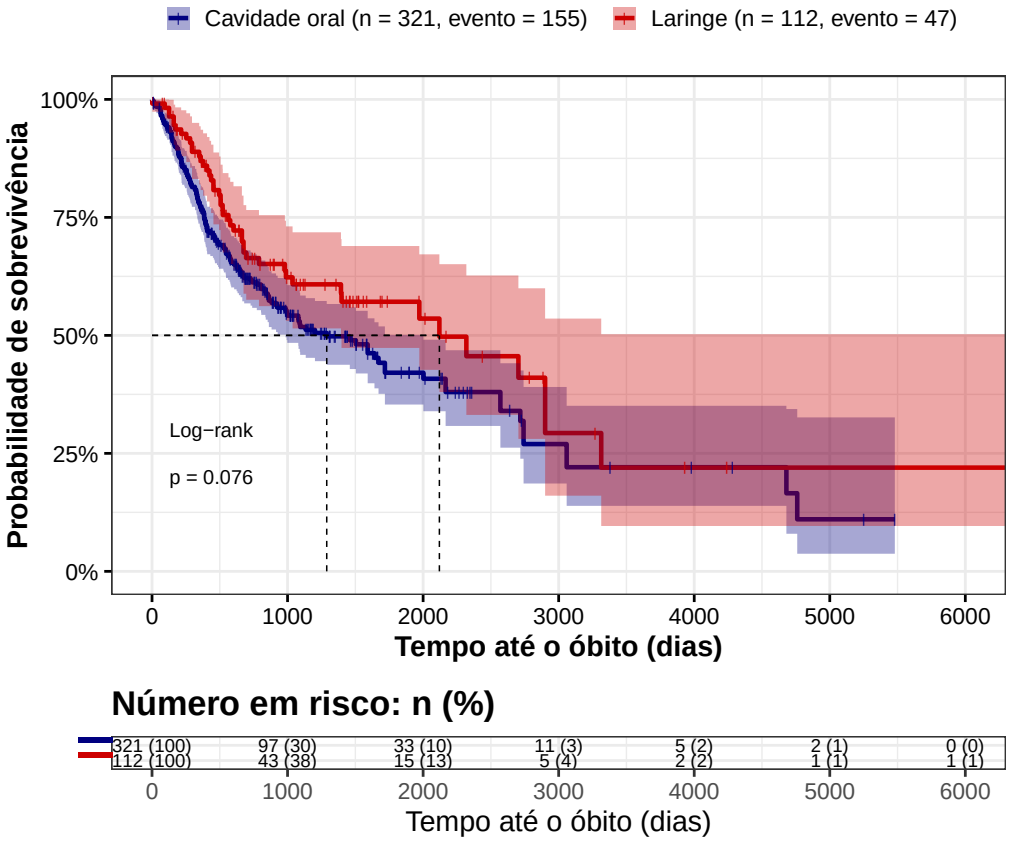


Figura 47 – Curvas de sobrevivência estimada pelo método de Kaplan-Meier de acordo com o câncer da cavidade oral e laringe.

A Tabela 13 apresenta o C-index dos modelos empregados em diferentes estadiamentos do câncer da cavidade oral e laringe nos conjuntos de teste. Em relação ao câncer da cavidade oral, abrangendo todos os estadiamento e o estadiamento avançado, a capacidade preditiva foi superior a 0,70 tanto no conjunto de treinamento quanto no teste do modelo de riscos proporcionais de Cox multivariado, comparado às árvores de sobrevivência e à RSF, o que sugere um desempenho satisfatório na previsão da sobrevivência dos pacientes, uma característica comum em modelos de sobrevivência. Adicionalmente, observou-se que o modelo atendeu ao pressuposto de riscos proporcionais a um nível de significância de 5% em todos os estadiamentos, exceto no estadiamento inicial, onde a convergência não foi alcançada.

Ao comparar a árvore de sobrevivência com o RSF, notou-se que o RSF apresentou um C-index mais elevado em ambos os conjuntos de dados para todos os casos, enquanto a árvore de sobrevivência demonstrou um desempenho intermediário.

Tabela 13 – Comparação do desempenho de modelos multivariados de Cox e árvores de sobrevivência e floresta de sobrevivência aleatória no câncer da cavidade oral e laringe, considerando os estadiamentos da doença.

Modelos	Cavidade oral	Laringe
	C-index	C-index
Todos os estadiamentos (I - IV)		
Modelo de riscos proporcionais de Cox	0,765	0,765
Árvore de sobrevivência	0,445	0,610
Floresta de sobrevivência aleatória	0,522	0,403
Estadiamento patológico inicial (I - II)		
Modelo de riscos proporcionais de Cox	–	–
Árvore de sobrevivência	0,321	–
Floresta de sobrevivência aleatória	0,679	–
Estadiamento patológico avançado (III - IV)		
Modelo de riscos proporcionais de Cox	0,727	–
Árvore de sobrevivência	0,530	0,543
Floresta de sobrevivência aleatória	0,529	0,476

Nota: Não avaliado (–).

Quanto ao câncer da laringe, as amostras não foram divididas aleatoriamente para o modelo de Cox devido à violação do pressuposto de proporcionalidade. No entanto, para a construção das árvores de sobrevivência, a divisão aleatória foi aplicada em todos os estadiamentos e no estadiamento avançado. O modelo de regressão de Cox multivariado ajustado para o câncer de laringe alcançou um índice de concordância de 76,5%, indicando um desempenho satisfatório na previsão da sobrevivência dos pacientes. Contudo, no estadiamento avançado, houve uma violação do pressuposto de proporcionalidade, o que impediu a análise. Além disso, o desempenho da

árvore de sobrevivência foi superior no conjunto de teste, mas inferior no conjunto de treinamento em relação ao RSF.

6.5.1 MiRNAs como preditores de sobrevivência em pacientes com câncer da cavidade oral

As estimativas da análise de regressão de Cox, tanto univariada quanto multivariada, para as características clinicopatológicas dos pacientes diagnosticados com câncer oral estão detalhadas na Tabela 14. Observa-se que os fatores que se mostraram estatisticamente significativos a um nível de 5% e que apresentaram risco de mortalidade em ambos os modelos, univariado e multivariado, considerando todos os estadiamentos, foram a idade (em anos) e o estadiamento (III - IV vs. I - II). No entanto, nos estadiamentos avançados, apenas a idade (em anos) manteve-se significativa. Esses resultados sugerem que a idade e o estadiamento patológico do tumor é importante para determinar o prognóstico de sobrevivência dos pacientes com câncer da cavidade oral.

Tabela 14 – Estimativas da regressão de Cox univariado e multivariado das caraterísticas clinicopatológicas do câncer da cavidade oral no conjunto treinamento.

Fatores	Cox univariado				Cox multivariado			
	$\hat{\beta}$	HR	$IC_{95\%}$(HR)	Valor p	$\hat{\beta}$	HR	$IC_{95\%}$(HR)	Valor p
Todos os estadiamentos (I - IV)								
Sexo (Feminino)	0,05	1,05	(0,72 -1,55)	0,7890	-0,14	0,87	(0,58 - 1,30)	0,4936
Idade (anos)	0,03	1,03	(1,01 - 1,05)	0,0004	0,03	1,03	(1,02 - 1,05)	0,0001
Estadiamento (III - IV)	0,64	1,90	(1,18 - 3,05)	0,0080	0,69	2,00	(1,24 - 3,23)	0,0042
Estadiamento patológico avançado (III - IV)								
Sexo (Feminino)	0,18	1,20	(0,55 - 1,27)	0,3920	0,03	1,03	(0,62 - 1,51)	0,8862
Idade (anos)	0,02	1,02	(1,00 - 1,04)	0,0485	0,02	1,02	(1,00 - 1,04)	0,0673

Nota: HR - *Razão de Risco*; IC - Intervalo de confiança.

Para o estadiamento patológico do câncer da cavidade oral, o teste Log-rank revelou diferença estatisticamente significativa ($p = 0,0072$), conforme ilustrado na Figura 48. De acordo com o modelo univariado de Cox, os pacientes nos estadiamentos avançados apresentam aproximadamente 89,7% maior risco de mortalidade em comparação com aqueles nos estadiamentos iniciais.

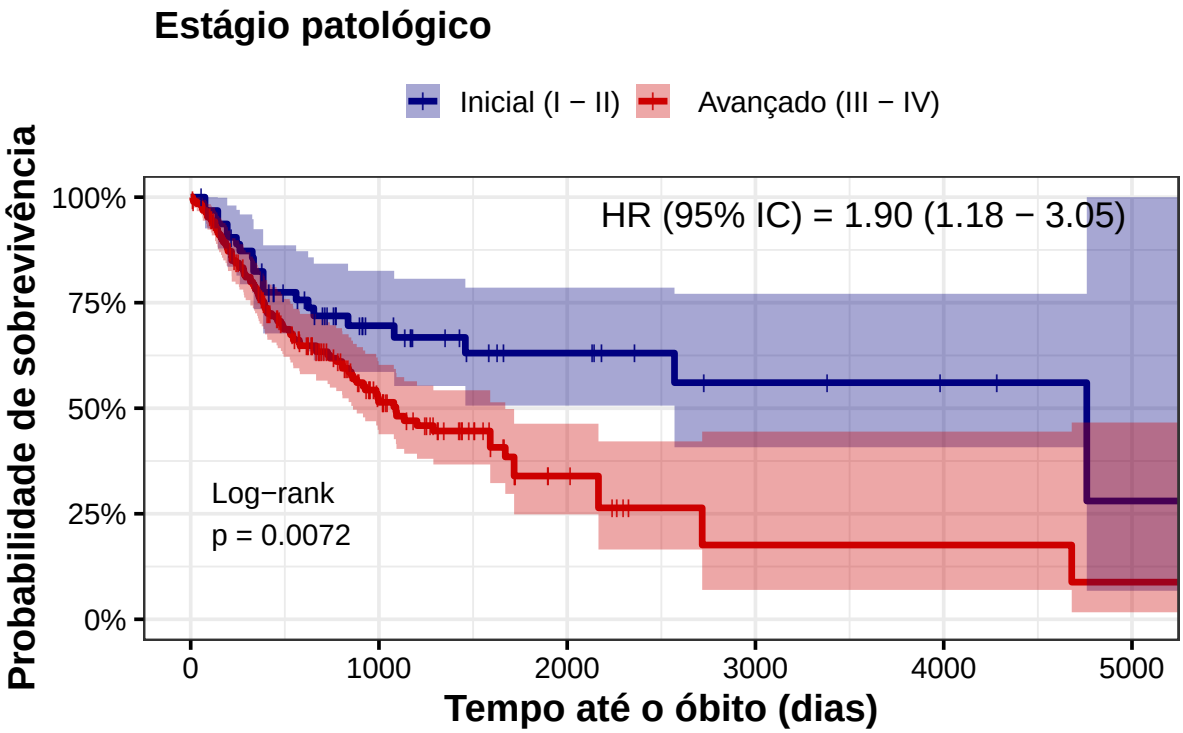


Figura 48 – Curva de sobrevivência estimada pelo método de Kaplan-Meier, teste de Log-Rank e modelo de Cox univariado estratificado por estadiamento patológico no conjunto treinamento do câncer da cavidade oral.

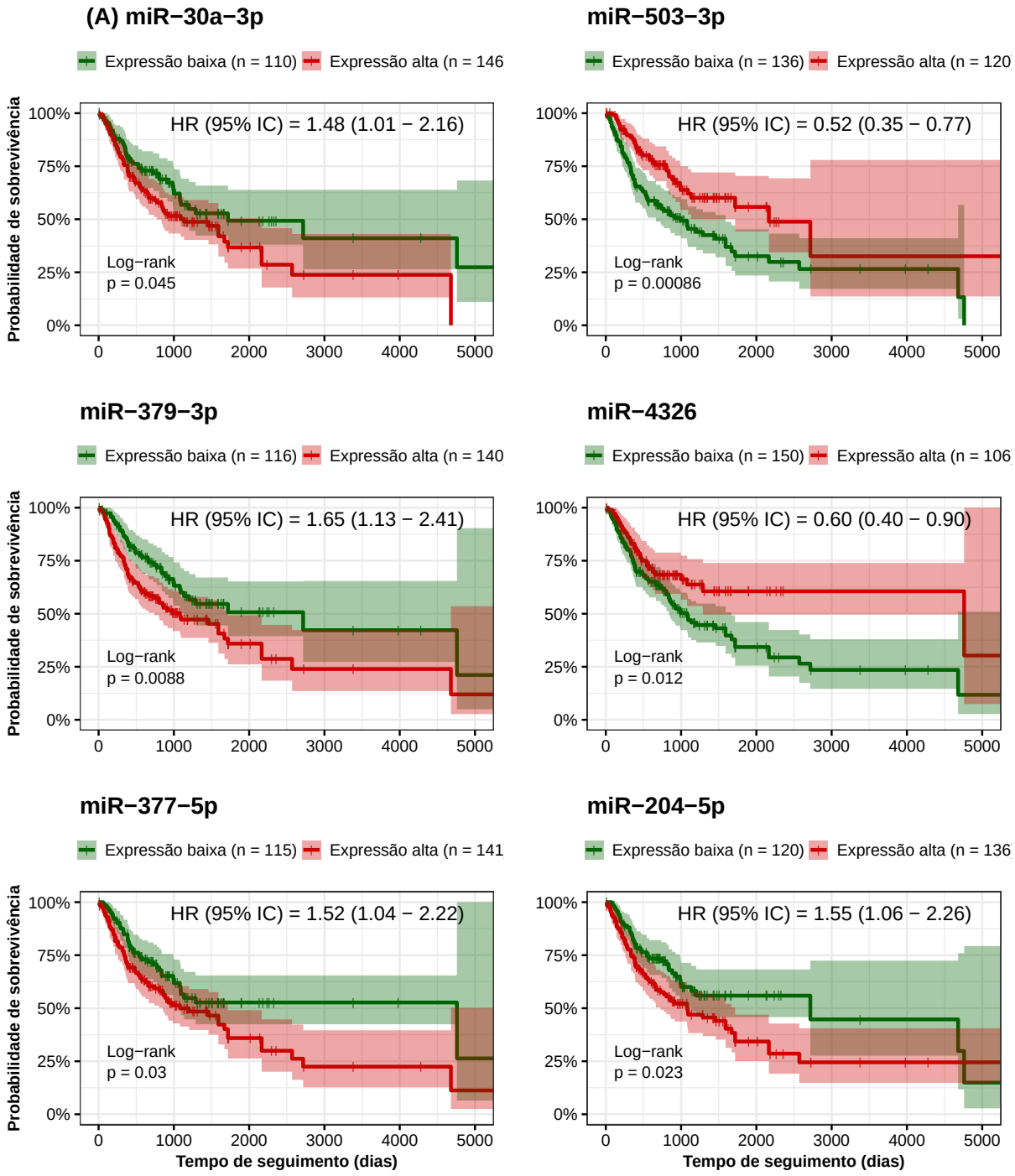
A Figura 49 exibe os resultados da análise de sobrevivência de Kaplan-Meier, do teste de Log-rank e do modelo univariado de riscos proporcionais de Cox. A mediana foi utilizada como ponto de corte para os miRNAs, abrangendo todos os estadiamentos (I - IV), iniciais (I - II) e avançados (III - IV) do câncer de cavidade oral. Valores acima da mediana estão destacados em vermelho (expressão alta), enquanto valores abaixo da mediana estão em verde (expressão diminuída), acompanhados por seus intervalos de confiança de 95%.

Através do teste de Log-rank, foram identificados miRNAs com diferença estatisticamente significativa ($p < 0,05$) em relação à sobrevivência de pacientes com câncer de cavidade oral, considerando diferentes estadiamentos (Figura 49 A). No total, seis miRNAs (miR-30a-3p, miR-4326, miR-377-5p, miR-379-3p, miR-204-5p e miR-503-3p) demonstraram capacidade prognóstica significativa em todos os estadiamentos entre grupos de alta e baixa expressão. Observam-se que os miRNAs com expressão aumentada (miR-4326 e miR-503-3p) apresentaram menor risco de câncer oral e maior taxa de sobrevivência em comparação com os pacientes que tinham expressão diminuída desses miRNAs. Estes últimos estavam significativamente associados a uma pior sobrevivência dos pacientes.

Ao focalizar especificamente nos estadiamentos iniciais, sete miRNAs (miR-381-3p, miR-503-5p, miR-1305, miR-195-5p, miR-204-5p, miR-379-3p e miR-105-5p) apresentaram diferença estatisticamente significativa. Pelo modelo de Cox univariado, a expressão diminuída

dos miR-381-3p, miR-195-5p, miR-204-5p e miR-379-3p, bem como a expressão aumentada dos miR-503-5p, miR-1305 e miR-105-5p, estavam significativamente associadas a uma pior sobrevivência dos pacientes com câncer de cavidade oral (Figura 49 B).

Entretanto, nos estadiamentos avançados, apenas o miR-503-5p apresentou significância estatística pelo teste de Log-rank, e sua expressão aumentada demonstrou um prognóstico favorável em termos da taxa de sobrevivência (Figura 49 C).



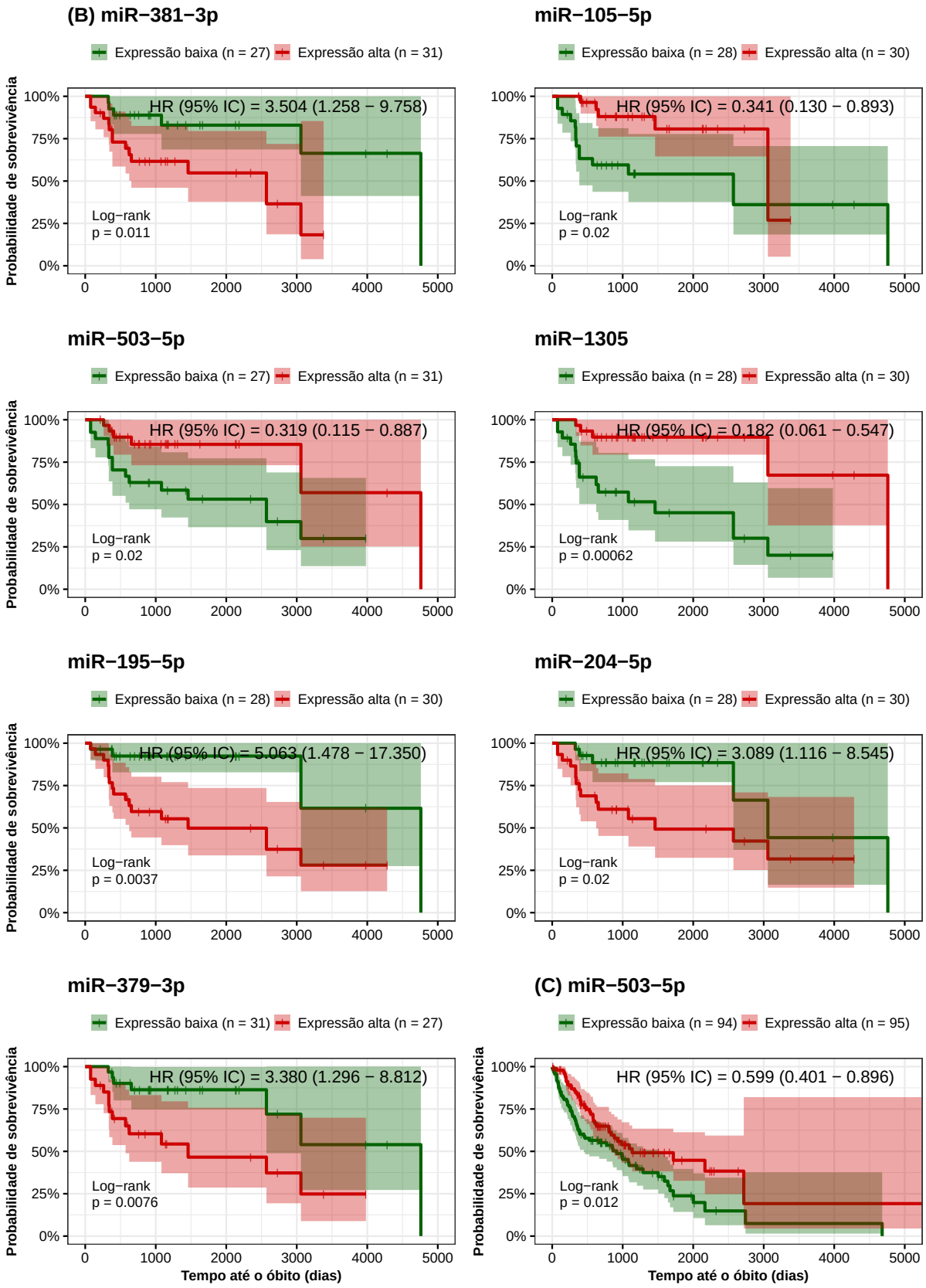


Figura 49 – Curvas de sobrevivência estimada pelo método de Kaplan-Meier, teste de Log-Rank e modelo de Cox univariado de acordo com a expressão de miRNAs em diferentes estadiamentos patológicos (A) inicial (B) e avançado (C) do câncer da cavidade oral no conjunto de treinamento.

Na Figura 50 estão apresentadas as estimativas dos parâmetros do modelo multivariado de riscos proporcionais de Cox considerando todos os estadiamentos (I - IV) e avançados (III - IV) do câncer da cavidade oral.

Na análise do modelo de Cox multivariado para todos os estadiamentos, identificaram-se 24 potenciais miRNAs prognósticos associados à sobrevivência de pacientes com câncer oral. Entre esses, as altas expressões de miR-30-3p ($HR = 2,33$), miR-210-3p ($HR = 4,94$), miR-376c-3p ($HR = 2,02$), miR-224-3p ($HR = 2,11$), miR-877-5p ($HR = 1,63$), miR-937-3p ($HR = 1,59$) e miR-204-5p ($HR = 3,08$) foram associados a um aumento no risco de morte e menor taxa de sobrevivência. Pelo contrario, a baixa expressão dos miR-455-3p, miR-30a-5p, miR-99a-5p, miR-615-3p, miR-378a-3p, miR-503-3p, miR-767-5p, miR-7705, miR-4326 foram associados a um bom prognóstico para pacientes com câncer oral. Os miRNAs (miR-136-3p, miR-503-5p, miR-378d, miR-379-5p, miR-4652-5p, miR-365a-5p, miR-381-5p) não apresentaram associação significativa com a taxa de risco da doença e, portanto, não influenciaram na ocorrência do evento de interesse (Figura 50 A).

Em relação ao estadiamento avançado do câncer da cavidade oral, foram observados 24 potenciais miRNAs prognósticos identificados pelo modelo de Cox multivariado. Dentre esses, miR-196b-5p, miR-503-5p, miR-4652-5p, miR-204-5p, miR-133b e miR-375 foram associados a um prognóstico desfavorável. Notavelmente, miR-133b destacou-se como um fator de risco significativo, com um risco 5,41 vezes maior de não sobreviver ao câncer da cavidade oral, enquanto miR-133a-3p evidenciou-se como o principal fator de proteção no câncer da cavidade oral no estadiamento avançado, enquanto os outros não alcançaram significância estatística (Figura 50 B).

Por outro lado, foram conduzidas análises de árvores de sobrevivência abrangendo todos os estadiamento do câncer da cavidade oral (I a IV), divididos em estadiamentos iniciais (I e II) e avançados (III e IV). Essas análises foram realizadas por meio dos algoritmos CART©, com a aplicação de poda utilizando o custo de complexidade (cp), além do algoritmo RSF, conforme representado na Figura 51.

Na Figura 51 (A), a árvore de sobrevivência baseada no método CART©, com poda ($cp = 0,035$), revelou o miR-503-3p (nó 1, $n = 256$) como um preditor significativo da sobrevivência em pacientes. Para valores de expressão iguais ou inferiores a 2,4 associados ao nó 3, observou-se um prognóstico desfavorável ($HR = 1,4$), resultando em 76 óbitos entre 130 casos. Além disso, o teste de Log-rank evidenciou diferenças significativas nas curvas de Kaplan-Meier para os nós 2 e 3 ($p < 0,0001$) e pelo modelo de Cox univariado, o nó 3 destacou-se com um prognóstico pior em termos de taxa de sobrevivência ($HR = 2,297$).

Contudo, na Figura 51 (B), são apresentados os miRNAs com importância superior a zero, identificados pelo método RSF. Notavelmente, o miR-503-3p destaca-se como um componente importante na predição da sobrevivência.

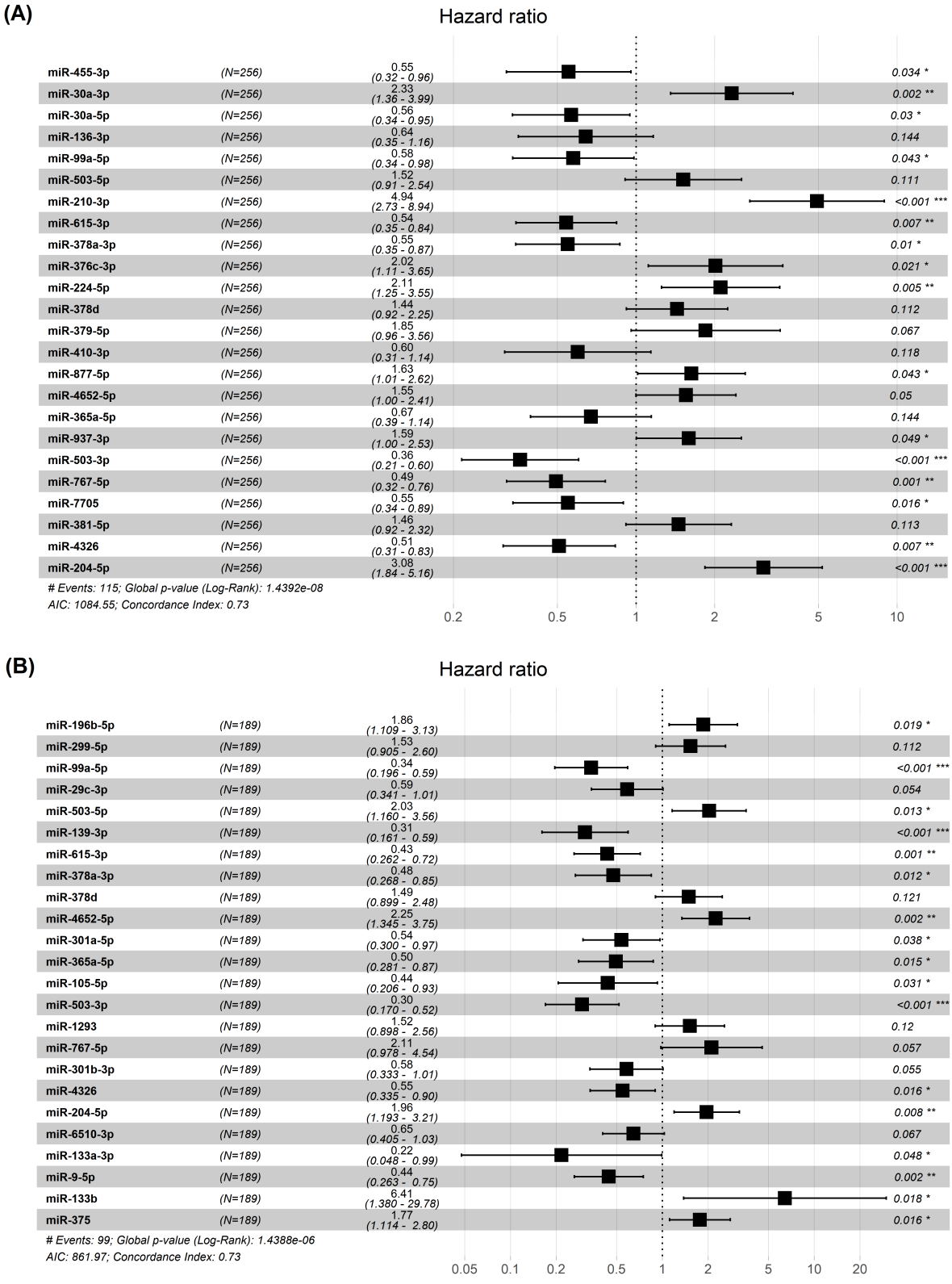
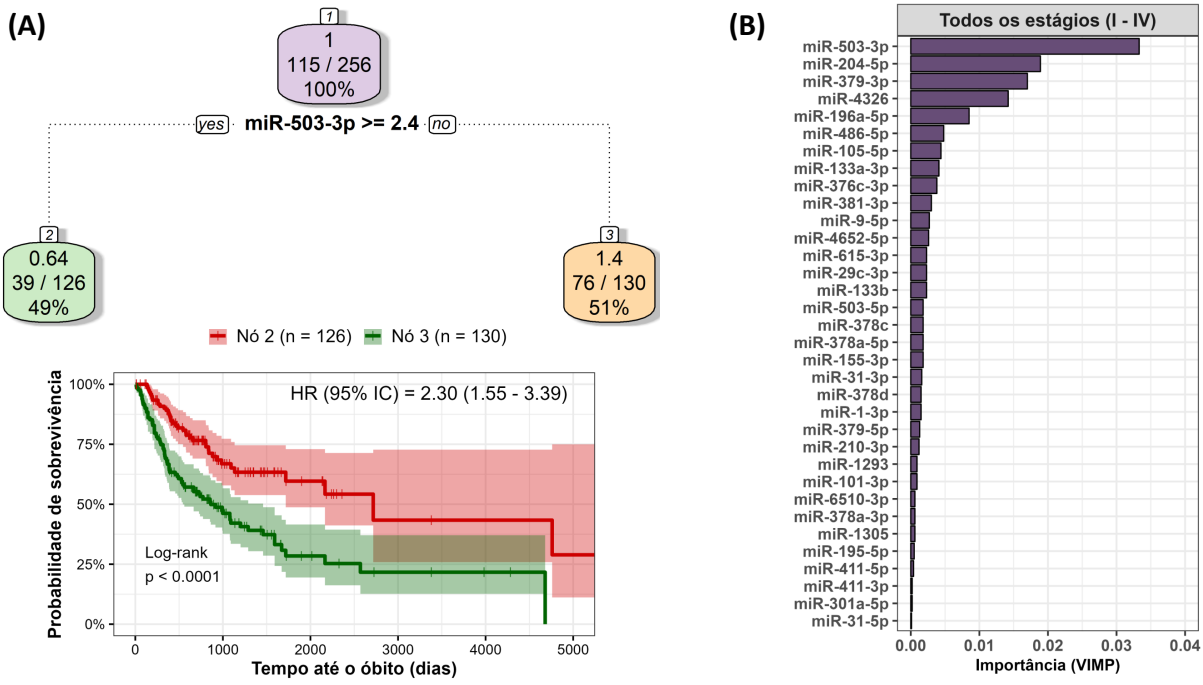


Figura 50 – Gráfico dos riscos do modelo multivariado de riscos proporcionais de Cox dos miRNAs associados com a sobrevivência dos pacientes com câncer oral considerando todos os estadiamentos (I - IV) e avançado (III - IV) no conjunto treinamento. Cada miRNA encontra-se categorizada (1: expressão aumentada e 0: expressão diminuída).

Na Figura 51 (C), observa-se que foram identificados três miRNAs (miR-1305, miR-105-5p e miR-376c-3p) como os principais preditores relevantes que afetam a sobrevivência de pacientes com câncer oral em estadiamento inicial, segundo o algoritmo CART© com poda ($cp = 0,121$). Quando o valor de expressão de miR-1305 é maior ou igual a 2,532, o Nó 2 ($n = 13$ pacientes) é alcançado, com um HR de 0,2, indicando que a expressão elevada do miR-1305 está associada a uma menor taxa de risco de morte. Por outro lado, se a expressão de miR-1305 for inferior a 2,532, miR-105-5p e miR-376c-3p se tornam os principais preditores da sobrevivência. Além disso, o nó 4 sugere que pacientes com uma expressão maior ou igual a 1,567 de miR-105-5p têm menor probabilidade de morte do que aqueles com expressão menor. Por sua vez, o nó 5 indica que pacientes com expressão de miR-376c-3p menor que 5,021 têm maior probabilidade de sobrevivência do que aqueles com expressão igual ou superior. No entanto, na Figura 51 (D), são apresentados os 20 miRNAs mais importantes ($VIMP > 0$) obtidos pelo RSF, sendo que miR-1305, miR-376c-3p e miR-224-5p foram os mais significativos na predição da sobrevivência.

Na Figura 51 (E), cinco miRNAs (miR-503-3p, miR-139-3p, miR-615-3p, miR-378a-3p e miR-195-5p) foram identificados como preditores relevantes na sobrevivência dos pacientes no estadiamento avançado do câncer oral pelo algoritmo CART© com poda ($cp = 0,0318$). Dessa forma, pode-se observar que quando o valor de expressão do miR-503-3p foi menor que 2,537, os pacientes apresentaram uma taxa de risco maior de mortalidade ($HR = 1,30$), mas se a expressão do miR-503-3p fosse maior ou igual 2,537 e se a expressão do miR-139-3p fosse maior a 5,982 apresentaram um bom prognóstico de sobrevivência.

Similarmente, a Figura 51 (F) destaca que o miR-503-3p é o mais importante na predição de sobrevivência no estadiamento avançado do câncer da cavidade oral obtido pelo RSF.



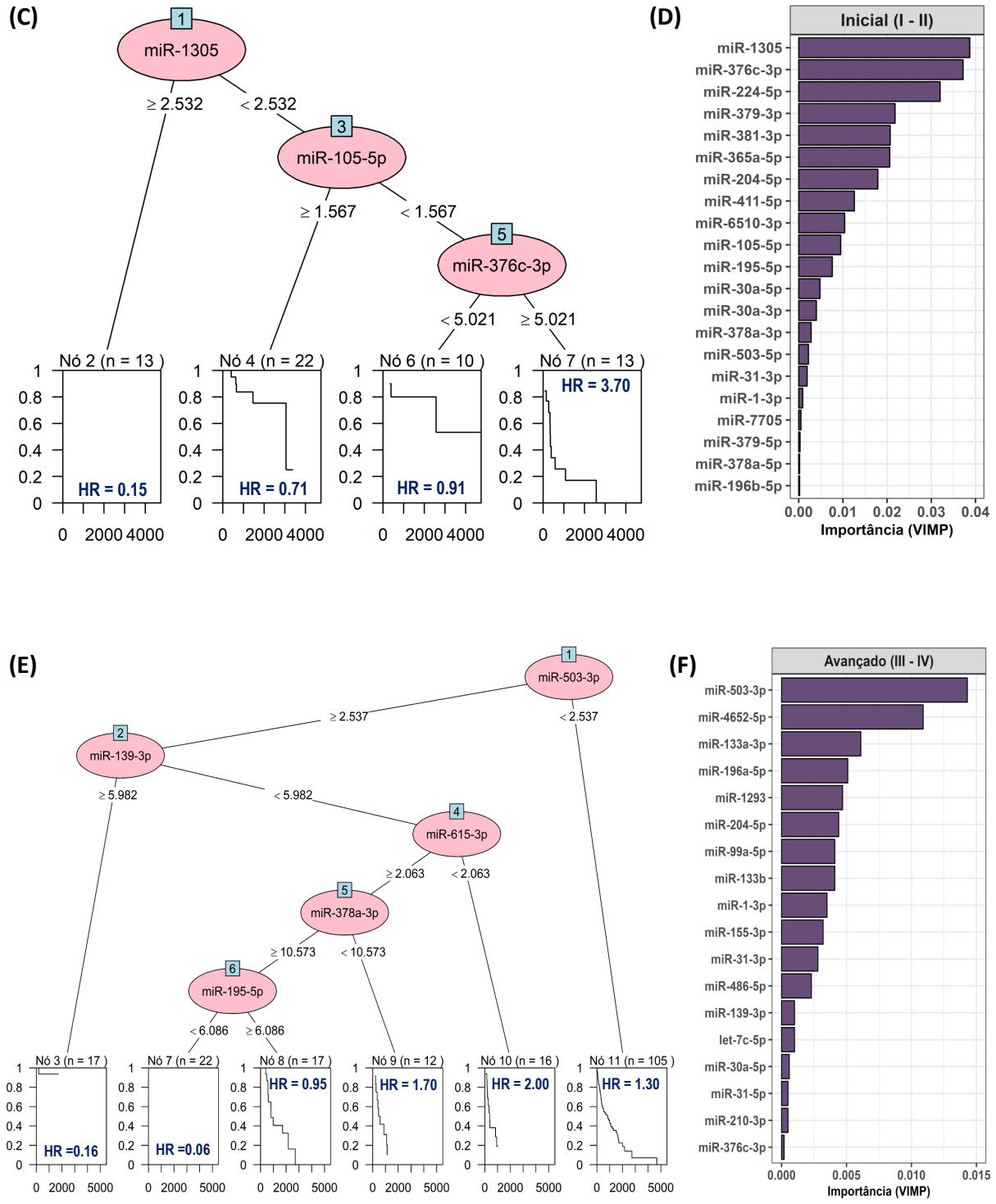


Figura 51 – Árvore de sobrevivência podada obtida pelo CART© e gráfico de barras da importância de miRNAs obtidos pela floresta aleatória na previsão da sobrevivência dos pacientes considerando todos os estadiamentos (A, B), inicial (C, D) e avançado (E, F) do câncer da cavidade oral no conjunto treinamento.

6.5.2 MiRNAs como preditores de sobrevivência em pacientes com câncer da laringe

As estimativas da análise de regressão de Cox, abrangendo tanto a análise univariada quanto a multivariada, para as características clinicopatológicas dos pacientes diagnosticados com câncer de laringe, estão apresentadas na Tabela 15. Nota-se que, apenas o sexo demonstrou significância estatística, com um $HR = 2,74$, indicando que as mulheres enfrentam um risco de mortalidade significativamente superior em comparação aos homens. Pelo teste Log-rank, evidenciou-se uma diferença estatisticamente significativa ($p = 0,0043$), conforme representado de maneira ilustrativa na Figura 52.

Tabela 15 – Estimativas de regressão de Cox univariada e multivariada das caraterísticas clinicopatológicas do câncer da laringe.

Fatores	Cox univariado				Cox multivariado			
	$\hat{\beta}$	HR	$IC_{95\%}$(HR)	Valor p	$\hat{\beta}$	HR	$IC_{95\%}$(HR)	Valor p
Todos os estadiamentos								
Sexo (Feminino)	1,01	2,74	(1,33 - 5,65)	0,0059	0,98	2,65	(0,77 - 1,72)	0,0208
Idade (anos)	0,02	1,02	(0,98 - 1,06)	0,3160	0,02	1,02	(0,98 - 1,05)	0,3604
Estadiamento (III - IV)	0,52	1,68	(0,80 - 3,57)	0,1730	0,06	1,06	(0,44 - 2,55)	0,8931

Nota: HR - Razão de Risco; IC - Intervalo de confiança.

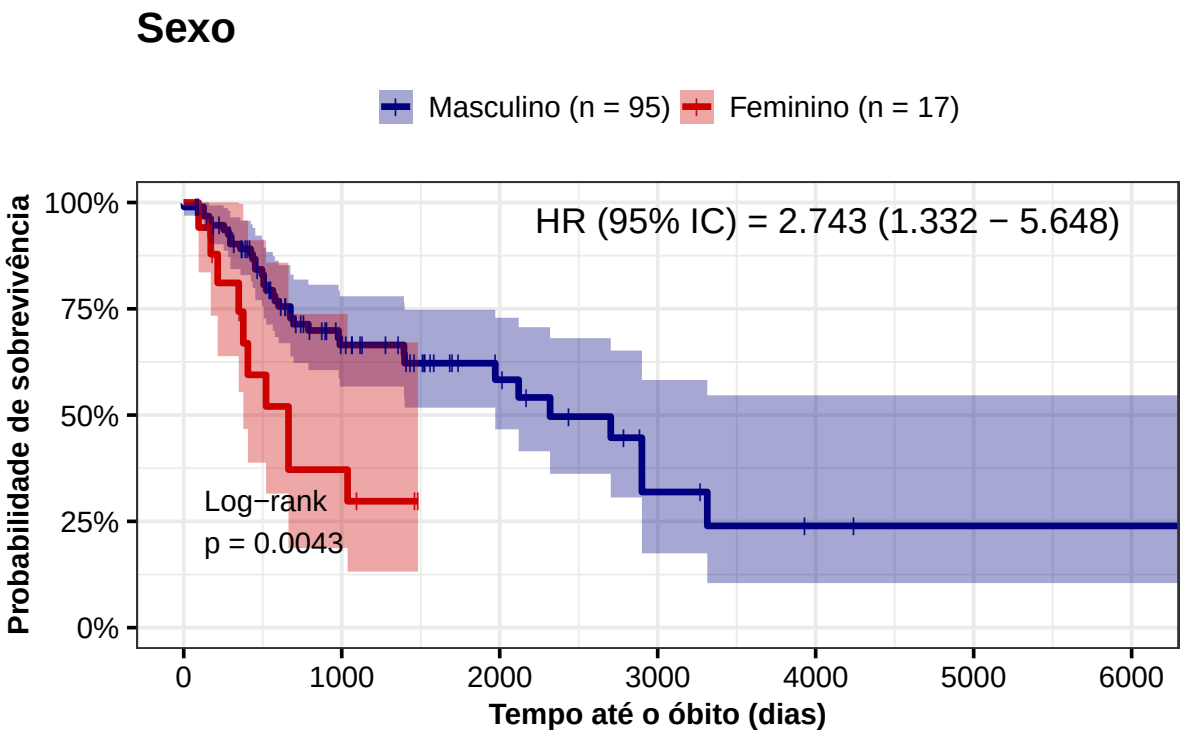


Figura 52 – Curvas de sobrevivências estimada pelo método de Kaplan-Meier para sobrevivência dos pacientes estratificado por estadiamento patológico no câncer da laringe.

Os resultados da análise de sobrevivência dos miRNAs, realizada por meio do teste de Log-Rank e do modelo de Cox univariado, estão representados na Figura 53. Nesse contexto, foram identificados nove miRNAs (miR-139-3p, miR-13p, miR-455-3p, miR-4326, miR-301a-3p, miR-6399-3p, miR-210-3p, miR-206, e miR-301-5p) que exibiram diferenças estatisticamente significativas pelo teste de Log-rank ($p < 0,05$), relacionando-se à sobrevivência de pacientes com câncer de laringe em diferentes estadiamentos da doença. Destes, apenas dois miRNAs (miR-455-3p e miR-1-3p) demonstraram significância no estadiamento avançado do câncer de laringe (Figura 53 A).

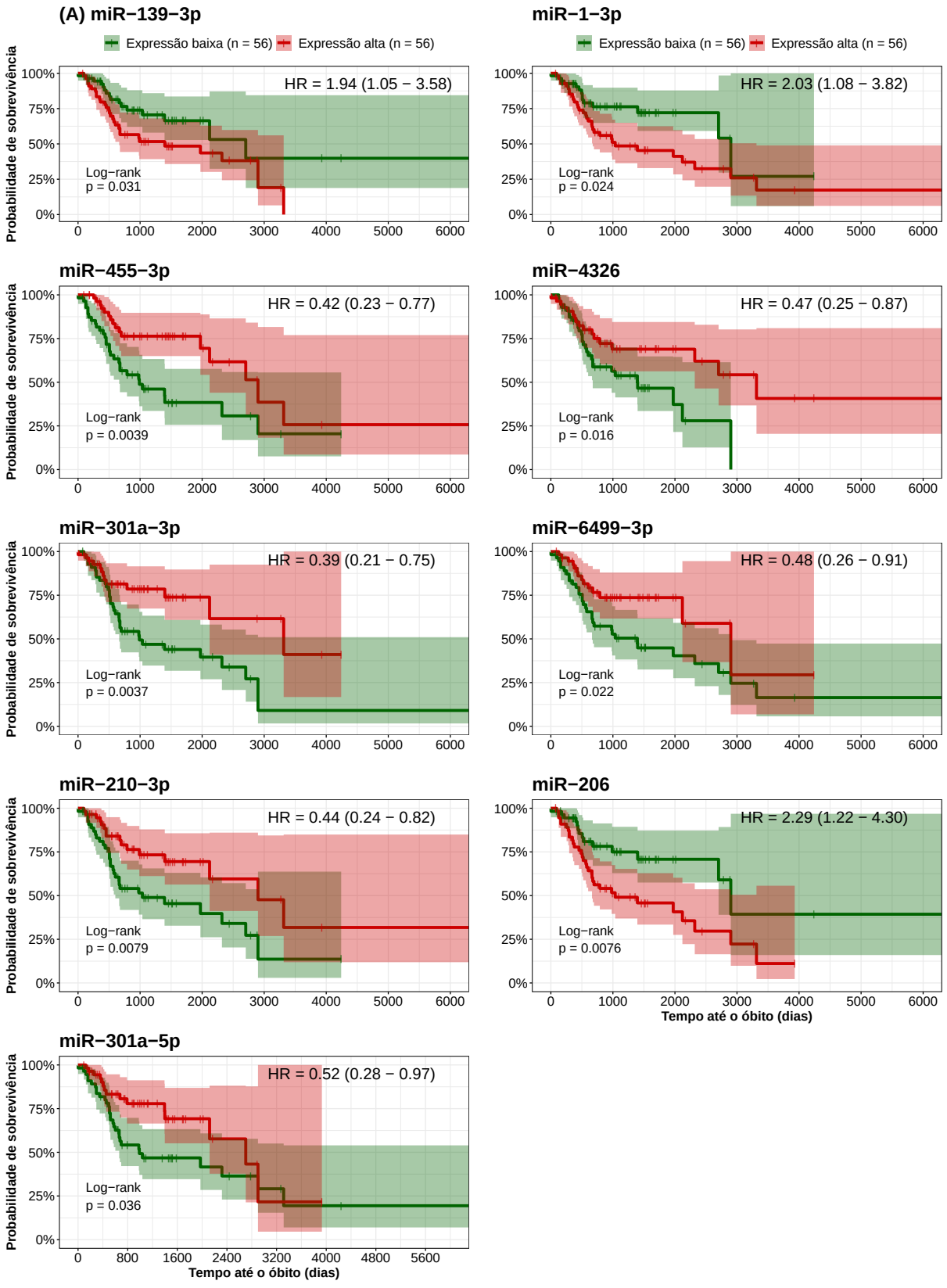
Adicionalmente, observam-se que a expressão aumentada dos miR-455-3p, miR-4326, miR-301a-3p, miR-6499-3p, miR-210-3p e miR-301a-5p está associada de maneira significativa a um prognóstico mais favorável em pacientes com câncer de laringe, abrangendo todos os estadiamentos da doença. Entretanto, no estadiamento avançado do câncer de laringe, observou-se que a elevada expressão do miR-455-3p está correlacionada a um prognóstico mais positivo, enquanto o miR-1-3p está associado a um prognóstico menos favorável, conforme evidenciado na Figura 53 (B).

Entretanto, a Figura 54 apresenta o modelo de regressão de Cox multivariado ajustado para o câncer de laringe, abrangendo todos os estadiamentos da doença. Em particular, destaca-se o miR-1-3p ($HR = 10,154$; $IC_{95\%} : 2,558 - 40,310$) como o principal fator de risco associado a um prognóstico desfavorável. Por outro lado, o miR-133b ($HR = 0,043$; $IC_{95\%} : 0,010 - 0,190$) emerge como o principal fator de proteção relacionado à sobrevivência de pacientes com câncer de laringe, enquanto os miR-378c, miR-378d, miR-6510-3p, miR-3170, miR-508-3p e miR-184 não foram significativas.

Na Figura 55 (A), é possível observar a árvore podada ($cp = 0,083$) de sobrevivência gerada pelo CART©, que identificou três miRNAs (miR-301a-3p, miR-378c e miR-4326) possivelmente relacionados à sobrevivência de pacientes com câncer de laringe em todos os estadiamento. O nó 7, com um total de 40 pacientes, indica que quando a expressão do miR-301a-3p é inferior a 4,507, a taxa de risco de mortalidade é de 1,9. Em contraste, o nó 4, contendo 38 pacientes, revela que quando a expressão do miR-301a-3p é igual ou superior a 4,507, e ao mesmo tempo maior ou igual a 3,699 para miR-378c, e maior e igual a 2,872 para miR-4326, a taxa de risco é de 0,19, indicando um bom prognóstico de sobrevivência. No entanto, na Figura 55 (B), o RSF identificou 24 miRNAs significativos que impactam positivamente na predição da sobrevivência dos pacientes. Dentre eles, os miR-301-3p, miR-455-3p e miR-301a-5p destacam-se como os mais relevantes para a predição da sobrevivência em todos os estadiamentos do câncer de laringe.

Em relação ao estadiamento avançado do câncer da laringe, na Figura 55 (C) observa-se que apenas dois miRNAs (miR-301a-3p e miR-4449) que foram identificados pelo CART© com poda ($cp = 0,066$) como possíveis preditores de sobrevivência. Dessa forma, quando o valor de expressão do miR-301a-3p foi maior ou igual a 4,507 o nó 2 foi alcançado com uma taxa de risco

de 0,55 o que níveis elevados desse miRNA pode estar relacionados a uma melhor prognóstico de sobrevivência. Entretanto, níveis baixos de miR-301a-3p e altos níveis de miR-4449 podem estar associados a uma sobrevivência pior.



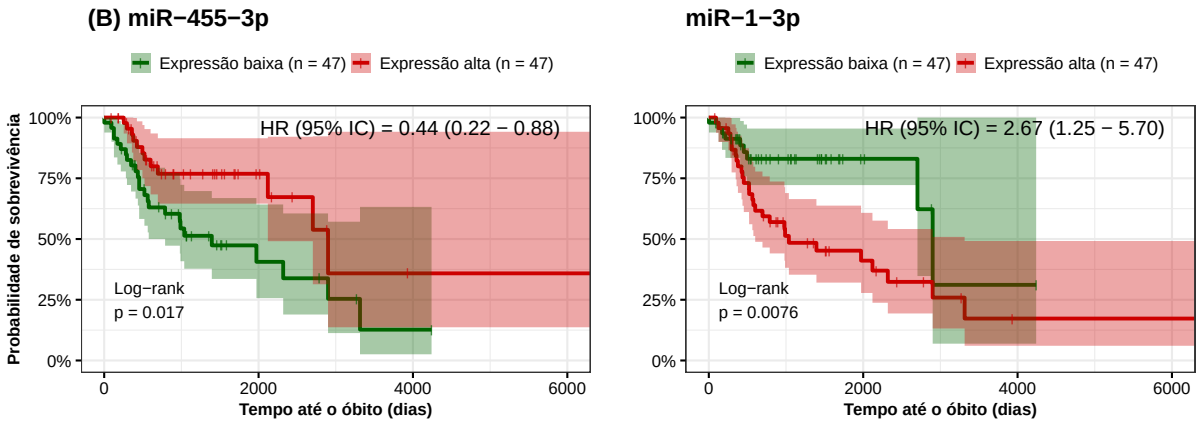


Figura 53 – Curvas de sobrevivência estimada pelo método de Kaplan-Meier, teste de Log-Rank e modelo de Cox univariado para sobrevivência dos pacientes de acordo com a expressão dos miRNAs considerando todos os estadiamentos (A) e o estadiamento avançado (B) do câncer da laringe.

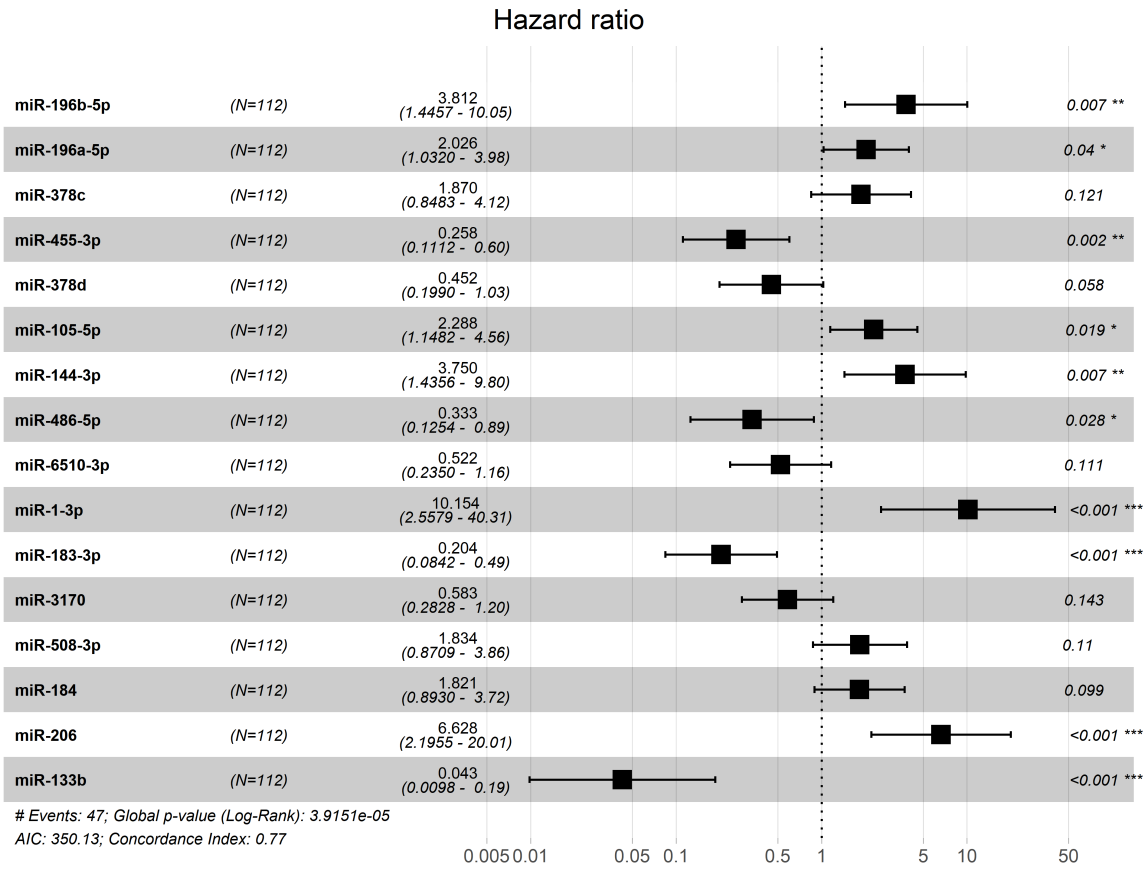


Figura 54 – Gráfico de riscos do modelo multivariado de riscos proporcionais de Cox dos miRNAs associados com a sobrevivência dos pacientes considerando todos os estadiamentos do câncer da laringe. Cada miRNA encontra-se categorizada (1: expressão aumentada e 0: expressão diminuída).

Contudo, na Figura 55 (D), o RSF identificou 12 miRNAs importantes que influenciam

positivamente na previsão da sobrevivência dos pacientes. Entre eles, os miR-301a-3p e miR-301a-5p destacam-se como os mais relevantes para a predição da sobrevivência nos estadiamentos avançados do câncer de laringe.

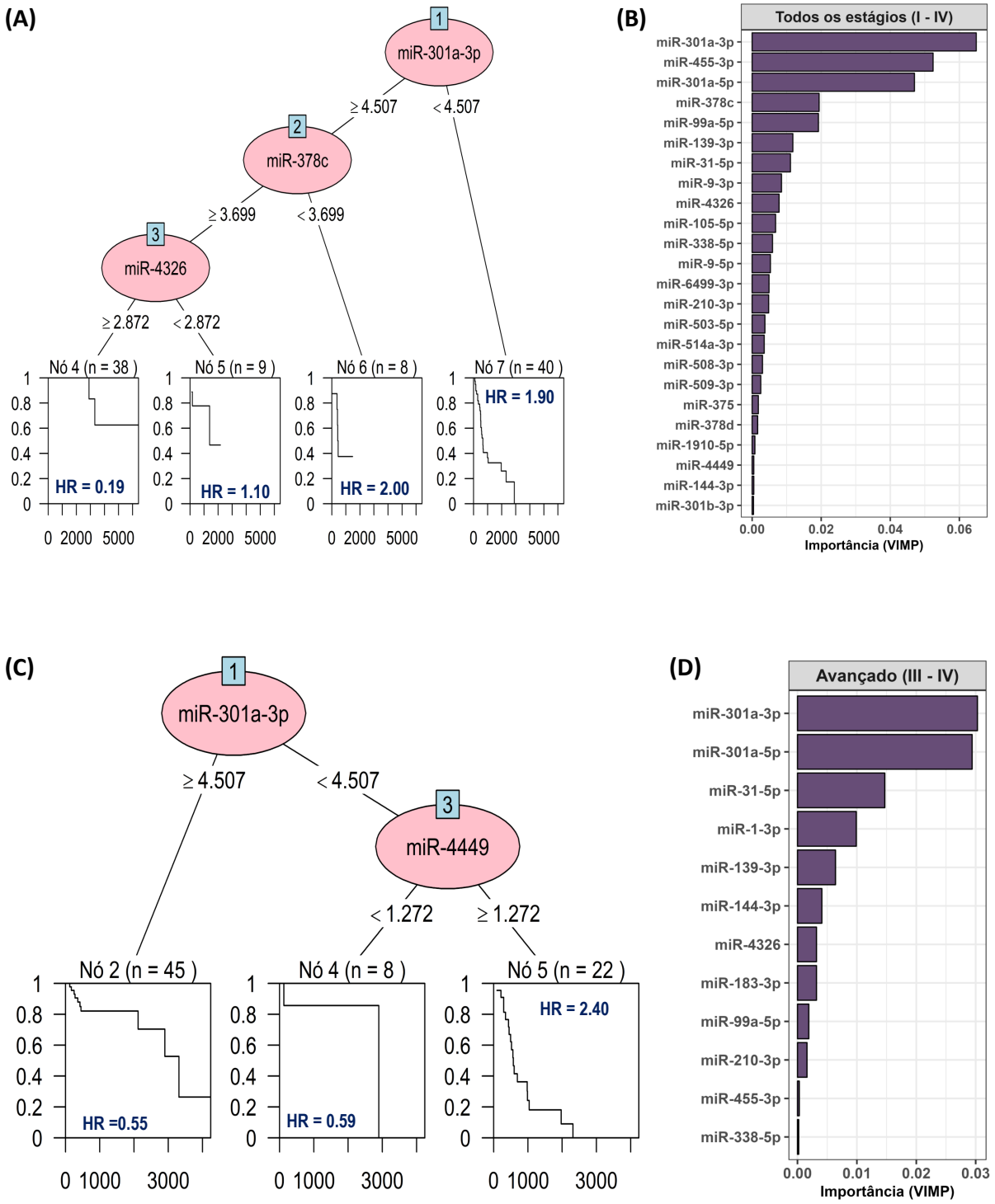


Figura 55 – Árvore de sobrevivência podada obtido pelo CART© e fator de importância de miRNAs obtidas via floresta aleatória na previsão da sobrevivência dos pacientes considerando todos os estadiamentos (A, B) e avançado (C, D) no conjunto treinamento do câncer da laringe, respectivamente.

7 Discussão

Este estudo teve como propósito a identificação de miRNAs e genes alvo que possam desempenhar papéis significativos como potenciais biomarcadores nos âmbitos do diagnóstico, prognóstico e tratamento dos cânceres de cabeça e pescoço, com ênfase nos casos relacionados à cavidade oral e à laringe. Para atingir esse objetivo, foram utilizadas diversas ferramentas de bioinformática, analisando as expressões de miRNA-Seq e mRNA-Seq, além de informações clinicopatológicas dos pacientes disponíveis no TCGA.

A maioria dos estudos relacionados aos miRNAs e mRNAs concentra-se na análise de expressão diferencial, comparando tumores primários com tecidos normais usando diferentes métodos de expressão diferencial. Na etapa inicial deste estudo, empregaram-se os métodos DESeq2, edgeR e o teste de Wilcoxon-Mann-Whitney para avaliar a eficácia na identificação de miRNAs desregulados.

Em relação ao câncer de cavidade oral, identificaram-se 58 miRNAs com DESEq2, 58 com edgeR e 60 com o teste de Wilcoxon-Mann-Whitney foram expressos diferencialmente entre tumores e tecidos normais adjacentes, a maioria com expressão diminuída. No câncer de laringe, DESEq2 identificou 44 miRNAs, edgeR identificou 42, e o teste de Wilcoxon-Mann-Whitney identificou 33 foram expressos diferencialmente entre tumores e tecidos normais adjacentes, predominantemente com expressão aumentada. Embora o teste de Wilcoxon-Mann-Whitney tenha identificado todos os DEmiRNAs detectados pelo DESeq2, a discrepância na quantidade sugere que a escolha do método pode influenciar significativamente os resultados. Li *et al.* (2022a) destacaram a eficácia do teste de Wilcoxon-Mann-Whitney com normalização edgeR, especialmente em grandes amostras, enquanto Fekete, Welker e Győrffy (2020) empregaram o teste com normalização e FDR-BH, evidenciando que os resultados desses métodos podem ser complementares. Os DEmiRNAs identificados foram validados usando ferramentas como dbDMEC e oncomiR, corroborando seu envolvimento em neoplasias.

Estudos específicos sobre CCP confirmaram a presença dos miRNAs identificados neste estudo ao serem comparadas o tumor primário com os tecidos normais independentemente do método utilizado. Por exemplo, Lopez *et al.* (2018) observaram superexpressão de miR-615-3p, miR-1269a, miR-455-3p, miR-196b-5p, miR-210-3p, miR-1293 e miR-503-5p, enquanto miR-101-3p, miR-29c-3p, miR-375, miR-376c-3p, miR-378c, miR-195-5p, miR-299-5p e miR-6510-3p estavam subexpressos. Em outro estudo, Panvongsa *et al.* (2021) identificaram a superexpressão

do miR-1910-5p, enquanto Qi *et al.* (2021) demonstraram que miR-411-5p e miR-410-3p estavam subexpressos. Além disso, a pesquisa conduzida por Oshima *et al.* (2021) revela que os miR-31-5p e miR-31-3p desempenham papéis como miRNAs oncogênicos em células CCP.

Em estudos relacionados ao câncer da cavidade oral, diversas pesquisas identificaram alterações nos níveis de diferentes miRNAs. Petronacci *et al.* (2019) observaram uma redução nos níveis de miR-6510-3p, miR-204-5p, miR-101-3p e miR-99a-5p. Taherkhani *et al.* (2022) reportaram uma superexpressão de miR-503-5p, miR-31-3p e miR-196a-5p em estadiamento iniciais, enquanto miR-411-5p, miR-30a-3p, miR-410-3p, miR-30a-5p e let-7c-5p estavam subexpressos. Yoon *et al.* (2014) evidenciaram que miR-375 estava subexpresso, Azar *et al.* (2020) observaram uma diminuição significativa em miR-133b, e Wang *et al.* (2018) destacaram o papel de miR-1-3p como supressor tumoral. Ukey *et al.* (2023) identificaram aumento na expressão de miR-9-5p e redução na expressão de miR-133a-3p, enquanto Gołąbek *et al.* (2023) relataram regulação positiva de miR-155-3p. Além disso, Piotrowski *et al.* (2021) e Taherkhani *et al.* (2022) encontraram evidências de uma superexpressão dos miR-34b-5p e miR-34c-5p. Em estudos anteriores, Chen, Zhang e Zhang (2019) observaram uma redução significativa do miR-1254 em tecidos e linhas celulares de carcinoma espinocelular oral, associando essa diminuição ao estadiamento do câncer e à metástase linfonodal. Adicionalmente, Gai *et al.* (2018) constataram que miR-512-3p estava regulado positivamente em pacientes com câncer oral, sugerindo-o como um novo biomarcador candidato para essa doença.

No caso do câncer de laringe, Luo *et al.* (2022) identificaram superexpressão dos miR-615-3p, miR-4652-5p, miR-503-5p, miR-455-3p, miR-196a-5p, miR-196b-5p e miR-301a-5p. Piotrowski *et al.* (2021) destacaram a superexpressão do miR-6510-3p, enquanto Lu *et al.* (2015) consideraram miR-301a-3p um oncogene. Além dessas descobertas, Fei *et al.* (2017) relataram uma diminuição na expressão dos miR-133b, miR-206, miR-375 e miR-378c. Zhao e Cui (2020) identificaram superexpressão dos miR-767-5p e miR-105-5p, enquanto Fang *et al.* (2019), destacaram a regulação negativa do miR-29c-3p, associando-o a diferentes estadiamento de progressão tumoral e prognóstico desfavorável em pacientes com carcinoma espinocelular de laringe (LSCC). Adicionalmente, Broseghini *et al.* (2023) observaram expressão aumentada de miR-196a-5p, miR-210-3p, miR-301a-3p, miR-503-3p e miR-9-5p, enquanto a expressão diminuída foi constatada nos miR-133a-3p, miR-133b, miR-144-3p, miR-206 e miR-29c-3p.

A análise de expressão diferencial de mRNAs e a predição de alvos, por meio de ferramentas de bioinformática específicas para cada tipo de câncer da cavidade oral e laringe, identificaram um total de 127 e 85 genes alvo candidatos, respectivamente. Dentre esses, foram selecionados aqueles cuja expressão apresentava correlação inversa com os DEmiRNAs, resultando em 103 genes alvo candidatos associados a 55 miRNAs e 63 genes alvo candidatos associados a 40 miRNAs. Notavelmente, a subexpressão do miR-29c-3p destacou-se como um regulador crucial de genes específicos em ambos os tipos de câncer, incluindo COL5A3, ITGA6, LAMC2, COL5A2, DNMT3B, SERPINH1, COL3A1, COL1A2, COL1A1, C1QTNF6,

COL7A1, COL10A, COL4A1, LAMC2, COL1A2, COL7A1, SLC2A14, COL5A2, SPARC, CD276, COL3A1, DNMT3B, C1QTNF6, PXDN, COL4A1 e COL1A1. O estudo de Wei *et al.* (2023a) confirmou que o miR-29c-3p tem como alvo específico o gene SERPINH1 no carcinoma espinocelular de esôfago (ESCC). Além disso, o miR-29c desempenha um papel regulador nas funções celulares, como proliferação, migração, invasão e ciclo celular do carcinoma espinocelular oral (OSCC), com SERPINH1 sendo um dos alvos identificados (WANG *et al.*, 2021).

No estadiamento avançado do câncer oral, o miR-133a-3p regula seis genes específicos (COL1A1, FBN2, FOXL2, NRIP3, ONECUT2 e SERPINH1). Estudos indicam que a expressão reduzida do miR-133a-3p está correlacionada com doença mais avançada e pior prognóstico em pacientes com tumores na cavidade oral (CRIMI *et al.*, 2020). Além disso, o miR-133a-3p inhibe a proliferação, invasão e migração celular, promovendo apoptose no ESCC (YIN *et al.*, 2019) e do OSCC através do direcionamento do COL1A1 (HE *et al.*, 2018).

Na fase inicial do câncer de laringe, a regulação do gene CLIC5 (*Chloride intracellular channel 5*) é predominantemente influenciada pela superexpressão de miR-105-5p, o que acelera a tumorigênese. Este miRNA também se destaca como um potencial biomarcador para o diagnóstico precoce de câncer de pulmão de células não pequenas (NSCLC) (DONG *et al.*, 2021). Além disso, CLIC5 emerge como um marcador molecular promissor para avaliar o prognóstico e a infiltração imunológica em pacientes com adenocarcinoma pulmonar (LUAD) (BIAN *et al.*, 2023). No estadiamento avançado, o regulador predominante é o miR-1-3p, que tem sido associado a diversos tipos de câncer, agindo como supressor tumoral e mostrando eficácia no diagnóstico e prognóstico de tumores (DAI *et al.*, 2023). A subexpressão de miR-1-3p inibe a proliferação e invasão de células cancerígenas gástricas, visando o gene STC2 (KE *et al.*, 2019), enquanto também regula inversamente a expressão do gene CDK14 em osteossarcoma, inibindo a proliferação celular e promovendo a apoptose (ZHANG *et al.*, 2023).

A análise de enriquecimento da via KEGG revela que, no câncer de cavidade oral e laringe, os genes regulados positivamente estão significativamente enriquecidos em vias como digestão e absorção de proteínas, AGE-RAGE em complicações diabéticas, amebíase, interação ECM-receptor e adesão focal. Essas vias também são observadas em condições como leucoplasia oral, CCP e LSCC (YAO *et al.*, 2023; CHEN *et al.*, 2021; MA *et al.*, 2021). Quanto aos processos biológicos, os genes envolvidos na organização de fibrilas de colágeno, como BMP1, COL12A1, COL1A1, COL1A2, COL3A1, COL5A2, COL5A3 e SERPINH1, são destacados no câncer de cavidade oral. No câncer de laringe, os genes COL1A1, COL1A2, COL3A1, COL5A2 e PXDN estão associados principalmente à resposta celular ao estímulo de aminoácidos e à organização de fibrilas de colágeno. O colágeno fibrilar, regulado positivamente em muitos tipos de câncer, desempenha um papel crucial na matriz extracelular (OUELLETTE *et al.*, 2021).

Em relação à ontologia de doenças, foi identificada uma associação entre cinco genes (COL1A1, COL7A1, ITGA6, LAMC2 e SERPINH1) e o câncer da cavidade oral com condições dermatológicas, incluindo epidermólise bolhosa, doença de pele vesiculobolhosa e doença bolhosa

da pele. Por outro lado, no câncer da laringe, os genes COL1A1, COL1A2, COL3A1 e COL5A2 mostraram uma associação mais proeminente com o síndrome de Ehlers-Danlos. Destaca-se que o COL1A2 (*Collagen Type I Alpha 2 Chain*) demonstrou regulação desregulada em vários tumores, incluindo o ESCC (LI *et al.*, 2020). Por sua vez, a expressão do COL3A1 (*collagen type III alpha 1*) desempenha um papel crucial na proliferação e migração de células tumorais, sendo desregulada em diversas doenças malignas, como evidenciado no Sarcoma de Ewing (TANG *et al.*, 2021). Níveis elevados de COL3A1 foram associados a prognósticos desfavoráveis para pacientes com CCP (SHEN *et al.*, 2023).

Adicionalmente, foram desenvolvidas as redes PPI para o câncer de cavidade oral e laringe. Nestas redes, identificaram-se dez principais genes centrais em ambos os tipos de câncer. Dentre esses genes, o COL1A1 (*Collagen Type I Alpha 1*) apresentou o maior grau de conectividade em ambas as situações. O COL1A1, pertencente à família do colágeno, desempenha um papel crucial na transição epitelial-mesenquimal, uma fase intimamente relacionada à formação de tumores malignos (LI *et al.*, 2022a). A expressão elevada do COL1A1 está associada à progressão e ao prognóstico desfavorável em diversos tipos de cânceres (LIU; LIAO; LI, 2017; HE *et al.*, 2018; LI *et al.*, 2019; HOU *et al.*, 2021; LI *et al.*, 2022a). Além disso, Yu *et al.* (2020) destacou a importância do COL1A1 na matriz extracelular e nos tecidos conjuntivos, associando-o ativamente ao tamanho do tumor e à profundidade da invasão no câncer gástrico. No estudo de Wu *et al.* (2022), COL1A1 e COL1A2 foram identificados como genes centrais no CCP, reforçando a importância desses genes na compreensão da biologia e progressão desses tipos de câncer.

Avaliando a eficácia diagnóstica dos miRNAs no câncer de cavidade oral e laringe, foi notado que a maioria dos miRNAs individuais demonstrou uma notável capacidade discriminatória entre o tumor primário e o tecido normal adjacente em todas as fases da doença. Para aprimorar a precisão do diagnóstico, foram conduzidas análises utilizando combinações de miRNAs, utilizando classificadores baseados em árvores de decisão com os algoritmos CTREE e RFT. Técnicas de reamostragem foram aplicadas por meio de uma validação cruzada em três cenários distintos, abrangendo todos os estadiamento patológicos, estadiamento iniciais e avançados tanto para o câncer de cavidade oral quanto para o de laringe. Em todos esses cenários, foi observado que o RFT, empregando a técnica de reamostragem SMOTE, apresentou um desempenho superior em comparação com o CTREE. Diversas pesquisas, incluindo as conduzidas por Sherafatian e Arjmand (2019), Zhang *et al.* (2020) e Li *et al.* (2022b), demonstraram resultados promissores ao utilizar a técnica de SMOTE para lidar com o desequilíbrio nas classes minoritárias ao durante a construção de modelos de classificação baseados em árvores de decisão, com o objetivo de identificar miRNAs associados ao câncer.

Assim, no câncer de cavidade oral, o algoritmo RFT ressaltou a importância dos miR-101-3p, miR-455-3p, miR-99a-5p e miR-30a-5p na discriminação dos tecidos, indicando seu potencial como biomarcadores diagnósticos. Em fases iniciais, os miR-99a-5p e miR-101-3p

foram identificados como particularmente relevantes, enquanto em estadiamento avançados, miR-30a-5p, miR-196b-5p, miR-455-3p, miR-30a-3p e miR-99a-5p emergiram como biomarcadores de significativa importância. Quanto ao câncer de laringe, tanto em todos os estadiamento patológicos quanto no estadiamento inicial, os mesmos miRNAs foram considerados relevantes, com miR-99a-5p destacando-se como o mais significativo, independente do estadiamento da doença.

Além disso, esses miRNAs foram identificados como biomarcadores em diversos tipos de cânceres. Por exemplo, o miR-101-3p, por exemplo, suprime a apoptose e aumenta a proliferação celular em células de câncer oral (WANG *et al.*, 2020), considerado um biomarcador diagnóstico em lúpus eritematoso sistêmico (ZHAO *et al.*, 2021), carcinoma hepatocelular (YANG *et al.*, 2018), mesotelioma maligno (FILETTI *et al.*, 2021) e transtorno de déficit de atenção e hiperatividade (MARTINEZ; PEPLOW, 2023). O miR-99a-5p também mostra potencial como biomarcador, podendo ser utilizado para detectar metástases linfonodais em pacientes com OSCC (COUSSEAU *et al.*, 2023), câncer gástrico (SAITO *et al.*, 2020) e câncer de mama (GARRIDO-CANO *et al.*, 2020). Da mesma forma, miR-455-3p foi identificado como um potencial biomarcador para a doença de Alzheimer (KUMAR; VIJAYAN; REDDY, 2017) e câncer de fígado (LEE *et al.*, 2021). Adicionalmente, o miR-30a foi considerado um biomarcador promissor e um potencial alvo terapêutico no desenvolvimento e progressão do câncer, oferecendo uma nova perspectiva sobre o tratamento da doença (JIANG; ZHANG; TANG, 2018).

Os miRNAs foram avaliados em sua capacidade prognóstica em diferentes estadiamentos dos cânceres da cavidade oral e laringe através de três abordagens distintas: o modelo de riscos proporcionais de Cox, a árvore de sobrevivência usando o algoritmo CART (com poda baseada no custo-complexidade) e RSF. O modelo de riscos proporcionais de Cox destacou-se pelo seu desempenho superior, superando tanto o CART quanto o RSF. No entanto, esses métodos foram utilizados de forma complementar na identificação de miRNAs.

No câncer de cavidade oral, miR-30a-3p, miR-379-3p, miR-204-5p, miR-4326 e miR-503-3p foram associados à sobrevivência dos pacientes em todos os estadiamentos patológicos. Destacou-se que níveis elevados de expressão de miR-503-3p estavam correlacionados a uma maior taxa de sobrevivência. Em estadiamentos iniciais, vários miRNAs, incluindo miR-381-3p, miR-503-3p, miR-1305, miR-195-5p, miR-204-5p, miR-379-3p e miR-105-5p, foram significativamente associados à sobrevivência dos pacientes. Entretanto, em estadiamentos avançados, a expressão elevada de miR-133b foi associada a um maior risco de morte, enquanto a de miR-133a-3p foi associada a um menor risco.

Em casos de câncer de laringe, miR-139-3p, miR-1-3p, miR-455-3p, miR-4326, miR-301a-3p, miR-6499-3p, miR-210-3p, miR-206 e miR-301a-5p como fatores associados à sobrevivência dos pacientes. Notavelmente, a expressão aumentada de miR-1-3p foi relacionada a um prognóstico desfavorável em todos os estadiamentos patológicos da doença. Além disso, apenas os miR-455-3p e miR-1-3p mostraram associação significativa com a sobrevivência

em estadiamentos avançados, com a expressão elevada de miR-1-3p correlacionada a um pior prognóstico.

Neste cenário, diversos estudos identificaram miRNAs desempenhando papéis cruciais na sobrevivência de pacientes afetados por distintos tipos de câncer. Um exemplo disso é evidenciado por Chen *et al.* (2020), que revela a capacidade inibitória do miR-30a-3p na invasão e migração de células no carcinoma de células renais (RCC), por meio da regulação do gene ATG12. Este estudo sugere o miR-30a-3p como um potencial indicador prognóstico no RCC, associando a baixa expressão a uma pior sobrevivência no câncer de mama triplo-negativo (TURASHVILI *et al.*, 2018). Em LUAD, o miR-30a-3p demonstrou ser um fator preditor de prognóstico valioso (YU *et al.*, 2019) e exerce impacto na atividade, proliferação e migração, regulando o IGF2BP1 e inibindo o desenvolvimento tumoral (XU *et al.*, 2023).

Outros estudos destacaram o papel do miR-30a-3p no tempo de sobrevivência de pacientes com câncer de laringe, onde sua ação supressora se manifesta a partir da regulação do gene BMPR2 (GAO *et al.*, 2023; CONRAD *et al.*, 2023). No adenocarcinoma de estômago (STAD), o miR-379-3p foi associado à sobrevida dos pacientes (QIAN *et al.*, 2022). Por sua vez, o miR-204-5p emerge como um supressor tumoral potencialmente terapêutico, correlacionando-se negativamente com o prognóstico desfavorável em pacientes com CCP (ZHUANG *et al.*, 2020). Além disso, a expressão reduzida de miR-381-3p está vinculada a um prognóstico significativamente mais desfavorável em pacientes com câncer de mama (YU *et al.*, 2021). O miR-381-3p, ao regular negativamente o gene NASP, demonstra sua eficácia na redução da viabilidade, migração, invasão e transição epitelial-mesenquimal das células CCP, destacando-se como um potencial alvo terapêutico para inibir a progressão dessa neoplasia (KONG *et al.*, 2020).

Contudo, a maioria dos miRNAs e mRNAs abordados neste estudo já foram previamente associados a cânceres de cabeça e pescoço, além de outros tipos de câncer. Apesar dessas associações conhecidas, existem ainda miRNAs e mRNAs que permanecem pouco explorados, indicando um potencial promissor para investigações futuras.

8 Conclusão

Com base nos resultados obtidos, foram identificados 58 DEmiRNAs no contexto do câncer da cavidade oral, destacando-se como os mais significantes: miR-455-3p, miR-196b-5p, miR-196a-5p, miR-503-5p, miR-210-3p, miR-101-3p, miR-381-3p, miR-30a-3p, miR-30a-5p e miR-378a-5p. Paralelamente, foram identificadas 1423 DEmRNAs, sendo CA9, LAMC2, INHBA, IL11, TGFBI, PRH2, PRB3, KLK1, NOSTRIN e LPO os mais notáveis neste cenário. Na análise do câncer de laringe, observou-se 44 DEmiRNAs, com miR-196b-5p, miR-196a-5p, miR-615-3p, miR-4652-5p, miR-455-3p, miR-338-5p, miR-139-3p, miR-378c, miR-378d e miR-99a-5p como os mais significativos. Adicionalmente, foram identificadas 1507 DEmRNAs, sendo ESM1, ZIC5, COL4A1, COL4A2, MMP11, EMP1, UNL3, SLC35C1, KAT2B e FCHO2 os mais proeminentes.

De maneira significativa, 29 miRNAs foram compartilhados entre os dois tipos de câncer, indicando um enriquecimento em várias vias KEGG, como carcinogênese viral, vias do câncer, câncer de próstata, desregulação transcricional e câncer de bexiga, sugerindo interconexões entre esses processos patológicos. Identificou-se também dez genes centrais, sendo COL1A1 e COL3A1 os mais importantes em ambos os tipos de cânceres, indicando que o miR-29c-3p é o principal regulador dos genes associados principalmente às vias enriquecidas, como digestão e absorção de proteínas. Adicionalmente, cinco miRNAs (miR-1269a, miR-767-5p, miR-381-5p, miR-133a-3p e miR-133b) estiveram especificamente ligados ao estadiamento avançado do câncer de cavidade oral, regulando genes como COL1A1, FBN2, FOXL2, NRIP3, ONECUT2 e SERPINH1. No câncer de laringe, o miR-105-5p associou-se ao estadiamento inicial, regulando exclusivamente o gene CLIC5.

Na avaliação dos algoritmos de classificação, o RFT superou o CTree para ambos os cânceres. Os miR-101-3p, miR-455-3p, miR-99a-5p e miR-30a-5p foram identificados como potenciais biomarcadores no diagnóstico do câncer de cavidade oral, distinguindo o tumor primário do tecido normal adjacente. No câncer de laringe, miR-99a-5p, miR-139-3p e miR-196b-5p emergiram como candidatos a biomarcadores independentemente do estadiamento da doença.

Em suma, o modelo de riscos proporcionais de Cox foi eficaz na identificação de biomarcadores prognósticos em dois tipos de câncer. No câncer de cavidade oral, a idade e o estadiamento avançado surgiram como fatores de risco para óbito. Além disso, observou-se que a superexpressão de miR-210-3p e miR-101-3p estava correlacionada a uma menor taxa de

sobrevivência, enquanto miR-503-3p estava associada a uma maior taxa de sobrevivência. Em estadiamento avançados, miR-133b foi vinculada a uma menor taxa de sobrevivência, enquanto miR-133a-3p foi associada a uma maior taxa de sobrevivência. No câncer de laringe, somente pacientes do sexo feminino e miR-1-3p apresentaram um prognóstico desfavorável.

Desta forma, os resultados obtidos neste estudo representam uma valiosa contribuição para o avanço do entendimento da biologia molecular dos cânceres de cavidade oral e laringe. Essas descobertas não apenas promovem novas perspectivas no diagnóstico, prognóstico e tratamento da doença, mas também abrem portas para o desenvolvimento de estratégias terapêuticas inovadoras e aprimoramentos na abordagem clínica dos pacientes afetados por cânceres de cabeça e pescoço. Contudo, é crucial ressaltar a necessidade de validar esses miRNAs por meio de estudos experimentais, a fim de avaliar sua relevância clínica.

8.1 Limitações do estudo

Limitações deste estudo não invalidam os resultados obtidos, mas fornecem informações sobre os desafios e as restrições enfrentadas pelos pesquisadores. As limitações incluem dificuldades na obtenção conjunta de dados de expressão de miRNAs maduros e mRNAs do TCGA, bem como a disponibilidade limitada de dados em comparação com outras plataformas, como o GEO e o *ArrayExpress*. Cada plataforma possui limitações distintas em termos de amostras, tipos de dados e abrangência de genes, o que pode restringir a quantidade de dados disponíveis para análise, dependendo dos objetivos do estudo.

Embora o TCGA forneça algumas informações clínicas, estas podem não ser suficientes para análises mais sofisticadas ou relacionadas a doenças específicas, sendo a qualidade dos dados clínicos crucial para a interpretação dos resultados. Em estudos retrospectivos, a obtenção de dados clínicos completos pode ser desafiadora, com informações incompletas identificadas em variáveis como tempo de seguimento, idade, estadiamento do tumor e outros.

A variação na expressão de miRNAs e mRNAs entre os indivíduos também dificultou a identificação de padrões de expressão relevantes para determinar a condição clínica da doença. Essa variação pode ser influenciada por fatores como variações genéticas individuais, fatores ambientais e a heterogeneidade tumoral.

8.2 Trabalhos futuros

Com base nos resultados desta tese, algumas direções promissoras são sugeridas para orientar futuros desenvolvimentos na pesquisa:

- Recomenda-se a realização de análises de classificação e sobrevivência considerando os genes alvo dos miRNAs;

- Recomenda-se aprimorar a análise de expressão diferencial por meio da integração de dados transcriptômicos de diversas plataformas, juntamente com os dados clínicos, utilizando diferentes métodos de redução de dimensionalidade. Para a classificação de tecidos, sugere-se a adoção de modelos alternativos de classificação binária que levem em consideração o desbalanceamento dos dados, empregando funções de ligação assimétricas. Esses modelos devem ser comparados com os algoritmos de classificação do aprendizado de máquina, mediante técnicas como reamostragem e validação cruzada;

- No contexto da análise de sobrevivência, é recomendável comparar métodos paramétricos de sobrevivência com algoritmos de sobrevivência utilizados no aprendizado de máquina, tais como máquinas de vectores de suporte e redes neuronais. Novamente, a utilização de técnicas como reamostragem e validação cruzada é essencial nesse processo. Essas análises, tanto para classificação quanto para sobrevivência, podem ser conduzidas sob as abordagens frequentista ou Bayesiana envolvendo a bioinformática, proporcionando uma compreensão abrangente e robusta dos resultados obtidos.

REFERÊNCIAS

ABDI, H. *et al.* Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics*, Sage Thousand Oaks, CA, v. 3, p. 103–107, 2007. 19

ADELSTEIN, D. *et al.* Nccn guidelines insights: head and neck cancers, version 2.2017. *Journal of the National Comprehensive Cancer Network*, Harborside Press, LLC, v. 15, n. 6, p. 761–770, 2017. 8

AHMAD, A.; IMRAN, M.; AHSAN, H. Biomarkers as biomedical bioindicators: Approaches and techniques for the detection, analysis, and validation of novel biomarkers of diseases. *Pharmaceutics*, MDPI, v. 15, n. 6, p. 1630, 2023. 8

AHMAD, J. *et al.* Microrna in carcinogenesis & cancer diagnostics: a new paradigm. *The Indian Journal of Medical Research*, Wolters Kluwer–Medknow Publications, v. 137, n. 4, p. 680, 2013. 11

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, Ieee, v. 19, n. 6, p. 716–723, 1974. 36

AL-HAMEED, K. A. A. Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications*, Semnan University, v. 13, n. 1, p. 3249–3255, 2022. 45

ALFARO, E.; GÁMEZ, M.; GARCÍA, N. *Ensemble classification methods with applications in R*. [S.l.]: John Wiley & Sons, 2018. 26

ALSAHAFI, E. *et al.* Clinical update on head and neck cancer: molecular biology and ongoing challenges. *Cell Death & Disease*, Nature Publishing Group UK London, v. 10, n. 8, p. 540, 2019. 1, 8

ALWANI, A.; BAJ-KRZYWORZEKA, M. Mikrorna jako cel terapeutyczny w chorobach nowotworowych. *Postępy Biochemii*, v. 67, n. 3, 2021. 11

AMBROS, V. *et al.* A uniform system for microrna annotation. *RNA*, Cold Spring Harbor Lab, v. 9, n. 3, p. 277–279, 2003. 10

AMIN, M. B. *et al. AJCC cancer staging manual*. [S.l.]: Springer, 2017. v. 1024. 8

ANDERS, S.; HUBER, W. Differential expression analysis for sequence count data. *Nature Precedings*, Nature Publishing Group, p. 1–1, 2010. 15, 16, 17, 19

ANDERSON, G. *et al.* An updated review on head and neck cancer treatment with radiation therapy. *Cancers*, MDPI, v. 13, n. 19, p. 4912, 2021. 8

ASSIS, A. F. *et al.* What is the transcriptome and how it is evaluated. In: *Transcriptomics in Health and Disease*. [S.l.]: Springer, 2022. p. 3–50. 12

ATHANASIOS, A. *et al.* Protein-protein interaction (ppi) network: recent advances in drug discovery. *Current Drug Metabolism*, Bentham Science Publishers, v. 18, n. 1, p. 5–10, 2017. 13

AUPÉRIN, A. Epidemiology of head and neck cancers: an update. *Current Opinion in Oncology*, LWW, v. 32, n. 3, p. 178–186, 2020. 7

AZAR, M. R. M. H. *et al.* Altered expression levels of mir-212, mir-133b and mir-27a in tongue squamous cell carcinoma (tscc) with clinicopathological considerations. *Gene Reports*, Elsevier, v. 19, p. 100622, 2020. 115

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. 19, 20, 21

BERTSIMAS, D. *et al.* Optimal survival trees. *Machine Learning*, Springer, v. 111, n. 8, p. 2951–3023, 2022. 37

BIAN, T. *et al.* Identification of clic5 as a prognostic biomarker and correlated immunomodulator for lung adenocarcinoma. *Combinatorial Chemistry & High Throughput Screening*, Bentham Science Publishers, v. 26, n. 14, p. 2452–2468, 2023. 116

BLIGHE, K. *et al.* Package 'pcatools.'. *REAGENT or RESOURCE SOURCE IDENTIFIER Deposited data GitHub repository with code and data for analysis GitHub https://github. com/CharlotteEPage/NFI_Disease_Ecology Raw data tables Mendeley https://data. mendeley. com/datasets/y869bhhmzr/1 Software and algorithms R version*, v. 4, n. 1, p. 9–11, 2019. 44

BORKOVEC, M.; MADIN, N. ggparty:"ggplot" visualizations for the "partykit" package. *R package version*, v. 1, n. 0, 2019. 48

BOZEC, A. *et al.* Current role of primary surgical treatment in patients with head and neck squamous cell carcinoma. *Current Opinion in Oncology*, LWW, v. 31, n. 3, p. 138–145, 2019. 8

BRAY, F. *et al.* Erratum: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, v. 70, n. 4, p. 313, 2020. 6

BREIMAN, L. Random forests. *Machine Learning*, Springer, v. 45, n. 1, p. 5–32, 2001. 26, 28, 29, 37

BRIEMAN, L. *et al. Classification and regression tree analysis*. [S.l.]: CRC Press: Boca Raton, FL, USA, 1984. 27, 28

BRIERLEY, J. D.; GOSPODAROWICZ, M. K.; WITTEKIND, C. *TNM classification of malignant tumours*. [S.l.]: John Wiley & Sons, 2017. 8

BROSEGHINI, E. *et al.* Diagnostic and prognostic value of micrornas in patients with laryngeal cancer: A systematic review. *Non-coding RNA*, MDPI, v. 9, n. 1, p. 9, 2023. 115

BUITRAGO, D. L.; LOVERING, R. C.; CAPORALI, A. Insights into online microrna bioinformatics tools. *Non-coding RNA*, MDPI, v. 9, n. 2, p. 18, 2023. 1

BUJA, A.; EYUBOGLU, N. Remarks on parallel analysis. *Multivariate Behavioral Research*, Taylor & Francis, v. 27, n. 4, p. 509–540, 1992. 24

CALIFF, R. M. Biomarker definitions and their applications. *Experimental Biology and Medicine*, SAGE Publications Sage UK: London, England, v. 243, n. 3, p. 213–221, 2018. 8

CANNING, M. *et al.* Heterogeneity of the head and neck squamous cell carcinoma immune landscape and its impact on immunotherapy. *Frontiers in Cell and Developmental Biology*, Frontiers Media SA, v. 7, p. 52, 2019. 8

CARLBERG, C.; VELLEUER, E. Introduction to cancer. In: ____. *Cancer Biology: How Science Works*. Cham: Springer International Publishing, 2021. p. 1–16. ISBN 978-3-030-75699-4. 4

CARLSON, M. *et al.* org. hs. eg. db: Genome wide annotation for human. *R package version*, v. 3, n. 2, p. 3, 2019. 47

CARVALHO, P. R.; MUNITA, C. S.; LAPOLLI, A. L. Validity studies among hierarchical methods of cluster analysis using cophenetic correlation coefficient. *Brazilian Journal of Radiation Sciences*, v. 7, n. 2A, 2019. 26

CASELLA, G.; BERGER, R. L. *Inferência estatística*. [S.l.]: Cengage Learning, 2010. 15, 16, 19

CATTELL, R. B. The scree test for the number of factors. *Multivariate Behavioral Research*, Taylor & Francis, v. 1, n. 2, p. 245–276, 1966. 24

CECCO, L. D. *et al.* Integrative mirna-gene expression analysis enables refinement of associated biology and prediction of response to cetuximab in head and neck squamous cell cancer. *Genes*, MDPI, v. 8, n. 1, p. 35, 2017. 2, 10

CHARRAD, M. *et al.* Nbclust: an r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, v. 61, p. 1–36, 2014. 45

CHAWLA, N. V. *et al.* Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. 30

CHEN, L. *et al.* Trends in the development of mirna bioinformatics tools. *Briefings in Bioinformatics*, Oxford University Press, v. 20, n. 5, p. 1836–1852, 2019. 1

____. Integrated data analysis implicates plau, serpine1, spp1, and mmp1 as prognostic factors in head and neck squamous cell carcinoma. *Journal of Postgraduate Medicine, Education and Research*, v. 55, n. 4, p. 160–170, 2021. 116

CHEN, R.; ZHANG, Y.; ZHANG, X. Mir-1254 functions as a tumor suppressor in oral squamous cell carcinoma by targeting cd36. *Technology in Cancer Research & Treatment*, SAGE Publications Sage CA: Los Angeles, CA, v. 18, p. 1533033819859447, 2019. 115

CHEN, Y.; WANG, X. mirdb: an online database for prediction of functional microrna targets. *Nucleic Acids Research*, Oxford University Press, v. 48, n. D1, p. D127–D131, 2020. 12

CHEN, Y. *et al.* mir-30a-3p inhibits renal cancer cell invasion and metastasis through targeting atg12. *Translational Andrology and Urology*, AME Publications, v. 9, n. 2, p. 646, 2020. 119

CHIN, C.-H. *et al.* cytohubba: identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology*, BioMed Central, v. 8, n. 4, p. 1–7, 2014. 47

CHOW, L. Q. Head and neck cancer. *New England Journal of Medicine*, Mass Medical Soc, v. 382, n. 1, p. 60–72, 2020. 6, 8

CIAMPI, A. *et al.* Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*, Elsevier, v. 4, n. 3, p. 185–204, 1986. 37

COLAPRICO, A. *et al.* Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research*, Oxford University Press, v. 44, n. 8, p. e71–e71, 2016. 40

COLEVAS, A. D. *et al.* Nccn guidelines insights: head and neck cancers, version 1.2018. *Journal of the National Comprehensive Cancer Network*, Harborside Press, LLC, v. 16, n. 5, p. 479–490, 2018. 8

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Editora Blucher, 2006. 33, 34, 35, 36

CONRAD, O. *et al.* Tumor-suppressive and immunomodulating activity of mir-30a-3p and mir-30e-3p in hnscc cells and tumoroids. *International Journal of Molecular Sciences*, MDPI, v. 24, n. 13, p. 11178, 2023. 119

CORMACK, R. M. A review of classification. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 134, n. 3, p. 321–353, 1971. 26

COUSSEAU, C. P. V. *et al.* mir-99a-5p as a biomarker for lymph node metastasis prediction in oral squamous cell carcinoma patients. *Head & Neck*, Wiley Online Library, v. 45, n. 10, p. 2489–2497, 2023. 118

COWLAND, J. B.; HOTHER, C.; GRØNBAEK, K. Micrornas and cancer. *Apmis*, Wiley Online Library, v. 115, n. 10, p. 1090–1106, 2007. 9

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. 35, 36

CRIMI, S. *et al.* Droplet digital pcr analysis of liquid biopsy samples unveils the diagnostic role of hsa-mir-133a-3p and hsa-mir-375-3p in oral cancer. *Biology*, MDPI, v. 9, n. 11, p. 379, 2020. 11, 116

CRISPIM, D. L. *et al.* Comparação de métodos de agrupamentos hierárquicos aglomerativos em indicadores de sustentabilidade em municípios do estado do pará. *Research, Society and Development*, v. 9, n. 2, p. e60922067–e60922067, 2020. 26

DAI, S. *et al.* The important role of mir-1-3p in cancers. *Journal of Translational Medicine*, Springer, v. 21, n. 1, p. 769, 2023. 116

DALMAN, M. R. *et al.* Fold change and p-value cutoffs significantly alter microarray interpretations. In: SPRINGER. *BMC Bioinformatics*. [S.l.], 2012. v. 13, p. 1–4. 18

DENISSOFF, A. *et al.* Exposure to alcohol and overall survival in head and neck cancer: A regional cohort study. *Head & Neck*, Wiley Online Library, v. 44, n. 10, p. 2109–2117, 2022. 7

D'HAESELEER, P. How does gene expression clustering work? *Nature Biotechnology*, Nature Publishing Group US New York, v. 23, n. 12, p. 1499–1501, 2005. 26

DÍEZ, J.; LÓPEZ-NOVOA, J. Introducción a la biología molecular. *Nefrología*, Elsevier, v. 13, n. 1, p. 14–22, 1993. 9

DILLIES, M.-A. *et al.* A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, Oxford University Press, v. 14, n. 6, p. 671–683, 2013. 17

DONG, X. *et al.* Plasma mir-1247-5p, mir-301b-3p and mir-105-5p as potential biomarkers for early diagnosis of non-small cell lung cancer. *Thoracic Cancer*, Wiley Online Library, v. 12, n. 4, p. 539–548, 2021. 116

D'SOUZA, W.; KUMAR, A. micrornas in oral cancer: moving from bench to bed as next generation medicine. *Oral Oncology*, Elsevier, v. 111, p. 104916, 2020. 11

FALZONE, L. *et al.* Identification of novel micrornas and their diagnostic and prognostic significance in oral cancer. *Cancers*, MDPI, v. 11, n. 5, p. 610, 2019. 46

FAN, J.; UPADHYE, S.; WORSTER, A. Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine*, Cambridge University Press, v. 8, n. 1, p. 19–20, 2006. 31

FANG, R. *et al.* Downregulation of mir-29c-3p is associated with a poor prognosis in patients with laryngeal squamous cell carcinoma. *Diagnostic Pathology*, Springer, v. 14, p. 1–7, 2019. 115

FAWCETT, T. An introduction to roc analysis. *Pattern Recognition Letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. 30, 31

FEI, Y. *et al.* Identification of mirna-mrna crosstalk in laryngeal squamous cell carcinoma. *Molecular Medicine Reports*, Spandidos Publications, v. 16, n. 4, p. 4179–4186, 2017. 115

FEKETE, J. T.; WELKER, Á.; GYŐRFFY, B. Mirna expression signatures of therapy response in squamous cell carcinomas. *Cancers*, MDPI, v. 13, n. 1, p. 63, 2020. 114

FILETTI, V. *et al.* Diagnostic and prognostic value of three micrornas in environmental asbestiform fibers-associated malignant mesothelioma. *Journal of Personalized Medicine*, MDPI, v. 11, n. 11, p. 1205, 2021. 118

FILHO, A. D. P. C. Uso de big data em saúde no brasil: perspectivas para um futuro próximo. *Epidemiologia e Serviços de Saúde*, SciELO Brasil, v. 24, p. 325–332, 2015. 19, 20

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. 22

FOULKES, A. S. *Applied statistical genetics with R*. [S.l.]: Springer-Verlag, 2009. 19

GAI, C. *et al.* Salivary extracellular vesicle-associated mirnas as potential biomarkers in oral squamous cell carcinoma. *BMC cancer*, BioMed Central, v. 18, n. 1, p. 1–11, 2018. 115

GALBIATTI, A. L. S. *et al.* Câncer de cabeça e pescoço: causas, prevenção e tratamento. *Brazilian Journal of Otorhinolaryngology*, SciELO Brasil, v. 79, p. 239–247, 2013. 8

GALILI, T. dendextend: an r package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, Oxford University Press, v. 31, n. 22, p. 3718–3720, 2015. 45

GAO, C.-H.; YU, G.; CAI, P. ggvenndiagram: an intuitive, easy-to-use, and highly customizable r package to generate venn diagram. *Frontiers in Genetics*, Frontiers, p. 1598, 2021. 47

GAO, T. *et al.* mir-30a-3p inhibits the proliferation of laryngeal cancer cells by targeting dnmt3a through regulating dna methylation of pten. *Oncologie*, De Gruyter, n. 0, 2023. 119

GARCIA-MORENO, A.; CARMONA-SAEZ, P. Computational methods and software tools for functional analysis of mirna data. *Biomolecules*, MDPI, v. 10, n. 9, p. 1252, 2020. 1

GARRIDO-CANO, I. *et al.* Circulating mir-99a-5p expression in plasma: A potential biomarker for early diagnosis of breast cancer. *International Journal of Molecular Sciences*, MDPI, v. 21, n. 19, p. 7427, 2020. 118

GAUCH, H. G.; JR, H. G. G. *Multivariate analysis in community ecology*. [S.l.]: Cambridge University Press, 1982. 26

GE, S. X.; JUNG, D.; YAO, R. Shinygo: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, Oxford University Press, v. 36, n. 8, p. 2628–2629, 2020. 48

GENTLEMAN, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, Springer, v. 5, n. 10, p. 1–16, 2004. 16, 40, 43

GIBBONS, J. D.; CHAKRABORTI, S. *Nonparametric statistical inference*. [S.l.]: CRC press, 2014. 21

GLUSMAN, G. *et al.* Optimal scaling of digital transcriptomes. *PloS One*, Public Library of Science San Francisco, USA, v. 8, n. 11, p. e77885, 2013. 17

GODARD, P.; EYLL, J. V. Pathway analysis from lists of micrornas: common pitfalls and alternative strategy. *Nucleic acids Research*, Oxford University Press, v. 43, n. 7, p. 3490–3497, 2015. 13

GOŁĄBEK, K. *et al.* mir-125b-5p, mir-155-3p, and mir-214-5p and target e2f2 gene in oral squamous cell carcinoma. *International Journal of Molecular Sciences*, MDPI, v. 24, n. 7, p. 6320, 2023. 115

GORMLEY, M. *et al.* Reviewing the epidemiology of head and neck cancer: definitions, trends and risk factors. *British Dental Journal*, Nature Publishing Group UK London, v. 233, n. 9, p. 780–786, 2022. 1

GOYAL, N. *et al.* Risk factors for head and neck cancer in more and less developed countries: Analysis from the inhance consortium. *Oral Diseases*, Wiley Online Library, v. 29, n. 4, p. 1565–1578, 2023. 7

GRAMBSCH, P. M.; THERNEAU, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, Oxford University Press, v. 81, n. 3, p. 515–526, 1994. 36

GRANDIS, E.-N. A. C. J.; WHO, J. T. T. S. P. *WHO classification of head and neck tumours*. [S.l.]: IARC. Lyon, 2017. 6

GRIFFITHS-JONES, S. *et al.* mirbase: microrna sequences, targets and gene nomenclature. *Nucleic Acids Research*, Oxford University Press, v. 34, n. suppl_1, p. D140–D144, 2006. 10

GU, Z. Complex heatmap visualization. *Imeta*, Wiley Online Library, v. 1, n. 3, p. e43, 2022. 45

GUO, M. *et al.* Epigenetic heterogeneity in cancer. *Biomarker Research*, BioMed Central, v. 7, n. 1, p. 1–19, 2019. 5

HAIR, J. F. *et al. Análisis multivariante*. [S.l.: s.n.], 2004. 23

____. *Análise multivariada de dados*. [S.l.]: Bookman editora, 2009. 25, 26

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. [S.l.]: Morgan Kaufmann, 2012. 26, 27

HARDIN, J. W. *et al. Generalized linear models and extensions*. [S.l.]: Stata Press, 2007. 16

HARRELL, F. E. *et al.* Evaluating the yield of medical tests. *Jama*, American Medical Association, v. 247, n. 18, p. 2543–2546, 1982. 38

HE, B. *et al.* Mir-133a-3p inhibits oral squamous cell carcinoma (oscc) proliferation and invasion by suppressing col1a1. *Journal of Cellular Biochemistry*, Wiley Online Library, v. 119, n. 1, p. 338–346, 2018. 116, 117

HE, X.; ZHANG, J. Why do hubs tend to be essential in protein networks? *PLoS Genetics*, Public Library of Science San Francisco, USA, v. 2, n. 6, p. e88, 2006. 13

HESKETH, R. *Introduction to Cancer Biology*. [S.l.]: Cambridge University Press, 2023. 4

HILBE, J. M. *Modeling count data*. [S.l.]: Cambridge University Press, 2014. 16

HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. de O. Análise de componentes principais: resumo teórico, aplicação e interpretação. *ES Engineering and Science*, v. 5, n. 1, p. 83–90, 2016. 23

HORN, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika*, Springer, v. 30, p. 179–185, 1965. 24

HOTHORN, T. *et al.* Package 'coin'. 2017. Disponível em: <http://coin.r-forge.r-project.org>. Acesso em: 24 junho 2021. 44

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. 28

HOTHORN, T.; ZEILEIS, A.; HOTHORN, M. T. Package 'partykit'. 2023. 48, 50

HOU, L. *et al.* Collagen type 1 alpha 1 chain is a novel predictive biomarker of poor progression-free survival and chemoresistance in metastatic lung cancer. *Journal of Cancer*, Ivyspring International Publisher, v. 12, n. 19, p. 5723, 2021. 117

HUBER, W. *et al.* Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, Nature Publishing Group, v. 12, n. 2, p. 115–121, 2015. 40, 43

HUSSON, F. *et al.* Package 'factominer'. *An R package*, v. 96, p. 698, 2016. 44

INCA. *Brasil - estimativa dos casos novos: Estimativas para o ano de 2023*. Rio de Janeiro: [s.n.], 2022. Disponível em: <https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa/estado-capital/brasil>. 1, 6

IRIMIE-AGHIORGHIESEI, A. I. *et al.* Prognostic value of mir-21: an updated meta-analysis in head and neck squamous cell carcinoma (hnscc). *Journal of Clinical Medicine*, MDPI, v. 8, n. 12, p. 2041, 2019. 11

ISHWARAN, H. *et al.* Random survival forests. 2008. 37, 38

ISHWARAN, H.; KOGALUR, U. B.; KOGALUR, M. U. B. Package 'randomforestsrc'. *Breast*, v. 6, n. 1, p. 854, 2023. 50

JAMAL-HANJANI, M. *et al.* Translational implications of tumor heterogeneity. *Clinical Cancer Research*, AACR, v. 21, n. 6, p. 1258–1266, 2015. 5, 8

JANITZA, S.; HORNUNG, R. On the overestimation of random forest's out-of-bag error. *PloS one*, Public Library of Science San Francisco, CA USA, v. 13, n. 8, p. e0201904, 2018. 28, 29

JASKOWIAK, P. A.; COSTA, I. G.; CAMPELLO, R. J. Clustering of rna-seq samples: Comparison study on cancer data. *Methods*, Elsevier, v. 132, p. 42–49, 2018. 26

JIANG, L.-h.; ZHANG, H.-d.; TANG, J.-h. Mir-30a: A novel biomarker and potential therapeutic target for cancer. *Journal of Oncology*, Hindawi, v. 2018, 2018. 118

JOHNSON, D. E. *et al.* Head and neck squamous cell carcinoma. *Nature Reviews Disease Primers*, Nature Publishing Group, v. 6, n. 1, p. 1–22, 2020. 1, 6

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. [S.l.]: Pearson London, UK, 2014. 23, 25, 26

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, Springer, v. 23, n. 3, p. 187–200, 1958. 24

KALFERT, D. *et al.* Multifunctional roles of mir-34a in cancer: a review with the emphasis on head and neck squamous cell carcinoma and thyroid cancer with clinical implications. *Diagnostics*, MDPI, v. 10, n. 8, p. 563, 2020. 11

KANTIDAKIS, G. *et al.* Survival prediction models since liver transplantation-comparisons between cox models and machine learning techniques. *BMC Medical Research Methodology*, Springer, v. 20, p. 1–14, 2020. 38

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. 35

KASSAMBARA, A. *et al.* Drawing survival curves using 'ggplot2'. *R package version 0.3*, v. 1, 2017. 49

KE, J. *et al.* Mir-1-3p suppresses cell proliferation and invasion and targets stc2 in gastric cancer. *European Review for Medical & Pharmacological Sciences*, v. 23, n. 20, 2019. 116

KHALID, S. *et al.* Predicting risk through artificial intelligence based on machine learning algorithms: A case of pakistani nonfinancial firms. *Complexity*, Hindawi, v. 2022, 2022. 27

KHAN, M. Y. *et al.* Automated prediction of good dictionary examples (gdex): a comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques. *Complexity*, Hindawi Limited, v. 2021, p. 1–18, 2021. ix, 26, 28, 29

KINGSFORD, C.; SALZBERG, S. L. What are decision trees? *Nature Biotechnology*, Nature Publishing Group, v. 26, n. 9, p. 1011–1013, 2008. 26

KLEIN, J. D.; GRANDIS, J. R. The molecular pathogenesis of head and neck cancer. *Cancer Biology & Therapy*, Taylor & Francis, v. 9, n. 1, p. 1–7, 2010. 4

KLEINBAUM, D. G.; KLEIN, M. *et al. Survival analysis: a self-learning text*. [S.l.]: Springer, 2012. v. 3. 33, 34, 35

KONG, F. *et al.* Mir-381-3p suppresses biological characteristics of cancer in head-neck squamous cell carcinoma cells by targeting nuclear autoantigenic sperm protein (nasp). *Bioscience, Biotechnology, and Biochemistry*, Taylor & Francis, v. 84, n. 4, p. 703–713, 2020. 119

KONINGS, H. *et al.* A literature review of the potential diagnostic biomarkers of head and neck neoplasms. *Frontiers in Oncology*, Frontiers Media SA, v. 10, p. 1020, 2020. 7

KOTSIANTIS, S.; PINTELAS, P. Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics*, v. 1, n. 1, p. 46–55, 2003. 30

KOWALSKI, L. P. *et al.* Survival trends of patients with oral and oropharyngeal cancer treated at a cancer center in são paulo, brazil. *Clinics*, SciELO Brasil, v. 75, 2020. 7

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. 10

KRZYWINSKI, M.; ALTMAN, N. Points of significance: Nonparametric tests. *Nature Methods*, v. 11, n. 5, p. 467–468, 2014. 21

KUHN, D. E. *et al.* Experimental validation of mirna targets. *Methods*, Elsevier, v. 44, n. 1, p. 47–54, 2008. 13

KUHN, M. *et al.* Package 'caret'. *The R Journal*, v. 223, n. 7, 2020. 48

KUMAR, S.; PATNAIK, S.; DIXIT, A. Predictive models for stage and risk classification in head and neck squamous cell carcinoma (hnscc). *PeerJ*, PeerJ Inc., v. 8, p. e9656, 2020. 2

KUMAR, S.; VIJAYAN, M.; REDDY, P. H. Microrna-455-3p as a potential peripheral biomarker for alzheimer's disease. *Human Molecular Genetics*, Oxford University Press, v. 26, n. 19, p. 3808–3822, 2017. 118

KUMAR, V. E. *et al.* Targeting epigenetic modifiers of tumor plasticity and cancer stem cell behavior. *Cells*, MDPI, v. 11, n. 9, p. 1403, 2022. 5

LANGMEAD, B.; HANSEN, K. D.; LEEK, J. T. Cloud-scale rna-sequencing differential expression analysis with myrna. *Genome Biology*, Springer, v. 11, n. 8, p. 1–11, 2010. 15

LE, T. T.; MOORE, J. H. treeheatr: an r package for interpretable decision tree visualizations. *Bioinformatics*, Oxford University Press, v. 37, n. 2, p. 282–284, 2021. 48

LEBLANC, M.; CROWLEY, J. Relative risk trees for censored survival data. *Biometrics*, JSTOR, p. 411–425, 1992. 37

____. Survival trees by goodness of split. *Journal of the American Statistical Association*, Taylor & Francis, v. 88, n. 422, p. 457–467, 1993. 37

LEE, H. M. *et al.* Detection of increased serum mir-122-5p and mir-455-3p levels before the clinical diagnosis of liver cancer in people with type 2 diabetes. *Scientific Reports*, Nature Publishing Group UK London, v. 11, n. 1, p. 23756, 2021. 118

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*, Elsevier, v. 75, n. 5, p. 843–854, 1993. 9

LEE, S.; LIM, H. Review of statistical methods for survival analysis using genomic data. *Genomics & Informatics*, Korea Genome Organization, v. 17, n. 4, 2019. 2, 34

LEI, Y. *et al.* Applications of single-cell sequencing in cancer research: progress and perspectives. *Journal of Hematology & Oncology*, Springer, v. 14, n. 1, p. 91, 2021. 5

LEMESHOW, S.; MAY, S.; HOSMER, D. W. *Applied survival analysis: regression modeling of time-to-event data*. [S.l.]: John Wiley & Sons, 2011. 33

LI, G. *et al.* High expression of collagen 1a2 promotes the proliferation and metastasis of esophageal cancer cells. *Annals of Translational Medicine*, AME Publications, v. 8, n. 24, 2020. 117

LI, J. *et al.* The clinical significance of collagen family gene expression in esophageal squamous cell carcinoma. *PeerJ*, PeerJ Inc., v. 7, p. e7705, 2019. 117

LI, R. *et al.* Gdcrnatools: an r/bioconductor package for integrative analysis of lncrna, mirna and mrna data in gdc. *Bioinformatics*, Oxford University Press, v. 34, n. 14, p. 2515–2517, 2018. 40, 43

LI, W. Volcano plots in analyzing differential expressions with mrna microarrays. *Journal of Bioinformatics and Computational Biology*, World Scientific, v. 10, n. 06, p. 1231003, 2012. 18

LI, X. *et al.* Integrated analysis of microrna (mirna) and mrna profiles reveals reduced correlation between microrna and target gene in cancer. *BioMed Research International*, Hindawi, v. 2018, 2018. 45

____. mir-34a-5p functions as a tumor suppressor in head and neck squamous cell cancer progression by targeting flotillin-2. *International Journal of Biological Sciences*, Ivyspring International Publisher, v. 17, n. 15, p. 4327, 2021. 11

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. 44, 114, 117

LI, Z. *et al.* Identifying key microrna signatures for neurodegenerative diseases with machine learning methods. *Frontiers in Genetics*, Frontiers, v. 13, p. 880997, 2022. 117

LING, C. X.; LI, C. Data mining for direct marketing: Problems and solutions. In: *Kdd*. [S.l.: s.n.], 1998. v. 98, p. 73–79. 30

LIU, J.; DANG, H.; WANG, X. W. The significance of intertumor and intratumor heterogeneity in liver cancer. *Experimental & Molecular Medicine*, Nature Publishing Group, v. 50, n. 1, p. e416–e416, 2018. 5

LIU, S.; LIAO, G.; LI, G. Regulatory effects of col1a1 on apoptosis induced by radiation in cervical cancer cells. *Cancer Cell International*, Springer, v. 17, p. 1–9, 2017. 117

LOPEZ, Y. O. N. *et al.* Characteristic mirna expression signature and random forest survival analysis identify potential cancer-driving mirnas in a broad range of head and neck squamous cell carcinoma subtypes. *Reports of Practical Oncology and Radiotherapy*, v. 23, n. 1, p. 6–20, 2018. 2, 10, 114

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. 16, 17, 18, 19, 43

LU, L. *et al.* A novel seven-mirna prognostic model to predict overall survival in head and neck squamous cell carcinoma patients. *Molecular Medicine Reports*, Spandidos Publications, v. 20, n. 5, p. 4340–4348, 2019. 2, 10

LU, Y. *et al.* Hsa-mir-301a-3p acts as an oncogene in laryngeal squamous cell carcinoma via target regulation of smad4. *Journal of Cancer*, Ivyspring International Publisher, v. 6, n. 12, p. 1260, 2015. 115

LUNARDON, N.; MENARDI, G.; TORELLI, N. Rose: a package for binary imbalanced learning. *R Journal*, v. 6, n. 1, 2014. 30, 48

LUO, K. *et al.* Identification of critical mirnas as novel diagnostic markers for laryngeal squamous cell carcinoma. *Disease Markers*, Hindawi, v. 2022, 2022. 46, 115

LYDIATT, W. M. *et al.* Head and neck cancers—major changes in the american joint committee on cancer eighth edition cancer staging manual. *CA: a Cancer Journal for Clinicians*, Wiley Online Library, v. 67, n. 2, p. 122–137, 2017. 8

MA, J. *et al.* Prediction of the mechanism of mirnas in laryngeal squamous cell carcinoma based on the mirna-mrna regulatory network. *PeerJ*, PeerJ Inc., v. 9, p. e12075, 2021. 116

MAIMON, O. Z.; ROKACH, L. *Data mining with decision trees: theory and applications*. [S.l.]: World scientific, 2014. v. 81. 33

MALENOVÁ, G.; ROWSON, D.; BOEVA, V. Exploring pathway-based group lasso exploring pathway-based group lasso for cancer survival analysis: a special case of multi-task learning. *Frontiers in Genetics*, Frontiers, p. 2276, 2021. 33

MANLY, B. F.; ALBERTO, J. A. N. *Multivariate statistical methods: a primer*. [S.l.]: Chapman and Hall/CRC, 2016. 23

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. 21

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. 35

MARIONI, J. C. *et al.* Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, Cold Spring Harbor Lab, v. 18, n. 9, p. 1509–1517, 2008. 17

MARTINEZ, B.; PEPLOW, P. V. Micrornas as potential biomarkers for diagnosis of attention deficit hyperactivity disorder. *Neural Regeneration Research*, Medknow, 2023. 118

MARZBAN, C. The roc curve and the area under it as performance measures. *Weather and Forecasting*, American Meteorological Society, v. 19, n. 6, p. 1106–1114, 2004. 31

MCDERMAID, A. *et al.* Interpretation of differential gene expression results of rna-seq data: review and integration. *Briefings in Bioinformatics*, Oxford University Press, v. 20, n. 6, p. 2044–2054, 2019. 2, 12

MCHUGH, M. L. The chi-square test of independence. *Biochemia Medica*, Medicinska naklada, v. 23, n. 2, p. 143–149, 2013. 22

MENARD, S. *Applied logistic regression analysis*. [S.l.]: Sage, 2002. 14

MINGOTI, S. A. Análise de dados através de métodos estatística multivariada: uma abordagem aplicada. In: *Análise de dados através de métodos estatística multivariada: uma abordagem aplicada*. [S.l.: s.n.], 2007. p. 295–295. 23, 26

MIRZA, B. *et al.* Machine learning and integrative analysis of biomedical big data. *Genes*, MDPI, v. 10, n. 2, p. 87, 2019. 2

MODY, M. D. *et al.* Head and neck cancer. *The Lancet*, Elsevier, v. 398, n. 10318, p. 2289–2299, 2021. 1, 6

MONARD, M. C.; BARANAUSKAS, J. A. Indução de regras e árvores de decisão. *Sistemas Inteligentes-fundamentos e Aplicações*, Editora Manole Ltda, v. 1, p. 115–139, 2003. 28

MONROY, L. G. D.; RIVERA, M. A. M.; DÁVILA, L. R. L. *Análisis estadístico de datos categóricos*. [S.l.]: Universidad Nacional de Colombia, 2018. 15

NAIR, J. *et al.* Gene and mirna expression changes in squamous cell carcinoma of larynx and hypopharynx. *Genes & cancer*, Impact Journals, LLC, v. 6, n. 7-8, p. 328, 2015. 2

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. 14

NETWORK, C. G. A. *et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, Nature Publishing Group, v. 517, n. 7536, p. 576, 2015. 1, 11

NIGRO, C. L. *et al.* Head and neck cancer: improving outcomes with a multidisciplinary approach. *Cancer Management and Research*, Dove Press, v. 9, p. 363, 2017. 8

O'BRIEN, J. *et al.* Overview of microrna biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology*, Frontiers Media SA, v. 9, p. 402, 2018. 12

OSHIMA, S. *et al.* Identification of tumor suppressive genes regulated by mir-31-5p and mir-31-3p in head and neck squamous cell carcinoma. *International Journal of Molecular Sciences*, MDPI, v. 22, n. 12, p. 6199, 2021. 115

OTASEK, D. *et al.* Cytoscape automation: empowering workflow-based network analysis. *Genome Biology*, Springer, v. 20, p. 1–15, 2019. 47

OUELLETTE, J. N. *et al.* Navigating the collagen jungle: the biomedical potential of fiber organization in cancer. *Bioengineering*, MDPI, v. 8, n. 2, p. 17, 2021. 116

PANVONGSA, W. *et al.* Plasma extracellular vesicle microrna-491-5p as diagnostic and prognostic marker for head and neck squamous cell carcinoma. *Cancer Science*, Wiley Online Library, v. 112, n. 10, p. 4257–4269, 2021. 2, 10, 114

PASI, F. *et al.* Radiation-induced circulating mirna expression in blood of head and neck cancer patients. *Radiation and Environmental Biophysics*, Springer, p. 1–8, 2020. 1

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. 22

PECORINO, L. *Molecular biology of cancer: mechanisms, targets, and therapeutics*. [S.l.]: Oxford University Press, 2021. 4

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. 8

PETRONACCI, C. M. C. *et al.* mirnas expression of oral squamous cell carcinoma patients: Validation of two putative biomarkers. *Medicine*, Wolters Kluwer Health, v. 98, n. 13, 2019. 115

PIOTROWSKI, I. *et al.* mirnas as biomarkers for diagnosing and predicting survival of head and neck squamous cell carcinoma patients. *Cancers*, Multidisciplinary Digital Publishing Institute, v. 13, n. 16, p. 3980, 2021. 2, 10, 115

PORCHERI, C.; MITSIADIS, T. A. Notch in head and neck cancer. In: *Notch Signaling in Embryology and Cancer*. [S.l.]: Springer, 2020. p. 81–103. 6

PROVINCE, M. A. 30 sequential methods of analysis for genome scans. Elsevier, 2001. 20

PURAM, S. V.; ROCCO, J. W. Molecular aspects of head and neck cancer therapy. *Hematology/Oncology Clinics*, Elsevier, v. 29, n. 6, p. 971–992, 2015. 7

QI, C.-L. *et al.* Integrated analysis of deregulation microrna expression in head and neck squamous cell carcinoma. *Medicine*, Wolters Kluwer Health, v. 100, n. 6, 2021. 2, 10, 115

QIAN, H. *et al.* Identification of mirna biomarkers for stomach adenocarcinoma. *BMC Bioinformatics*, Springer, v. 23, n. 1, p. 181, 2022. 119

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, Springer, v. 1, n. 1, p. 81–106, 1986. 27

____. Program for machine learning. *C4.5*, Morgan Kaufmann Pub, 1993. 27

QUIRICO, L.; ORSO, F. The power of micrornas as diagnostic and prognostic biomarkers in liquid biopsies. *Cancer Drug Resistance*, OAE Publishing Inc, v. 3, n. 2, p. 117, 2020. 5

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. 16, 40, 43, 44

RANA, R.; SINGHAL, R. *et al.* Chi-square test and its application in hypothesis testing. *Journal of the Practice of Cardiovascular Sciences*, Medknow Publications, v. 1, n. 1, p. 69, 2015. 22

REINHART, B. J. *et al.* The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *Nature*, Nature Publishing Group, v. 403, n. 6772, p. 901–906, 2000. 10

RENCHER, A. C. *Methods of Multivariate Analysis, 2nd edn.* New York: A John Wiley & Sons, 2002. v. 2. 23, 25, 26

RIPLEY, B. *et al.* Package 'mass'. *Cran R*, v. 538, p. 113–120, 2013. 49

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. 46

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. 16, 17, 43, 44

ROBINSON, M. D.; SMYTH, G. K. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, Oxford University Press, v. 9, n. 2, p. 321–332, 2008. 15

RODRIGUES, R. M. *et al.* How pathological criteria can impact prognosis of tongue and floor of the mouth squamous cell carcinoma. *Journal of Applied Oral Science*, SciELO Brasil, v. 28, 2019. 7

ROSA, G. J. d. M.; ROCHA, L. B. d.; FURLAN, L. R. Estudos de expressão gênica utilizando-se microarrays: delineamento, análise, e aplicações na pesquisa zootécnica. *Revista Brasileira de Zootecnia*, SciELO Brasil, v. 36, p. 186–209, 2007. 20

SAITO, R. *et al.* mir-99a-5p as possible diagnostic and prognostic marker in patients with gastric cancer. *Journal of Surgical Research*, Elsevier, v. 250, p. 193–199, 2020. 118

SALAHUDDIN, S. *et al.* Human immunodeficiency virus is associated with poor overall survival among patients with head and neck cancer. *Clinical Infectious Diseases*, Oxford University Press US, v. 76, n. 8, p. 1449–1458, 2023. 7

SALINAS-RODRÍGUEZ, A.; MANRIQUE-ESPINOZA, B.; SOSA-RUBÍ, S. G. Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud. *Salud pública de México*, SciELO Public Health, v. 51, p. 397–406, 2009. 15

SCHOENFELD, D. Partial residuals for the proportional hazards regression model. *Biometrika*, Oxford University Press, v. 69, n. 1, p. 239–241, 1982. 37

SHAW, R.; BEASLEY, N. Aetiology and risk factors for head and neck cancer: United kingdom national multidisciplinary guidelines. *The Journal of Laryngology & Otology*, Cambridge University Press, v. 130, n. S2, p. S9–S12, 2016. 1, 6

SHEN, Y. *et al.* Col3a1: Potential prognostic predictor for head and neck cancer based on immune-microenvironment alternative splicing. *Cancer Medicine*, Wiley Online Library, v. 12, n. 4, p. 4882–4894, 2023. 117

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. 117

SIDDIQUI, I. A. *et al.* Resveratrol nanoformulation for cancer prevention and therapy. *Annals of the New york Academy of Sciences*, Wiley Online Library, v. 1348, n. 1, p. 20–31, 2015. 4

SIEGEL, R. L. *et al.* Cancer statistics, 2021. *Ca Cancer J Clin*, v. 71, n. 1, p. 7–33, 2021. 6

SIEGMUND, K. D.; LAIRD, P. W.; LAIRD-OFFRINGA, I. A. A comparison of cluster analysis methods using dna methylation data. *Bioinformatics*, Oxford University Press, v. 20, n. 12, p. 1896–1904, 2004. 26

SMYTH, G. K.; YANG, Y. H.; SPEED, T. Statistical issues in cdna microarray data analysis. *Functional genomics: methods and protocols*, Springer, p. 111–136, 2003. 18

SONESON, C.; DELORENZI, M. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, BioMed Central, v. 14, n. 1, p. 1–18, 2013. 12

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, 2021. 1, 6

TAHERKHANI, A. *et al.* Pathogenesis and prognosis of primary oral squamous cell carcinoma based on micrornas target genes: a systems biology approach. *Genomics & Informatics*, Korea Genome Organization, v. 20, n. 3, 2022. 115

TANG, M. *et al.* Col3a1 and its related molecules as potential biomarkers in the development of human ewing's sarcoma. *BioMed Research International*, Hindawi, v. 2021, 2021. 117

TASTSOGLOU, S. *et al.* Diana-mirpath v4. 0: expanding target-based mirna functional analysis in cell-type and tissue contexts. *Nucleic Acids Research*, Oxford University Press, p. gkad431, 2023. 46

THABTAH, F. *et al.* Data imbalance in classification: Experimental evaluation. *Information Sciences*, Elsevier, v. 513, p. 429–441, 2020. 30

THERNEAU, T. *et al.* rpart: Recursive partitioning and regression trees. 2022. 50

THERNEAU, T. M.; LUMLEY, T. Survival analysis. *R Top Doc*, v. 128, n. 10, p. 28–33, 2015. 49

THOMAIDOU, A. C. *et al.* Promising biomarkers in head and neck cancer: The most clinically important mirnas. *International Journal of Molecular Sciences*, MDPI, v. 23, n. 15, p. 8257, 2022. 11

TORGO, L.; TORGO, M. L. Package 'dmwr'. *Comprehensive R archive network*, Citeseer, 2013. 48

TURASHVILI, G. *et al.* Novel prognostic and predictive microrna targets for triple-negative breast cancer. *The FASEB Journal*, Wiley Online Library, v. 32, n. 11, p. 5937–5954, 2018. 119

TURKMAN, M. A. A.; SILVA, G. L. Modelos lineares generalizados-da teoria à prática. *Sociedade Portuguesa de Estatística, Lisboa*, 2000. 14, 15

UKEY, S. *et al.* Study of microrna (mir-221-3p, mir-133a-3p, and mir-9-5p) expressions in oral submucous fibrosis and squamous cell carcinoma. *Indian Journal of Clinical Biochemistry*, Springer, v. 38, n. 1, p. 73–82, 2023. 115

VAHABI, M.; BLANDINO, G.; AGOSTINO, S. D. Micrornas in head and neck squamous cell carcinoma: a possible challenge as biomarkers, determinants for the choice of therapy and targets for personalized molecular therapies. *Translational Cancer Research*, AME Publications, v. 10, n. 6, p. 3090, 2021. 1, 11

VERHOEVEN, K. J.; SIMONSEN, K. L.; MCINTYRE, L. M. Implementing false discovery rate control: increasing your power. *Oikos*, Wiley Online Library, v. 108, n. 3, p. 643–647, 2005. 19

VILLE, B. D. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 5, n. 6, p. 448–455, 2013. 26

WALAYAT, A.; YANG, M.; XIAO, D. Therapeutic implication of mirna in human disease. *Antisense Therapy*, IntechOpen, 2018. 9

WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, v. 54, n. 3, p. 426–482, 1943. 18

WALTERS, S. J. *What is a Cox model?* [S.l.]: Citeseer, 1999. 35

WANG, C. *et al.* Mir-29c inhibits the metastasis of oral squamous cell carcinoma and promotes its cell cycle arrest by targeting serpinh1. *Annals of Translational Medicine*, AME Publications, v. 9, n. 18, 2021. 116

WANG, H. *et al.* mir-101-3p and mir-199b-5p promote cell apoptosis in oral cancer by targeting bicc1. *Molecular and Cellular Probes*, Elsevier, v. 52, p. 101567, 2020. 118

WANG, Y. *et al.* Mir-34a suppresses hnscc growth through modulating cell cycle arrest and senescence. *Neoplasma*, v. 64, n. 4, p. 543–553, 2017. 11

WANG, Z. *et al.* Microrna-1-3p inhibits the proliferation and migration of oral squamous cell carcinoma cells by targeting dkk1. *Biochemistry and Cell Biology*, NRC Research Press, v. 96, n. 3, p. 355–364, 2018. 115

WEI, D. *et al.* mir-29c-3p represses the angiogenesis of esophageal squamous cell carcinoma by targeting serpinh1 to regulate the wnt signaling pathway. *Acta Cirúrgica Brasileira*, SciELO Brasil, v. 38, p. e385223, 2023. 116

WEI, M. *et al.* Predicting therapeutic responses in head and neck squamous cell carcinoma from tp53 mutation detected by cell-free dna. *Translational Cancer Research*, AME Publications, v. 12, n. 12, p. 3604, 2023. 11

WEI, T. *et al.* Package 'corrplot'. *Statistician*, v. 56, n. 316, p. e24, 2017. 45

WEISS, C. N.; ITO, K. A macro view of micrornas: the discovery of micrornas and their role in hematopoiesis and hematologic disease. *International Review of Cell and Molecular Biology*, Elsevier, v. 334, p. 99–175, 2017. 9

WICKHAM, H.; CHANG, W.; WICKHAM, M. H. Package 'ggplot2'. *Create Elegant Data Visualisations using the Grammar of Graphics*, v. 2, n. 1, p. 1–189, 2016. 43, 47, 49, 50

WIENKE, A. *Frailty models in survival analysis*. [S.l.]: Chapman and Hall/CRC, 2010. 33

WILCOXON, F. *Individual comparisons by ranking methods. Biom Bull 1 (6): 80–83*. [S.l.]: JSTOR, 1945. 21

WILLIAMS, G. *Data mining with Rattle and R: The art of excavating data for knowledge discovery*. [S.l.]: Springer Science & Business Media, 2011. 50

WONG, N. W. *et al.* Oncomir: an online resource for exploring pan-cancer microrna dysregulation. *Bioinformatics*, Oxford University Press, v. 34, n. 4, p. 713–715, 2018. 45

WU, F. *et al.* A prognostic model for oral squamous cell carcinoma using 7 genes related to tumor mutational burden. *BMC Oral Health*, Springer, v. 22, n. 1, p. 152, 2022. 117

WU, T. *et al.* clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The innovation*, Elsevier, v. 2, n. 3, 2021. 47

XU, F. *et al.* dbdemc 3.0: functional exploration of differentially expressed mirnas in cancers of human and model organisms. *Genomics, Proteomics and Bioinformatics*, Oxford University Press, v. 20, n. 3, p. 446–454, 2022. xiv, 45, 46

XU, J. *et al.* mir-30a-3p/igf2bp1 affect the proliferation and migration of lung adenocarcinoma cells. *J Clin Images Med Case Rep*, v. 4, n. 10, p. 2628, 2023. 119

YAKA, H. *et al.* A comparison of machine learning algorithms for estimation of higher heating values of biomass and fossil fuels from ultimate analysis. *Fuel*, Elsevier, v. 320, p. 123971, 2022. 32

YANG, C.-C. *et al.* Differential impact of age on survival in head and neck cancer according to classic cox regression and decision tree analysis. *Clinical Otolaryngology*, Wiley Online Library, v. 44, n. 3, p. 244–253, 2019. 7

YANG, X. *et al.* Diagnostic value of strand-specific mi rna-101-3p and mi rna-101-5p for hepatocellular carcinoma and a bioinformatic analysis of their possible mechanism of action. *FEBS Open Bio*, Wiley Online Library, v. 8, n. 1, p. 64–84, 2018. 118

YAO, L. *et al.* Analysis of transcriptome expression profiling data in oral leukoplakia and early and late-stage oral squamous cell carcinoma. *Oncology Letters*, Spandidos Publications, v. 25, n. 4, p. 1–11, 2023. 116

YEUNG, K. Y.; RUZZO, W. L. Principal component analysis for clustering gene expression data. *Bioinformatics*, Oxford University Press, v. 17, n. 9, p. 763–774, 2001. 23

YIN, Y. *et al.* mir-133a-3p suppresses cell proliferation, migration, and invasion and promotes apoptosis in esophageal squamous cell carcinoma. *Journal of Cellular Physiology*, Wiley Online Library, v. 234, n. 8, p. 12757–12770, 2019. 116

YOON, A. J. *et al.* Prognostic value of mir-375 and mir-214-3p in early stage oral squamous cell carcinoma. *American Journal of Translational Research*, e-Century Publishing Corporation, v. 6, n. 5, p. 580, 2014. 115

YU, C. *et al.* Identification of key genes and signaling pathways associated with the progression of gastric cancer. *Pathology & Oncology Research*, Springer, v. 26, p. 1903–1919, 2020. 117

YU, N. *et al.* Identification of tumor suppressor mirnas by integrative mirna and mrna sequencing of matched tumor–normal samples in lung adenocarcinoma. *Molecular oncology*, Wiley Online Library, v. 13, n. 6, p. 1356–1368, 2019. 119

YU, Y.-Z. *et al.* mir-381-3p suppresses breast cancer progression by inhibition of epithelial–mesenchymal transition. *World Journal of Surgical Oncology*, Springer, v. 19, p. 1–11, 2021. 119

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. 28

ZHANG, G. *et al.* mir-1-3p inhibits osteosarcoma cell proliferation and cell cycle progression while promoting cell apoptosis by targeting cdk14 to inactivate wnt/beta-catenin signaling. *Molecular Biotechnology*, Springer, p. 1–14, 2023. 116

ZHANG, J. M. *et al.* An interpretable framework for clustering single-cell rna-seq datasets. *BMC Bioinformatics*, BioMed Central, v. 19, n. 1, p. 1–12, 2018. 26

ZHANG, L. *et al.* Functional analysis of mir-767-5p during the progression of hepatocellular carcinoma and the clinical relevance of its dysregulation. *Histochemistry and Cell Biology*, Springer, v. 154, p. 231–243, 2020. 117

ZHAO, X.; CUI, L. A robust six-mirna prognostic signature for head and neck squamous cell carcinoma. *Journal of Cellular Physiology*, Wiley Online Library, v. 235, n. 11, p. 8799–8811, 2020. 115

ZHAO, X. *et al.* mir-101-3p negatively regulates inflammation in systemic lupus erythematosus via mapk1 targeting and inhibition of the nf-κb pathway. *Molecular Medicine Reports*, Spandidos Publications, v. 23, n. 5, p. 1–13, 2021. 118

ZHUANG, Z. *et al.* Microrna-204-5p is a tumor suppressor and potential therapeutic target in head and neck squamous cell carcinoma. *Theranostics*, Ivyspring International Publisher, v. 10, n. 3, p. 1433, 2020. 119

APÊNDICES

A Variáveis clínicas, miRNA-Seq e mRNA-Seq

A.1 Análise bivariada das características clinicopatológicas

Tabela 16 – Distribuição das características clínico-patológicas em relação ao tecido normal e tumoral nos sítios anatômicos do CCP.

Características	Categorias	Sítio anatômico do tumor									
		Cavidade oral					Laringe				
		NAT		T		Valor p	NAT		T		Valor p
		$n = 28$		$n = 293$			$n = 12$		$n = 100$		
		f_a	f_p(%)	f_a	f_p(%)		f_a	f_p(%)	f_a	f_p(%)	
Sexo	Feminino	11	39,29	99	33,79	0,558*	2	16,67	15	15,00	1,000**
	Masculino	17	60,71	194	66,21		10	83,33	85	85,00	
Raça/Cor	Branco	26	92,86	247	84,30	1,000**	10	83,33	77	77,00	0,151**
	Preto	2	7,14	19	6,48		0	0,00	14	14,00	
	Outros	0	0,00	9	3,07		1	8,33	2	2,00	
Alcoolismo	Sim	16	57,14	186	63,48	0,448*	5	41,67	63	63,00	0,206*
	Não	12	42,86	103	35,15		7	58,33	35	35,00	
Historia do câncer	Sim	2	7,14	12	4,10	0,350**	11	91,67	96	96,00	0,439**
	Não	26	92,86	281	95,90		1	8,33	4	4,00	
Histologia	CCE	26	92,86	251	85,67	0,350**	11	91,67	87	87,00	1,000**
	CCEQ	2	7,14	39	13,31		1	8,33	10	10,00	
	CCENQ	0	0,00	0	0,00		0	0,00	1	1,00	
	CCEB	0	0,00	3	1,02		0	0,00	2	2,00	
Estadiamento patológico	I - II	11	39,29	73	24,91	0,099*	6	50,00	88	88,00	0,004*
	III - IV	17	60,71	220	75,09		6	50,00	12	12,00	
Estadiamento T	T1 - T2	15	53,57	123	41,98	0,237*	6	50,00	19	19,00	0,025*
	T3 - T4	13	46,43	170	58,02		6	50,00	81	81,00	
Estadiamento N	NX	6	21,43	22	7,51	0,059*	5	41,67	6	6,00	0,001**
	N0	10	35,71	118	40,27		5	41,67	40	40,00	
	N1 - N3	12	42,86	153	52,22		2	16,67	54	54,00	
Estadiamento M	MX	3	10,71	34	11,60	0,042**	0	0,00	8	8,00	1,000**
	M0	1	3,57	118	40,27		0	0,00	38	38,00	
	M1	0	0,00	0	0,00		1	8,33	1	1,00	
Estado vital	Vivo	8	28,57	158	53,92	0,010*	4	33,33	61	61,00	0,117**
	Óbito (evento)	20	71,43	135	46,08		8	66,67	39	39,00	

Nota: *Valor p obtido pelo teste Qui-quadrado de Pearson, **Valor p obtido pelo teste exato de Fisher.

A.2 Identificação de DEmiRNAs pelo DESeq2: oral e laringe

Tabela 17 – Estimativas dos parâmetros do modelo Binomial Negativa da expressão dos DEmiRNAs na comparação dos tecidos (tumoral vs. normal) na cavidade oral.

miRNAs	Estimativas		Erro-Padrão ($\hat{\beta}_1$)	Teste Wald	Exp($\hat{\beta}_1$)	$IC_{95\%}$(RR)	FDR-BH
	Normal ($\hat{\beta}_0$)	Tumoral ($\hat{\beta}_1$)			RR		
Superexpresso (↑)							
miR-455-3p	8,11	2,79	0,19	14,62	6,92	5,29 – 8,88	$6{,}52\times10^{-46}$
miR-196b-5p	5,65	3,65	0,26	14,07	12,52	8,64 – 17,55	$1{,}23\times10^{-42}$
miR-196a-5p	4,59	3,77	0,31	12,03	13,58	8,63 – 20,37	$1{,}43\times10^{-31}$
miR-503-5p	2,71	2,65	0,24	10,87	6,38	4,50 – 8,87	$6{,}20\times10^{-26}$
miR-210-3p	9,66	2,77	0,26	10,56	6,81	4,68 – 9,57	$1{,}55\times10^{-24}$
miR-31-5p	5,91	3,19	0,31	10,41	9,15	5,87 – 13,60	$6{,}36\times10^{-24}$
miR-615-3p	-0,77	4,18	0,41	10,23	19,57	10,97 – 36,75	$3{,}40\times10^{-23}$
miR-224-5p	7,36	2,12	0,22	9,90	4,36	3,22 – 5,77	$7{,}73\times10^{-22}$
miR-877-5p	1,15	2,35	0,25	9,26	5,25	3,66 – 7,54	$2{,}65\times10^{-19}$
miR-1910-5p	-1,56	4,28	0,47	9,21	19,48	9,77 – 42,58	$4{,}16\times10^{-19}$
miR-4652-5p	0,19	3,79	0,42	9,09	13,47	7,44 – 23,78	$1{,}24\times10^{-18}$
miR-301a-5p	1,65	2,58	0,29	9,07	6,25	4,15 – 9,23	$1{,}44\times10^{-18}$
miR-365a-5p	0,97	2,50	0,28	9,06	5,51	3,74 – 8,12	$1{,}53\times10^{-18}$
miR-937-3p	1,75	2,64	0,30	8,88	5,93	3,89 – 8,83	$6{,}77\times10^{-18}$
miR-105-5p	2,68	5,53	0,65	8,57	46,51	17,13 – 102,89	$1{,}03\times10^{-16}$
miR-503-3p	-0,35	3,01	0,37	8,16	8,38	4,96 – 14,56	$2{,}67\times10^{-15}$
miR-1269a	3,96	4,92	0,61	8,09	18,65	7,29 – 39,47	$4{,}81\times10^{-15}$
miR-31-3p	6,05	2,42	0,31	7,92	5,35	3,44 – 7,94	$1{,}80\times10^{-14}$
miR-1293	2,63	2,57	0,33	7,91	5,95	3,72 – 9,12	$1{,}99\times10^{-14}$
miR-767-5p	2,47	5,25	0,67	7,82	38,33	13,45 – 87,46	$3{,}92\times10^{-14}$
miR-7705	0,36	2,17	0,31	7,12	4,66	3,02 – 7,26	$6{,}60\times10^{-12}$
miR-4664-3p	-0,12	2,29	0,35	6,57	4,70	2,91 – 7,76	$2{,}58\times10^{-10}$
miR-1305	-0,62	2,76	0,42	6,52	7,15	3,91 – 13,44	$3{,}50\times10^{-10}$
miR-301b-3p	-0,41	2,37	0,38	6,22	5,32	3,08 – 9,41	$2{,}22\times10^{-09}$
miR-4326	2,75	2,01	0,34	5,91	4,02	2,47 – 6,25	$1{,}43\times10^{-08}$
miR-155-3p	-0,52	2,37	0,42	5,69	5,13	2,85 – 9,38	$4{,}59\times10^{-08}$
miR-9-5p	11,21	2,12	0,42	5,08	4,34	2,35 – 7,34	$1{,}18\times10^{-06}$
Subexpresso (↓)							
miR-101-3p	17,06	-2,29	0,15	-15,34	0,20	0,17 – 0,25	$2{,}49\times10^{-50}$
miR-381-3p	12,14	-3,62	0,26	-13,69	0,08	0,06 – 0,12	$1{,}93\times10^{-40}$
miR-30a-3p	17,01	-2,89	0,22	-12,90	0,14	0,10 – 0,18	$4{,}97\times10^{-36}$
miR-30a-5p	15,75	-2,37	0,19	-12,38	0,19	0,15 – 0,25	$3{,}19\times10^{-33}$
miR-378a-5p	11,31	-2,20	0,18	-12,35	0,22	0,17 – 0,28	$3{,}91\times10^{-33}$
miR-378c	8,06	-2,76	0,23	-12,04	0,15	0,11 – 0,20	$1{,}37\times10^{-31}$
miR-299-5p	7,73	-2,81	0,24	-11,90	0,14	0,10 – 0,19	$5{,}94\times10^{-31}$
miR-136-3p	7,80	-2,37	0,21	-11,41	0,19	0,14 – 0,25	$1{,}87\times10^{-28}$
miR-99a-5p	12,78	-2,54	0,23	-10,97	0,17	0,12 – 0,23	$2{,}13\times10^{-26}$
miR-29c-3p	15,07	-2,46	0,22	-10,98	0,18	0,13 – 0,24	$2{,}13\times10^{-26}$
miR-411-5p	8,31	-2,79	0,26	-10,81	0,15	0,09 – 0,20	$1{,}15\times10^{-25}$
miR-375	16,13	-4,18	0,40	-10,47	0,06	0,03 – 0,09	$3{,}42\times10^{-24}$
miR-139-3p	8,22	-2,32	0,23	-10,27	0,20	0,14 – 0,27	$2{,}61\times10^{-23}$
miR-378a-3p	14,53	-2,36	0,23	-10,22	0,20	0,14 – 0,26	$3{,}83\times10^{-23}$
miR-195-5p	8,82	-2,14	0,21	-10,16	0,23	0,17 – 0,30	$6{,}52\times10^{-23}$
miR-376c-3p	7,15	-2,27	0,22	-10,16	0,21	0,15 – 0,28	$6{,}52\times10^{-23}$
miR-30c-2-3p	8,60	-2,41	0,24	-10,07	0,19	0,13 – 0,26	$1{,}57\times10^{-22}$
miR-378d	4,60	-2,82	0,29	-9,62	0,14	0,09 – 0,21	$1{,}05\times10^{-20}$
let-7c-5p	14,88	-2,27	0,24	-9,56	0,21	0,15 – 0,28	$1{,}82\times10^{-20}$
miR-379-5p	14,36	-2,08	0,22	-9,53	0,24	0,17 – 0,32	$2{,}22\times10^{-20}$
miR-411-3p	5,29	-2,47	0,27	-9,31	0,18	0,12 – 0,26	$1{,}80\times10^{-19}$
miR-410-3p	8,33	-2,33	0,25	-9,30	0,20	0,14 – 0,28	$1{,}98\times10^{-19}$
miR-379-3p	4,54	-2,52	0,28	-9,09	0,17	0,12 – 0,25	$1{,}24\times10^{-18}$
miR-486-5p	11,97	-2,18	0,25	-8,58	0,22	0,15 – 0,31	$9{,}32\times10^{-17}$
miR-377-5p	5,64	-2,03	0,24	-8,45	0,25	0,17 – 0,34	$2{,}10\times10^{-16}$
miR-381-5p	4,20	-2,42	0,35	-6,86	0,18	0,11 – 0,29	$3{,}70\times10^{-11}$
miR-1-3p	14,02	-3,02	0,51	-5,97	0,12	0,06 – 0,23	$1{,}01\times10^{-08}$
miR-204-5p	6,14	-2,40	0,41	-5,85	0,19	0,10 – 0,31	$1{,}94\times10^{-08}$
miR-6510-3p	7,50	-2,08	0,36	-5,75	0,24	0,14 – 0,38	$3{,}46\times10^{-08}$
miR-133a-3p	13,48	-2,68	0,50	-5,31	0,16	0,07 – 0,29	$3{,}57\times10^{-07}$
miR-133b	11,31	-2,69	0,54	-5,00	0,16	0,07 – 0,30	$1{,}70\times10^{-06}$

Nota: RR - Risco Relativo; IC - Intervalo de confiança; FDR-BH - *False Discovery Rate* de Benjamini e Hochberg.

Tabela 18 – Estimativas dos coeficientes do modelo Binomial Negativa da expressão dos DEmiRNAs em amostras de tecido tumoral primário em comparação com os tecidos normais adjacentes ao tumor dos pacientes com câncer da laringe.

miRNAs	Estimativas		Erro-Padrão ($\hat{\beta}_1$)	Teste Wald	Exp($\hat{\beta}_1$)	IC$_{95\%}$(RR)	FDR-BH
	Normal ($\hat{\beta}_0$)	Tumoral ($\hat{\beta}_1$)			RR		
Superexpresso (↑)							
miR-196b-5p	6,52	3,36	0,34	9,83	10,27	6,27 – 15,93	$5{,}02 \times 10^{-20}$
miR-196a-5p	3,73	4,73	0,53	8,90	26,49	12,08 – 51,37	$1{,}12 \times 10^{-16}$
miR-615-3p	-4,02	7,26	0,72	7,56	153,38	25,66 – 5567,35	$2{,}79 \times 10^{-12}$
miR-4652-5p	1,21	3,45	0,45	7,54	10,90	5,82 – 19,98	$3{,}04 \times 10^{-12}$
miR-455-3p	8,12	2,48	0,34	7,30	5,60	3,42 – 8,66	$1{,}60 \times 10^{-11}$
miR-767-5p	1,76	6,27	0,90	6,98	77,32	18,34 – 229,02	$1{,}32 \times 10^{-10}$
miR-105-5p	2,23	6,45	0,89	6,77	87,23	19,44 – 264,22	$5{,}06 \times 10^{-10}$
miR-301a-3p	2,97	2,33	0,34	6,75	5,05	3,10 – 7,92	$5{,}34 \times 10^{-10}$
miR-193b-3p	7,26	2,00	0,31	6,53	4,01	2,59 – 5,94	$2{,}01 \times 10^{-09}$
miR-9-5p	10,64	4,16	0,64	6,47	17,93	6,67 – 39,30	$2{,}77 \times 10^{-09}$
miR-503-5p	2,00	2,52	0,40	6,21	5,74	3,27 – 9,73	$1{,}27 \times 10^{-08}$
miR-1910-5p	-1,79	4,24	0,68	6,06	18,88	6,63 – 72,80	$2{,}55 \times 10^{-08}$
miR-210-3p	9,93	2,26	0,38	6,03	4,80	2,77 – 7,75	$2{,}89 \times 10^{-08}$
miR-301a-5p	1,96	2,25	0,43	5,10	4,75	2,54 – 8,45	$3{,}73 \times 10^{-06}$
miR-1269a	4,91	3,85	0,76	5,08	14,47	4,38 – 36,21	$4{,}05 \times 10^{-06}$
miR-301b-3p	-0,50	2,97	0,63	4,97	7,81	3,38 – 19,31	$6{,}84 \times 10^{-06}$
miR-31-5p	5,76	2,59	0,54	4,81	6,02	2,69 – 11,75	$1{,}52 \times 10^{-05}$
miR-3662	0,20	2,62	0,59	4,67	6,13	2,70 – 13,72	$2{,}74 \times 10^{-05}$
miR-1293	1,72	2,48	0,53	4,64	5,57	2,57 – 11,09	$3{,}04 \times 10^{-05}$
miR-509-3p	5,05	3,44	0,77	4,50	10,88	2,57 – 31,42	$5{,}47 \times 10^{-05}$
miR-4326	2,31	2,08	0,48	4,33	4,24	2,08 – 7,95	$1{,}09 \times 10^{-04}$
miR-183-3p	0,60	2,07	0,48	4,27	4,20	2,16 – 8,22	$1{,}09 \times 10^{-04}$
miR-9-3p	2,52	2,61	0,63	4,09	6,09	2,33 – 13,46	$2{,}74 \times 10^{-04}$
miR-3170	-1,18	2,82	0,66	4,05	7,05	2,74 – 21,29	$3{,}08 \times 10^{-04}$
miR-6499-3p	1,71	2,56	0,57	3,82	5,90	2,32 – 13,07	$7{,}29 \times 10^{-04}$
miR-4449	0,13	2,19	0,62	3,53	4,56	1,88 – 10,72	$1{,}81 \times 10^{-03}$
miR-1305	-1,49	2,99	0,82	3,38	7,95	2,36 – 30,07	$2{,}95 \times 10^{-04}$
miR-508-3p	6,11	3,81	0,76	3,10	14,01	3,13 – 41,34	$6{,}93 \times 10^{-04}$
miR-514a-3p	4,64	4,39	0,77	3,09	20,98	4,60 – 62,77	$1{,}07 \times 10^{-04}$
miR-184	3,12	2,45	0,72	2,95	5,45	1,69 – 13,58	$1{,}07 \times 10^{-02}$
Subexpresso (↓)							
miR-338-5p	6,33	-2,43	0,27	-8,91	0,19	0,13 – 0,27	$1{,}12 \times 10^{-16}$
miR-139-3p	8,45	-2,83	0,36	-7,91	0,14	0,08 – 0,22	$3{,}13 \times 10^{-13}$
miR-378c	7,42	-2,31	0,30	-7,66	0,20	0,13 – 0,30	$1{,}94 \times 10^{-12}$
miR-378d	4,13	-2,82	0,40	-7,03	0,14	0,08 – 0,24	$9{,}70 \times 10^{-11}$
miR-99a-5p	12,95	-2,12	0,31	-6,74	0,23	0,15 – 0,35	$5{,}62 \times 10^{-10}$
miR-29c-3p	15,23	-2,05	0,34	-6,11	0,24	0,15 – 0,37	$2{,}00 \times 10^{-08}$
miR-375	16,14	-3,38	0,61	-5,53	0,10	0,04 – 0,20	$4{,}95 \times 10^{-07}$
miR-144-3p	7,84	-2,37	0,46	-5,14	0,19	0,10 – 0,35	$3{,}19 \times 10^{-06}$
miR-486-5p	11,32	-2,02	0,40	-5,10	0,25	0,14 – 0,41	$3{,}73 \times 10^{-06}$
miR-6510-3p	8,37	-2,48	0,52	-4,76	0,18	0,08 – 0,35	$1{,}87 \times 10^{-05}$
miR-1-3p	13,65	-3,89	0,89	-4,39	0,07	0,02 – 0,19	$8{,}60 \times 10^{-05}$
miR-133a-3p	12,67	-3,39	0,87	-3,90	0,10	0,02 – 0,27	$5{,}46 \times 10^{-04}$
miR-206	13,94	-2,22	0,95	-2,35	0,21	0,05 – 0,65	$4{,}61 \times 10^{-02}$
miR-133b	10,84	-3,70	0,97	-2,35	0,08	0,02 – 0,24	$4{,}61 \times 10^{-02}$

Nota: RR - Risco Relativo; IC - Intervalo de confiança; FDR-BH - *False Discovery Rate* de Benjamini e Hochberg.

A.3 Identificação de DEmiRNAs com edgeR

Tabela 19 – Estimativas dos coeficientes do modelo pelo edgeR em amostras de tecido tumoral primário em comparação com os tecidos normais adjacentes ao tumor dos pacientes com câncer da laringe.

Oral						Laringe		
miRNAs	$\log_2(FC)$	FDR-BH	miRNAs	$\log_2(FC)$	FDR-BH	miRNAs	$\log_2(FC)$	FDR-BH
Superexpresso (↑)			Subexpresso (↓)			Superexpresso (↑)		
miR-455-3p	2,86	1,57E-30	miR-101-3p	-2,21	6,49E-71	miR-196b-5p	3,39	3,26E-12
miR-196b-5p	3,63	1,78E-24	miR-381-3p	-3,37	5,63E-66	miR-615-3p	5,68	5,28E-12
miR-615-3p	4,03	1,87E-24	miR-30a-3p	-2,77	1,78E-54	miR-4652-5p	3,37	1,08E-09
miR-1910-5p	4,04	1,08E-19	miR-30a-5p	-2,27	4,71E-46	miR-196a-5p	4,77	4,04E-09
miR-503-5p	2,69	1,86E-19	miR-375	-4,12	1,40E-45	miR-1910-5p	4,00	4,43E-09
miR-877-5p	2,34	2,10E-18	miR-378c	-2,57	2,06E-45	miR-455-3p	2,52	2,82E-08
miR-196a-5p	3,80	7,75E-18	miR-299-5p	-2,65	7,01E-45	miR-301a-3p	2,37	2,92E-08
miR-365a-5p	2,50	9,34E-17	miR-378a-5p	-2,01	1,19E-42	miR-503-5p	2,54	2,21E-07
miR-503-3p	2,91	3,09E-16	miR-99a-5p	-2,48	3,05E-40	miR-193b-3p	2,08	2,60E-07
miR-210-3p	2,79	8,00E-16	miR-29c-3p	-2,41	8,97E-40	miR-301b-3p	2,95	3,74E-06
miR-301a-5p	2,58	3,09E-15	miR-136-3p	-2,22	2,35E-38	miR-210-3p	2,28	5,05E-06
miR-224-5p	2,15	4,43E-15	miR-411-5p	-2,58	1,15E-36	miR-9-5p	4,19	1,89E-05
miR-31-5p	3,22	1,43E-14	miR-139-3p	-2,22	2,30E-32	miR-301a-5p	2,20	2,31E-05
miR-937-3p	2,68	1,01E-13	miR-195-5p	-2,06	7,87E-32	miR-767-5p	6,33	3,04E-05
miR-4652-5p	3,68	1,43E-13	miR-378a-3p	-2,18	5,91E-31	miR-105-5p	6,47	4,54E-05
miR-7705	2,13	1,58E-12	miR-376c-3p	-2,11	2,46E-30	miR-3662	2,67	5,83E-05
miR-4664-3p	2,20	1,32E-10	miR-30c-2-3p	-2,29	4,25E-30	miR-183-3p	2,01	8,19E-05
miR-1293	2,59	2,71E-10	let-7c-5p	-2,22	1,98E-29	miR-1293	2,44	0,000237
miR-1305	2,62	2,77E-10	miR-378d	-2,59	3,88E-28	miR-7-5p	2,03	0,000332
miR-301b-3p	2,27	5,62E-10	miR-410-3p	-2,17	5,37E-26	miR-31-5p	2,64	0,000351
miR-31-3p	2,43	6,30E-10	miR-411-3p	-2,22	1,18E-24	miR-4326	2,14	0,000559
miR-155-3p	2,23	3,10E-08	miR-379-3p	-2,30	2,29E-24	miR-1269a	3,88	0,000661
miR-105-5p	5,59	3,86E-08	miR-34c-3p	-2,32	1,75E-23	miR-592	2,70	0,001769
miR-767-5p	5,29	3,35E-07	miR-486-5p	-2,03	5,03E-21	miR-6499-3p	2,52	0,001902
miR-1269a	4,28	1,65E-06	miR-34c-5p	-2,09	8,20E-18	miR-9-3p	2,59	0,00202
miR-512-3p	3,26	1,45E-05	miR-381-5p	-2,19	8,84E-14	miR-514a-3p	4,47	0,00216
miR-9-5p	2,17	3,55E-05	miR-1-3p	-2,87	7,13E-13	miR-4449	2,18	0,002303
			miR-6510-3p	-2,10	7,68E-12	miR-508-3p	3,87	0,005615
			miR-204-5p	-2,33	8,35E-12	miR-509-3p	3,51	0,008841
			miR-133a-3p	-2,51	9,49E-10	miR-184	2,48	0,013866
			miR-133b	-2,47	1,93E-08	Subexpresso (↓)		
						miR-338-5p	-2,34	2,27E-21
						miR-139-3p	-2,73	1,91E-18
						miR-378c	-2,19	4,74E-16
						miR-99a-5p	-2,02	3,26E-12
						miR-375	-3,28	1,10E-10
						miR-144-3p	-2,29	2,40E-08
						miR-34b-5p	-2,85	5,24E-08
						miR-1-3p	-3,80	1,01E-07
						miR-6510-3p	-2,38	3,89E-07
						miR-133a-3p	-3,31	5,05E-06
						miR-133b	-3,61	6,28E-06
						miR-206	-2,13	0,015439

Nota: RR - Risco Relativo; IC - Intervalo de confiança; FDR-BH - *False Discovery Rate* de Benjamini e Hochberg.

A.4 Identificação de DEmiRNAs pelo WMW: oral e laringe

Tabela 20 – Estatísticas descritivas da expressão dos miRNAs identificados na cavidade oral comparando os tecidos usando o teste de Wilcoxon-Mann-Whitney.

miRNAs	Normal adjacente				Tumor primário				FC	$\log_2$ (FC)	FDR-BH
	Mín.	Med.	Média	Máx.	Mín.	Med.	Média	Máx.			
Superexpresso (↑)											
miR-455-3p	14,77	50,48	62,01	178,76	40,42	371,07	449,65	1782,44	7,25	2,86	$1{,}58\times10^{-14}$
miR-196b-5p	0,82	9,11	11,88	57,50	6,34	98,00	146,89	1173,80	12,36	3,63	$8{,}46\times10^{-14}$
miR-615-3p	0,00	0,00	0,12	1,06	0,00	1,97	2,53	19,39	20,66	4,37	$1{,}78\times10^{-13}$
miR-503-5p	0,00	1,11	1,47	5,49	0,00	6,79	9,77	80,99	6,65	2,73	$6{,}08\times10^{-13}$
miR-1910-5p	0,00	0,00	0,08	0,63	0,00	1,06	1,60	18,71	20,21	4,34	$1{,}50\times10^{-12}$
miR-877-5p	0,00	0,51	0,50	1,28	0,00	2,17	2,69	13,12	5,35	2,42	$3{,}27\times10^{-12}$
miR-196a-5p	0,26	1,69	5,59	41,69	0,46	55,48	77,83	743,52	13,92	3,80	$6{,}23\times10^{-12}$
miR-210-3p	4,66	75,67	188,54	622,87	15,51	946,44	1302,18	10161,32	6,91	2,79	$1{,}54\times10^{-11}$
miR-365a-5p	0,00	0,40	0,47	1,98	0,00	2,10	2,66	13,57	5,69	2,51	$3{,}33\times10^{-11}$
miR-301a-5p	0,00	0,44	0,70	2,67	0,00	3,34	4,46	27,16	6,37	2,67	$3{,}37\times10^{-11}$
miR-224-5p	3,11	37,41	38,73	147,35	6,80	137,42	171,91	925,66	4,44	2,15	$5{,}38\times10^{-11}$
miR-503-3p	0,00	0,14	0,17	0,88	0,00	1,13	1,51	14,82	8,73	3,13	$5{,}52\times10^{-11}$
miR-31-5p	0,00	9,11	14,01	73,51	0,24	96,94	130,79	821,22	9,33	3,22	$2{,}18\times10^{-10}$
miR-937-3p	0,00	0,68	0,80	3,79	0,00	2,94	5,00	52,47	6,23	2,64	$2{,}95\times10^{-10}$
miR-4652-5p	0,00	0,00	0,28	2,51	0,00	2,36	3,74	45,40	13,38	3,74	$4{,}68\times10^{-10}$
miR-1293	0,00	0,56	1,45	12,29	0,00	4,89	8,81	103,93	6,08	2,60	$2{,}79\times10^{-09}$
miR-7705	0,00	0,22	0,29	1,10	0,00	1,12	1,37	6,81	4,74	2,25	$3{,}11\times10^{-09}$
miR-31-3p	0,15	7,52	15,57	133,20	0,24	64,90	84,25	414,19	5,41	2,44	$2{,}06\times10^{-08}$
miR-4664-3p	0,00	0,00	0,23	1,74	0,00	0,79	1,07	7,33	4,57	2,19	$2{,}47\times10^{-08}$
miR-130a-5p	0,00	0,21	0,30	0,88	0,00	0,99	1,23	7,00	4,15	2,05	$2{,}74\times10^{-08}$
miR-1305	0,00	0,00	0,14	1,32	0,00	0,71	1,05	9,97	7,32	2,87	$6{,}24\times10^{-08}$
miR-301b-3p	0,00	0,15	0,17	0,52	0,00	0,59	0,92	11,18	5,36	2,42	$1{,}37\times10^{-07}$
miR-1254	0,00	0,19	0,25	0,73	0,00	0,74	0,99	5,46	4,01	2,00	$4{,}92\times10^{-07}$
miR-155-3p	0,00	0,00	0,16	0,76	0,00	0,50	0,85	9,98	5,24	2,39	$1{,}09\times10^{-06}$
miR-105-5p	0,00	0,29	1,45	12,38	0,00	1,27	71,88	1190,82	49,44	5,63	$2{,}60\times10^{-03}$
miR-767-5p	0,00	0,27	1,27	12,99	0,00	0,75	51,03	692,51	40,28	5,33	$4{,}80\times10^{-03}$
miR-512-3p	0,00	0,19	0,41	1,61	0,00	0,60	4,16	861,30	10,24	3,36	$5{,}56\times10^{-03}$
miR-1269a	0,00	1,11	5,65	79,51	0,00	6,80	110,03	2407,94	19,49	4,28	$5{,}97\times10^{-03}$
Subexpresso (↓)											
miR-101-3p	7781,50	31068,69	30678,11	56406,34	1021,51	5774,33	6611,94	25141,23	0,22	-2,21	$1{,}28\times10^{-14}$
miR-99a-5p	419,91	1395,99	1591,89	4628,09	5,72	221,22	284,90	2774,48	0,18	-2,48	$1{,}58\times10^{-14}$
miR-30a-5p	8689,88	22293,01	92445,29	67585,75	1333,25	4488,50	6079,46	72166,75	0,21	-2,27	$5{,}87\times10^{-14}$
let-7c-5p	1601,90	6137,70	6879,89	20886,35	81,00	1019,04	1480,19	13203,69	0,22	-2,22	$8{,}46\times10^{-14}$
miR-29c-3p	1560,83	5950,76	7883,37	22452,83	113,75	959,61	1479,94	16276,65	0,19	-2,41	$1{,}10\times10^{-13}$
miR-136-3p	12,17	33,54	47,66	291,51	0,00	7,90	10,15	56,61	0,21	-2,23	$1{,}12\times10^{-13}$
miR-195-5p	26,99	86,35	101,41	253,82	2,19	18,36	24,32	361,15	0,24	-2,06	$1{,}21\times10^{-13}$
miR-204-5p	3,51	13,08	16,10	46,50	0,00	0,92	3,18	125,37	0,20	-2,34	$1{,}36\times10^{-13}$
miR-30a-3p	1345,21	9540,52	12052,53	32385,05	193,23	1106,94	1763,19	27032,68	0,15	-2,77	$1{,}44\times10^{-13}$
miR-139-3p	16,90	49,09	65,42	180,75	0,62	9,45	14,05	92,88	0,21	-2,22	$2{,}92\times10^{-13}$
miR-30c-2-3p	12,73	64,72	85,46	386,86	0,77	11,22	17,49	439,59	0,20	-2,29	$1{,}06\times10^{-12}$
miR-410-3p	9,88	45,09	67,39	485,77	0,00	10,96	14,94	104,24	0,22	-2,17	$3{,}93\times10^{-12}$
miR-411-5p	7,02	32,69	64,45	622,51	0,26	7,31	10,74	136,52	0,17	-2,59	$5{,}91\times10^{-12}$
miR-378c	5,71	25,32	55,29	308,06	0,83	6,45	9,29	139,46	0,17	-2,57	$1{,}27\times10^{-11}$
miR-375	36,25	5256,62	16419,95	94479,81	6,34	228,03	944,60	37560,44	0,06	-4,12	$3{,}13\times10^{-11}$
miR-381-3p	41,04	261,85	891,61	11104,69	6,76	60,06	86,18	1447,18	0,10	-3,37	$3{,}63\times10^{-11}$
miR-376c-3p	3,48	20,83	29,74	225,44	0,23	5,35	6,87	47,53	0,23	-2,11	$4{,}20\times10^{-11}$
miR-378a-3p	682,48	2493,12	4913,28	26322,94	139,97	735,52	1085,49	21467,75	0,22	-2,18	$1{,}83\times10^{-10}$
miR-379-3p	0,00	2,60	4,77	41,35	0,00	0,63	0,94	4,91	0,20	-2,34	$9{,}08\times10^{-09}$
miR-299-5p	3,16	17,94	44,97	295,43	0,22	5,21	7,16	51,85	0,16	-2,65	$2{,}00\times10^{-08}$
miR-378d	0,00	3,03	4,96	25,18	0,00	0,50	0,79	14,44	0,16	-2,65	$4{,}92\times10^{-08}$
miR-378a-5p	76,37	243,09	527,37	2807,07	29,38	112,76	131,23	1251,18	0,25	-2,01	$7{,}74\times10^{-08}$
miR-486-5p	80,01	438,37	863,89	4210,33	14,44	131,18	211,17	2720,51	0,24	-2,03	$4{,}01\times10^{-07}$
miR-411-3p	0,22	4,43	7,71	96,44	0,00	1,27	1,63	11,58	0,21	-2,24	$9{,}85\times10^{-07}$
miR-6510-3p	0,31	32,79	43,08	248,19	0,00	3,89	10,01	424,83	0,23	-2,11	$2{,}19\times10^{-06}$
miR-1-3p	7,24	1404,12	3557,30	26453,68	0,25	57,07	485,55	12885,46	0,14	-2,87	$2{,}70\times10^{-05}$
miR-133b	0,88	174,88	513,38	3750,02	0,00	12,85	92,59	1110,37	0,18	-2,47	$1{,}65\times10^{-04}$
miR-34b-5p	0,00	0,32	7,85	192,21	0,00	1,18	1,95	18,58	0,25	-2,01	$2{,}80\times10^{-04}$
miR-133a-3p	5,18	812,53	2414,14	16852,44	0,00	51,15	423,74	10314,10	0,18	-2,51	$4{,}31\times10^{-04}$
miR-381-5p	0,00	1,18	3,75	37,07	0,00	0,39	0,78	13,99	0,21	-2,26	$5{,}30\times10^{-04}$
miR-34c-3p	1,47	4,47	67,00	1444,65	0,25	9,65	13,42	74,61	0,20	-2,32	$1{,}23\times10^{-03}$
miR-34c-5p	1,14	8,92	81,62	1723,30	0,00	12,52	19,12	202,23	0,23	-2,09	$3{,}66\times10^{-02}$

Nota: Mín. - Mínimo; Med. - Mediana; Máx. - Máximo; FC - *Fold Change*; FDR-BH - *False Discovery Rate* de Benjamini e Hochberg.

Tabela 21 – Estatísticas descritivas da expressão dos miRNAs identificados na laringe comparando os tecidos (tumor primário vs. normal adjacente) usando o teste de Wilcoxon-Mann-Whitney.

miRNAs	Normal adjacente				Tumor primário				FC	$\log_2$ (FC)	FDR-BH
	Mín.	Med.	Média	Máx.	Mín.	Med.	Média	Máx.			
Superexpresso (↑)											
miR-196b-5p	5,18	20,51	21,41	52,64	4,35	181,70	224,26	1037,82	10,47	3,39	$1{,}89\times10^{-05}$
miR-615-3p	0,00	0,00	0,01	0,18	0,00	1,55	2,25	14,92	151,04	7,24	$2{,}09\times10^{-05}$
miR-4652-5p	0,00	0,34	0,55	1,79	0,00	4,47	6,08	31,31	11,11	3,47	$2{,}09\times10^{-05}$
miR-301a-3p	0,40	1,30	1,80	4,29	0,81	6,85	9,48	41,96	5,28	2,40	$3{,}15\times10^{-05}$
miR-455-3p	21,43	48,95	64,44	173,81	19,73	288,77	368,77	1288,57	5,72	2,52	$3{,}15\times10^{-05}$
miR-193b-3p	20,93	29,72	34,96	70,14	10,93	120,18	148,03	644,15	4,23	2,08	$3{,}81\times10^{-05}$
miR-503-5p	0,00	0,55	0,91	2,41	0,56	3,43	5,58	35,97	6,13	2,62	$3{,}81\times10^{-05}$
miR-1910-5p	0,00	0,00	0,07	0,40	0,00	1,03	1,29	6,81	19,76	4,30	$5{,}25\times10^{-05}$
miR-196a-5p	0,29	1,74	3,05	9,48	0,00	57,80	83,80	443,81	27,44	4,78	$1{,}24\times10^{-04}$
miR-210-3p	21,97	126,00	228,70	1109,08	33,32	800,34	1112,12	3780,62	4,86	2,28	$1{,}38\times10^{-04}$
miR-301b-3p	0,00	0,00	0,17	0,71	0,00	0,93	1,31	13,63	7,88	2,98	$2{,}12\times10^{-04}$
miR-301a-5p	0,21	0,57	0,90	3,06	0,00	2,88	4,35	26,75	4,81	2,27	$2{,}61\times10^{-04}$
miR-3662	0,00	0,20	0,26	0,57	0,00	0,97	1,68	11,16	6,36	2,67	$7{,}66\times10^{-04}$
miR-183-3p	0,00	0,32	0,35	0,85	0,00	1,12	1,51	11,01	4,31	2,11	$8{,}81\times10^{-04}$
miR-4326	0,00	0,63	1,14	5,71	0,00	4,28	5,11	22,85	4,48	2,16	$1{,}51\times10^{-03}$
miR-1293	0,00	0,82	0,76	1,69	0,00	2,48	4,34	34,10	5,72	2,52	$1{,}55\times10^{-03}$
miR-9-5p	88,95	310,56	370,30	991,59	114,72	1040,08	6745,03	123078,78	18,22	4,19	$1{,}82\times10^{-03}$
miR-592	0,18	0,78	0,94	1,87	0,00	2,04	6,18	387,41	6,59	2,72	$5{,}24\times10^{-03}$
miR-31-5p	0,54	8,69	12,56	45,42	0,45	43,09	78,64	771,85	6,26	2,65	$6{,}92\times10^{-03}$
miR-4449	0,00	0,18	0,25	0,79	0,00	0,69	1,20	11,53	4,85	2,28	$9{,}53\times10^{-03}$
miR-767-5p	0,00	0,49	0,77	3,95	0,00	5,72	62,84	1363,48	81,16	6,34	$1{,}25\times10^{-02}$
miR-1269a	0,00	1,83	7,15	33,01	0,00	40,80	105,65	937,86	14,78	3,89	$1{,}30\times10^{-02}$
miR-6499-3p	0,00	0,54	0,75	2,29	0,00	1,50	4,50	90,36	5,98	2,58	$1{,}51\times10^{-02}$
Subexpresso (↓)											
miR-99a-5p	812,32	1691,28	1797,74	2594,52	30,62	327,87	441,92	2098,61	0,25	-2,02	$1{,}89\times10^{-05}$
miR-139-3p	13,14	79,01	79,25	149,22	0,00	8,61	11,98	87,31	0,15	-2,73	$2{,}09\times10^{-05}$
miR-6510-3p	12,57	76,37	75,49	189,11	0,00	6,42	14,43	226,47	0,19	-2,39	$3{,}67\times10^{-05}$
miR-375	1286,13	6312,88	16576,73	89935,96	10,71	459,54	1706,51	19308,01	0,10	-3,28	$4{,}37\times10^{-05}$
miR-338-5p	2,29	15,56	18,31	41,47	0,00	2,96	3,60	17,00	0,20	-2,35	$4{,}37\times10^{-05}$
miR-378c	5,93	15,25	38,28	201,82	1,11	6,89	8,37	31,50	0,22	-2,19	$4{,}81\times10^{-04}$
miR-1-3p	6,22	36,36	2849,71	17666,97	0,61	5,83	204,85	4824,76	0,07	-3,80	$3{,}82\times10^{-03}$
miR-144-3p	0,29	26,83	50,65	232,54	0,28	6,05	10,38	77,00	0,20	-2,29	$5{,}11\times10^{-03}$
miR-133a-3p	2,83	22,24	1458,48	9591,29	0,00	5,22	147,51	2881,53	0,10	-3,31	$8{,}75\times10^{-03}$
miR-133b	0,57	5,00	408,91	3130,60	0,00	0,87	33,47	560,16	0,08	-3,61	$9{,}97\times10^{-03}$

Nota: Mín. - Mínimo; Med. - Mediana; Máx. - Máximo; FC - *Fold Change*; FDR-BH - *False Discovery Rate* de Benjamini e Hochberg.

A.5 Identificação de DEmRNAs com DESeq2: oral e laringe

Tabela 22 – Estimativas dos coeficientes do modelo pelo Deseq2 em amostras de tecido tumoral primário em comparação com os tecidos normais adjacentes ao tumor dos pacientes com câncer da laringe.

Oral						Laringe					
mRNAs	$\hat{\beta}_1$	FDR-BH	mRNAs	$\hat{\beta}_1$	FDR-BH	mRNAs	$\hat{\beta}_1$	FDR-BH	mRNAs	$\hat{\beta}_1$	FDR-BH
Superexpresso (↑)			Subexpresso (↓)			Superexpresso (↑)			Subexpresso (↓)		
LAMC2	4,35	1,8E-61	GPD1L	-2,85	1,2E-49	ZIC5	6,95	1,33E-38	EMP1	-3,90	3,8E-29
BMP1	2,02	8,2E-43	SH3BGRL2	-3,89	3,8E-38	COL4A1	3,49	9,52E-33	GPD1L	-2,61	2,72E-23
HMGA2	4,89	1,1E-42	RRAGD	-2,79	8,6E-30	STC2	3,92	1,21E-19	CPEB3	-2,08	7,76E-17
HOXA10	4,15	2,1E-41	TNFRSF11A	-2,33	2,0E-29	FADS1	3,12	1,88E-19	SORBS2	-3,15	6,37E-13
COL10A1	5,06	7,8E-38	SORBS2	-3,18	2,4E-29	ADAM12	4,24	2,75E-19	ZNF185	-2,49	3,52E-12
IGF2BP2	2,53	2,5E-37	NR3C2	-3,81	8,2E-29	NETO2	2,17	3,75E-19	EHD3	-2,02	4,65E-11
CEP55	2,09	4,6E-35	SHROOM3	-2,27	1,8E-27	CD276	2,04	2,53E-17	CFL2	-2,15	3,5E-10
COL12A1	2,99	2,1E-34	SLC25A34	-3,46	3,7E-27	FBN2	4,93	2,46E-15	PKIA	-3,07	4,45E-09
DNMT3B	2,43	3,0E-34	TSPAN12	-2,76	3,8E-27	EIF5A2	2,15	9,48E-15	TTC9	-2,51	7,56E-09
SERPINH1	2,12	3,8E-33	ITM2A	-2,62	3,7E-26	RAB3B	5,17	2,96E-14	ALPK3	-2,64	3,03E-08
HOXA11	4,98	3,7E-32	EMP1	-2,37	6,3E-26	COL7A1	2,31	4,9E-14	NR3C2	-2,41	2,12E-07
C1QTNF6	2,55	2,2E-31	ENPP4	-3,02	9,4E-24	COL1A1	3,62	1,78E-13	CLIC5	-2,77	1,3E-06
HOXC8	3,72	1,2E-30	NBEA	-3,15	1,5E-22	SOX11	4,42	4,23E-13	KBTBD12	-4,12	1,71E-06
IGF2BP1	5,71	9,2E-30	TGFBR3	-2,56	5,6E-22	HOXA10	2,83	4,73E-13	RBFOX1	-3,75	3,67E-06
COL5A2	3,05	1,5E-28	RNF150	-3,16	7,0E-21	COL3A1	3,45	2,92E-12	TRDN	-4,85	4,11E-06
SERPINE1	3,06	3,3E-27	INMT	-2,59	8,1E-21	LHX2	3,75	7,41E-12	PLEKHB1	-2,03	6,58E-05
HOXB7	3,67	4,0E-27	CFL2	-2,07	2,5E-20	DNMT3B	2,06	1,06E-11	XIRP2	-5,29	0,000115
ADAM12	3,21	2,6E-26	CLIC5	-3,45	3,3E-20	NRIP3	2,43	1,12E-11	CFTR	-2,33	0,000683
COL4A1	2,16	4,3E-26	AR	-2,69	6,3E-19	PXDN	2,54	2,02E-11	PAX1	-3,31	0,000761
COL7A1	2,23	7,0E-26	PPARGC1A	-3,31	2,8E-18	ONECUT2	3,50	2,84E-11	PRUNE2	-2,21	0,000796
ITGA6	2,10	9,2E-26	USP13	-2,05	4,2E-18	COL1A2	3,12	5,64E-11	ADCY2	-2,11	0,001693
E2F7	2,03	4,5E-25	MITF	-2,15	1,5E-17	C1QTNF6	2,40	3,15E-10	PLCXD3	-2,08	0,002758
COL5A3	2,34	7,2E-24	NTRK3	-3,38	2,7E-17	COL5A2	2,81	6,34E-10	SYNPO2L	-2,02	0,009688
NRIP3	2,36	4,8E-23	PKIA	-3,14	3,2E-17	TBX2	2,17	2,49E-09	BMP3	-2,04	0,027177
STC2	2,58	5,5E-23	DMD	-2,35	9,2E-17	PAX3	5,51	4,14E-09			
HTR1D	3,62	1,2E-22	PDE6A	-2,86	9,3E-17	LAMC2	2,78	9,48E-09			
ONECUT2	3,02	2,9E-22	ETV1	-2,35	1,3E-16	TRPA1	2,92	1,3E-08			
FBN2	3,48	3,8E-22	ID4	-2,05	5,0E-16	CDK14	2,49	5,75E-08			
FOXL2	3,75	4,0E-22	LPL	-2,65	1,4E-15	FOXL2	3,72	1,57E-07			
PMEPA1	2,05	6,0E-21	AMOT	-3,12	2,4E-15	SERPINE1	2,51	3,05E-07			
COL1A1	2,97	2,7E-20	SYNGR1	-2,00	1,0E-14	SPARC	2,06	3,08E-07			
FN1	3,16	2,5E-19	LYVE1	-2,11	3,4E-14	RIMS2	2,88	4,31E-07			
IGF2BP3	3,04	2,6E-17	GFRA1	-2,66	6,9E-13	SLC2A14	3,37	9,22E-07			
COL3A1	2,77	3,2E-17	RASD1	-2,05	1,9E-12	DCC	2,52	0,000164			
SOST	3,89	4,9E-16	TMEM100	-2,39	3,8E-12	ANO4	2,41	0,000442			
SOX11	2,80	9,8E-16	ENPP5	-2,89	4,4E-12	ELAVL2	2,49	0,000897			
COL1A2	2,50	4,9E-15	LIMCH1	-2,11	5,6E-12	CDH2	2,01	0,000908			
ZIC5	3,00	9,8E-15	SEMA3E	-2,79	5,6E-12	OLFM4	2,34	0,005512			
EN1	2,50	1,4E-13	NR4A1	-2,10	9,7E-12	BEND4	2,29	0,009898			
CELSR3	2,11	3,2E-13	LONRF2	-3,03	2,1E-11						
CXCL11	3,39	5,8E-13	PTX3	-2,48	7,8E-10						
SULF1	2,09	1,1E-12	RBFOX1	-3,30	2,3E-09						
HOXB6	2,56	1,6E-12	ANK1	-2,09	2,4E-09						
IFIT2	2,17	1,8E-12	TRDN	-3,79	2,5E-08						
EN2	2,36	2,7E-11	FAM135B	-2,09	3,2E-08						
DISP2	2,12	5,8E-11	FLNC	-2,36	4,2E-08						
ONECUT1	3,31	5,0E-10	PLD5	-2,63	3,4E-07						
RAB3B	2,42	7,2E-10	PAX1	-2,70	5,0E-07						
ELAVL2	2,29	1,9E-08	EPHA7	-2,47	7,5E-06						
HOXA9	2,17	2,0E-07	XKR4	-2,04	1,2E-05						
ADD2	2,08	5,8E-06	PDE11A	-2,12	5,7E-05						
			TLX1	-2,34	1,2E-03						

Nota: RR - Risco Relativo; IC - Intervalo de confiança; FDR-BH - *False Discovery Rate* de Benjamini e Hochberg.

A.6 Identificação dos DEmiRNAs nos estadiamento iniciais e avançados do câncer da cavidade oral e laringe

Tabela 23 – DEmiRNAs identificados com o teste WMW nos estadiamento iniciais e avançados e áreas sob a curva (AUC), especificidade (TNR) e sensibilidade (TPR) na capacidade de discriminação do tumor primário (T) e tecido normal adjacente ao tumor (NAT) na cavidade oral.

miRNAs	Cavidade oral													
	estadiamento inicial ($n = 84$)							estadiamento avançado ($n = 237$)						
	Mediana		Valor p	AUC ($IC_{95}\%$)	*Cuttof*	TNR	TPR	Mediana		Valor p	AUC ($IC_{95}\%$)	*Cuttof*	TNR	TPR
	T	NAT						T	NAT					
miR-455-3p	10,38	7,43	$1{,}5\times10^{-06}$	95,3 (90,0 - 100,0)	9,1	90,9	90,4	10,67	7,86	$3{,}9\times10^{-11}$	98,1 (96,6 - 99,7)	9,1	100,0	92,7
miR-196b-5p	8,41	5,31	$2{,}2\times10^{-06}$	94,5 (89,6 - 99,4)	7,0	100,0	83,6	8,81	5,30	$1{,}4\times10^{-10}$	96,8 (93,3 - 100,0)	6,6	94,1	90,5
miR-196a-5p	7,61	2,86	$6{,}6\times10^{-05}$	87,5 (78,4 - 96,7)	6,7	100,0	64,4	7,93	3,23	$6{,}1\times10^{-09}$	92,4 (87,3 - 97,4)	5,9	94,1	81,8
miR-503-5p	4,79	2,74	$2{,}5\times10^{-06}$	94,3 (87,6 - 100,0)	4,0	90,9	90,4	4,96	2,47	$2{,}7\times10^{-09}$	93,3 (88,2 -98,4)	3,6	88,2	86,8
miR-210-3p	11,60	7,58	$2{,}1\times10^{-05}$	90,0 (80,0 - 100,0)	8,3	72,7	98,6	12,00	8,73	$3{,}0\times10^{-08}$	90,3 (84,8 - 95,9)	10,9	88,2	79,5
miR-31-5p	8,47	4,57	$4{,}7\times10^{-05}$	88,3 (79,6 - 97,0)	7,1	90,9	78,1	8,75	5,53	$2{,}4\times10^{-07}$	87,6 (82,3 - 93,0)	6,9	94,1	80,9
miR-615-3p	3,33	0,00	$1{,}00\times10^{-05}$	91,4 (83,9 - 98,9)	0,8	81,8	91,8	3,22	0,00	$3{,}3\times10^{-10}$	95,8 (92,8 - 98,8)	1,4	94,1	92,7
miR-224-5p	9,00	6,46	$5{,}3\times10^{-05}$	88,0 (77,3 - 98,8)	7,5	72,7	89,0	9,24	7,28	$3{,}2\times10^{-08}$	90,3 (85,6 - 94,9)	8,4	100,0	74,1
miR-877-5p	3,31	1,38	$1{,}9\times10^{-06}$	94,8 (89,8 - 99,7)	2,0	90,9	90,4	3,38	1,89	$2{,}6\times10^{-08}$	90,6 (86,2 - 94,9)	2,6	100,0	76,8
miR-1910-5p	2,09	0,00	$3{,}0\times10^{-05}$	89,2 (80,6 - 97,9)	1,1	90,9	80,8	2,45	0,00	$1{,}2\times10^{-09}$	94,2 (91,1 - 97,3)	0,9	94,1	88,2
miR-4652-5p	3,19	0,00	$4{,}6\times10^{-04}$	82,8 (71,5 - 94,1)	2,1	90,9	76,7	3,55	0,00	$4{,}4\times10^{-08}$	89,8 (85,8 -93,7)	1,9	100,0	80,0
miR-301a-5p	3,64	1,10	$2{,}8\times10^{-05}$	89,4 (80,2 - 98,7)	2,1	81,8	87,7	3,99	2,05	$1{,}1\times10^{-07}$	88,7 (82,6 - 94,7)	3,0	94,1	74,1
miR-365a-5p	3,26	1,51	$7{,}2\times10^{-06}$	92,2 (85,9 - 98,6)	1,9	90,9	91,8	3,35	1,49	$1{,}1\times10^{-07}$	88,7 (83,0 - 94,4)	2,4	94,1	78,2
miR-937-3p	3,77	2,35	0,0010	81,1 (69,8 - 92,3)	2,8	81,8	76,7	3,81	1,84	$1{,}8\times10^{-08}$	91,0 (85,9 - 96,1)	2,8	94,1	76,8
miR-105-5p	1,73	1,03	0,0430	69,0 (53,8 - 84,2)	1,2	72,7	69,9	2,93	1,70	0,0340	65,4 (55,2 - 75,7)	5,3	100,0	41,8
miR-503-3p	2,71	0,00	$1{,}7\times10^{-06}$	95,0 (90,4 - 99,6)	1,2	90,9	93,2	2,41	0,89	$6{,}7\times10^{-07}$	86,2 (80,3 - 92,1)	1,5	94,1	75,5
miR-1269a	4,77	1,93	0,0870	66,1 (50,3 - 82,0)	6,4	90,9	41,1	5,04	2,62	0,0240	66,5 (58,3 -74,6)	5,4	100,0	49,5
miR-31-3p	7,97	4,23	$3{,}3\times10^{-05}$	89,0 (79,7 - 98,4)	5,8	81,8	86,3	8,16	5,59	$1{,}7\times10^{-05}$	81,4 (72,8 - 89,9)	6,9	94,1	73,2
miR-1293	4,44	1,10	$8{,}9\times10^{-05}$	86,9 (75,9 - 97,9)	2,9	81,8	83,6	4,39	2,13	$2{,}0\times10^{-06}$	84,6 (74,7 - 94,6)	3,4	88,2	75,5
miR-767-5p	1,82	0,00	0,0900	65,5 (49,2 - 81,8)	0,8	63,6	67,1	2,37	1,56	0,0400	64,9 (54,6 - 75,3)	5,0	100,0	41,4
miR-7705	2,65	0,95	$3{,}7\times10^{-05}$	88,8 (81,0 - 96, 6)	2,3	100,0	68,5	2,45	1,07	$3{,}1\times10^{-06}$	84,0 (76,8 - 91,2)	1,7	88,2	74,1
miR-4664-3p	2,05	0,00	0,0033	77,6 (60,7 - 94,5)	1,2	81,8	78,1	2,17	0,00	$2{,}6\times10^{-07}$	87,5 (82,1 - 92,9)	1,6	100,0	71,8
miR-1305	1,91	0,00	$1{,}9\times10^{-04}$	85,0 (75,5 - 94,5)	1,7	100,0	60,3	2,04	0,00	$2{,}5\times10^{-05}$	80,5 (72,3 - 88,8)	0,9	88,2	76,4
miR-301b-3p	1,68	0,00	$7{,}1\times10^{-05}$	87,2 (80,1 - 94,2)	1,1	100,0	74,0	1,90	0,96	0,0001	77,7 (69,9 - 85,5)	1,7	100,0	56,8
miR-4326	4,25	2,13	$1{,}1\times10^{-03}$	80,6 (71,2 - 89,9)	3,8	100,0	64,4	4,43	2,63	$6{,}9\times10^{-04}$	74,7 (65,2 - 84,3)	3,1	76,5	74,1
miR-155-3p	1,52	0,00	$4{,}1\times10^{-03}$	76,8 (62,8 - 90,9)	0,7	72,7	80,8	1,64	0,00	$5{,}7\times10^{-05}$	79,2 (70,0 - 88,5)	1,5	94,1	56,8
miR-9-5p	10,50	11,00	0,6100	54,9 (37,6 - 72,2)	10,7	72,7	53,4	10,82	10,38	0,3300	57,2 (45,5 - 68,8)	12,2	94,1	28,6
miR-101-3p	14,72	17,08	$1{,}1\times10^{-07}$	100 (100 - 100,0)	16,2	100,0	100,0	14,55	17,13	$1{,}1\times10^{-10}$	97,0 (93,2 - 100,0)	15,6	94,1	90,0
miR-381-3p	8,07	10,21	$3{,}1\times10^{-05}$	89,2 (77,8 - 100,0)	8,9	81,8	83,6	7,86	10,11	$5{,}2\times10^{-08}$	89,7 (81,6 - 97,7)	8,6	94,1	76,4
miR-30a-3p	12,25	15,65	$1{,}1\times10^{-06}$	95,9 (91,6 - 100,0)	13,9	90,9	90,4	12,18	14,79	$7{,}3\times10^{-10}$	94,9 (89,8 - 99,9)	14,1	88,2	92,3
miR-30a-5p	14,21	16,79	$1{,}6\times10^{-06}$	95,1 (90,6 - 99,7)	15,2	100,0	87,7	14,22	16,51	$7{,}8\times10^{-11}$	97,4 (95,0 - 99,7)	15,2	100,0	84,5
miR-378a-5p	9,06	10,19	$1{,}7\times10^{-03}$	79,6 (63,9 - 95,2)	10,2	54,5	94,5	8,83	9,99	$6{,}0\times10^{-06}$	83,0 (72,0 - 93,9)	9,8	64,7	88,6
miR-378c	5,13	6,33	$2{,}2\times10^{-05}$	89,9 (81,5 - 98,3)	5,6	100,0	67,1	4,83	6,88	$3{,}0\times10^{-08}$	90,4 (82,7 - 98,0)	5,9	88,2	85,0
miR-299-5p	4,40	5,71	$5{,}4\times10^{-04}$	82,6 (69,5 - 95,6)	5,2	90,9	72,6	4,55	6,47	$3{,}4\times10^{-06}$	83,9 (72,7 - 95,0)	6,0	70,6	89,5
miR-136-3p	5,04	7,13	$1{,}3\times10^{-06}$	95,5 (88,6 - 100,0)	6,4	90,9	94,5	5,12	7,28	$7{,}3\times10^{-10}$	94,9 (91,3 - 98,4)	5,9	100,0	78,2
miR-99a-5p	9,97	12,59	$1{,}2\times10^{-07}$	99,8 (99,2 - 100,0)	11,8	100,0	98,6	9,83	12,71	$1{,}9\times10^{-10}$	96,4 (92,9 - 99,9)	11,5	88,2	95,0
miR-29c-3p	12,30	14,90	$4{,}8\times10^{-06}$	93,0 (86,7 - 99,3)	13,3	100,0	74,0	11,92	14,79	$1{,}7\times10^{-10}$	96,5 (92,9 - 100,0)	13,8	94,1	93,2
miR-411-5p	5,22	7,17	$4{,}7\times10^{-07}$	97,4 (92,9 - 100,0)	6,6	90,9	97,3	4,94	7,46	$1{,}9\times10^{-07}$	87,9 (80,4 - 95,5)	5,9	82,4	77,3
miR-375	9,86	14,50	$1{,}2\times10^{-05}$	91,2 (84,0 - 98,3)	13,1	90,9	83,6	9,95	14,45	$1{,}6\times10^{-07}$	88,2 (75,5 - 100,0)	12,9	88,2	92,3
miR-139-3p	5,75	8,24	$6{,}5\times10^{-06}$	92,4 (85,0 - 99,8)	6,5	90,9	79,5	5,24	7,62	$6{,}8\times10^{-10}$	94,9 (91,6 - 98,3)	6,8	94,1	88,6
miR-378a-3p	11,69	13,65	$2{,}1\times10^{-05}$	90,0 (81,3 - 98,7)	13,1	72,7	91,8	11,59	13,31	$3{,}6\times10^{-07}$	87,1 (79,2 - 94,9)	12,6	76,5	84,5
miR-195-5p	6,28	8,78	$3{,}8\times10^{-07}$	97,8 (94,9 - 100,0)	7,7	100,0	95,9	6,24	8,40	$1{,}7\times10^{-09}$	93,9 (89,8 - 98,0)	7,1	94,1	81,4
miR-376c-3p	4,49	6,56	$5{,}3\times10^{-05}$	88,0 (74,5 - 100,0)	5,3	90,9	82,2	4,55	6,56	$1{,}8\times10^{-08}$	91,0 (83,6 - 98,5)	5,8	88,2	88,6
miR-30c-2-3p	5,49	9,02	$8{,}7\times10^{-07}$	96,3 (92,4 - 100,0)	7,2	100,0	91,8	5,60	7,56	$1{,}1\times10^{-08}$	91,6 (85,6 - 97,6)	7,1	88,2	88,6
miR-378d	1,95	3,61	$2{,}1\times10^{-03}$	78,9 (61,3 - 96,5)	3,6	54,5	95,9	1,57	3,81	$3{,}1\times10^{-06}$	83,9 (70,7 - 97,0)	2,9	76,5	88,6
let-7c-5p	12,04	14,83	$1{,}5\times10^{-07}$	99,4 (98,0 - 100,0)	13,8	100,0	93,2	12,03	14,62	$1{,}5\times10^{-09}$	94,1 (89,2 - 98,9)	13,7	88,2	89,1
miR-379-3p	1,89	4,08	$6{,}4\times10^{-05}$	87,5 (71,4 - 100,0)	3,1	81,8	90,4	1,92	3,38	$6{,}4\times10^{-06}$	82,8 (72,1 - 93,5)	2,6	88,2	72,7
miR-411-3p	2,67	4,41	$4{,}4\times10^{-04}$	83,1 (68,2 - 97,9)	4,2	63,6	98,6	2,66	3,94	$1{,}7\times10^{-04}$	77,4 (64,6 - 90,2)	3,3	76,5	71,4
miR-410-3p	5,54	8,08	$4{,}4\times10^{-07}$	97,5 (93,0 - 100,0)	7,2	90,9	97,3	5,51	7,32	$1{,}8\times10^{-07}$	88,0 (81,4 - 94,6)	6,6	94,1	80,0
miR-379-5p	11,81	13,93	$2{,}5\times10^{-07}$	98,5 (96,2 - 100,0)	13,0	100,0	89,0	11,91	13,74	$9{,}8\times10^{-07}$	85,7 (76,6 - 94,7)	12,4	94,1	66,8
miR-486-5p	9,19	10,28	$2{,}7\times10^{-03}$	78,2 (65,5 - 90,9)	9,2	100,0	54,8	9,16	11,10	$1{,}2\times10^{-05}$	81,9 (70,5 - 93,3)	10,9	58,8	93,2
miR-377-5p	3,22	4,72	$2{,}5\times10^{-04}$	84,4 (73,1 - 95,7)	4,2	81,8	78,1	3,29	4,95	$3{,}1\times10^{-08}$	90,3 (85,1 - 95,6)	4,0	94,1	75,9
miR-381-5p	1,23	1,89	$6{,}4\times10^{-02}$	67,2 (48,7 - 85,8)	1,3	81,8	54,8	1,49	2,73	$5{,}4\times10^{-04}$	75,1 (60,8 - 89,3)	2,6	64,7	80,9
miR-1-3p	8,91	10,10	$4{,}0\times10^{-02}$	69,4 (51,8 - 87,0)	6,7	100,0	39,7	7,52	12,51	$1{,}1\times10^{-04}$	78,1 (66,1 - 90,1)	12,5	58,8	90,0
miR-204-5p	2,45	5,35	$7{,}5\times10^{-07}$	96,5 (92,0 - 100,0)	4,8	90,9	95,9	2,26	6,35	$3{,}2\times10^{-10}$	95,7 (93,2 - 98,3)	4,7	100,0	88,2
miR-6510-3p	4,13	6,80	$1{,}2\times10^{-02}$	73,6 (53,1 - 94,1)	6,8	54,5	93,2	4,07	7,51	$7{,}9\times10^{-06}$	82,5 (70,6 - 94,5)	6,2	70,6	86,8
miR-133a-3p	9,50	8,65	$1{,}4\times10^{-01}$	36,0 (16,9 - 55,0)	8,7	54,5	56,2	7,61	12,10	$5{,}4\times10^{-04}$	75,2 (62,2 - 88,2)	11,7	58,8	84,5
miR-133b	6,83	7,82	$8{,}5\times10^{-02}$	66,3 (48,3 - 84,2)	4,0	100,0	37,0	5,40	9,61	$3{,}8\times10^{-04}$	75,9 (63,4 - 88,4)	9,4	58,8	83,2

Nota: IC - Intervalo de confiança.

Tabela 24 – DEmiRNAs identificados com o teste WMW nos estadiamentos iniciais e avançados e áreas sob a curva (AUC), especificidade (TNR) e sensibilidade (TPR) na capacidade de discriminação do tumor primário (T) e tecido normal adjacente ao tumor (NAT) na laringe.

miRNAs	Laringe													
	Estadiamento inicial ($n = 18$)							Estadiamento avançado ($n = 94$)						
	Mediana		Valor p	AUC ($IC_{95}\%$)	*Cuttof*	TNR	TPR	Mediana		Valor p	AUC ($IC_{95}\%$)	*Cuttof*	TNR	TPR
	T	NAT						T	NAT					
miR-196b-5p	9,78	5,91	0,0001	100,0 (100,0 - 100,0)	7,3	100,0	100,0	9,61	6,94	0,0003	94,3 (89,2 - 99,4)	7,9	100,0	87,5
miR-196a-5p	8,19	2,83	0,0032	91,7 (77,4 - 100,0)	5,7	100,0	83,3	7,88	3,95	0,0037	85,6 (78,0 - 93,2)	5,1	100,0	83,0
miR-615-3p	3,19	0,00	0,0017	95,8 (87,7 - 100,0)	0,3	100,0	91,7	2,76	0,00	0,0002	95,0 (91,1 - 98,9)	1,0	100,0	89,8
miR-4652-5p	4,31	1,10	0,0001	100,0 (100,0 - 100,0)	2,7	100,0	100,0	4,201	2,01	0,0006	92,4 (85,4 - 99,4)	3,1	100,0	81,8
miR-455-3p	9,43	7,32	0,0069	88,9 (69,2 - 100,0)	8,4	83,3	91,7	10,28	7,80	0,0004	93,2 (85,1 - 100,0)	8,2	83,3	93,2
miR-767-5p	7,02	1,68	0,0150	86,8 (69,4 - 100,0)	4,3	100,0	75,0	3,949	1,629	0,0330	76,1 (65,0 - 87,2)	2,4	100,0	65,9
miR-105-5p	8,16	2,48	0,0097	87,5 (70,0 - 100,0)	2,7	100,0	83,3	4,21	2,39	0,3800	60,8 (49,7 - 71,8)	3,4	100,0	53,4
miR-301a-3p	4,57	2,78	0,0069	88,9 (72,3 - 100,0)	4,3	100,0	66,7	4,93	2,79	0,0002	96,2 (90,8 - 100,0)	4,0	100,0	84,1
miR-193b-3p	8,35	6,72	0,0004	97,2 (90,7 - 100,0)	7,6	100,0	83,3	9,02	7,47	0,0026	86,9 (78,3 - 95,6)	8,2	100,0	76,1
miR-9-5p	11,95	10,46	0,1000	75,0 (51,3 - 98,7)	11,6	100,0	66,7	12,06	10,05	0,0092	82,0 (66,8 - 97,2)	11,0	83,3	72,7
miR-503-5p	3,81	1,92	0,0057	91,7 (78,5 - 100,0)	3,2	100,0	83,3	4,02	1,77	0,0010	90,5 (77,8 - 100,0)	2,0	66,7	98,9
miR-1910-5p	2,54	0,00	0,0012	98,6 (94,8 - 100,0)	1,7	100,0	91,7	2,40	0,00	0,0007	91,7 (86,5 - 96,8)	0,8	100,0	84,1
miR-210-3p	11,31	8,95	0,0047	90,3 (76,1 - 100,0)	10,2	83,3	83,3	11,70	9,08	0,0048	84,7 (67,0 - 100,0)	10,2	83,3	85,2
miR-301a-5p	3,31	1,96	0,0410	80,6 (58,4 - 100,0)	2,9	83,3	66,7	3,66	1,84	0,0013	89,4 (78,2 - 100,0)	2,3	83,3	90,9
miR-1269a	7,79	2,84	0,0750	77,1 (54,6 - 99,6)	6,0	100,0	66,7	7,27	3,37	0,0820	71,4 (58,2 - 84,6)	7,2	100,0	54,5
miR-301b-3p	2,22	0,92	0,1400	72,2 (48,1 - 96,3)	2,0	100,0	58,30	2,34	0,00	0,0014	89,1 (77,2 - 100,0)	0,5	83,3	89,8
miR-31-5p	6,54	5,93	0,1800	70,8 (43,6 - 98,0)	6,2	66,7	75,0	7,69	4,88	0,0096	81,8 (71,6 - 92,1)	6,3	100,0	68,2
miR-3662	2,10	1,22	0,0240	83,3 (64,1 - 100,0)	1,9	100,0	66,70	2,33	0,85	0,0037	85,6 (76,9 - 94,4)	1,7	100,0	71,6
miR-1293	2,75	1,90	0,1000	75,0 (51,2 - 98,8)	2,8	100,0	50,0	3,50	2,30	0,0180	79,2 (69,0 - 89,3)	3,0	100,0	68,2
miR-509-3p	4,87	3,67	0,3800	63,9 (35,9 - 91,9)	4,2	83,3	58,3	2,82	3,36	0,8800	51,9 (32,5 - 71,3)	2,9	66,7	56,8
miR-4326	4,10	0,99	0,0980	75,0 (45,4 - 100,0)	2,9	83,3	83,3	4,26	2,04	0,0083	82,4 (72,8 - 92,0)	3,5	100,0	67,0
miR-183-3p	1,34	2,95	0,0210	84,7 (66,3 - 100,2)	2,0	100,0	75,0	2,44	1,32	0,0110	81,2 (67,2 - 95,3)	2,2	100,0	61,4
miR-9-3p	2,85	2,51	0,5500	59,7 (32,8 -86,6)	3,6	100,0	41,7	3,18	2,59	0,3400	61,7 (46,3 - 77,2)	2,8	83,3	59,1
miR-3170	1,92	0,53	0,0560	78,5 (57,7 - 99,3)	1,8	100,0	58,3	1,63	0,00	0,0018	88,2 (81,7 - 94,6)	0,8	100,0	77,3
miR-6499-3p	2,46	1,55	0,2400	68,1 (35,9 - 100,0)	1,8	66,7	75,0	2,95	2,01	0,0450	74,6 (57,4 - 91,9)	3,0	100,0	47,7
miR-4449	1,91	0,40	0,1400	72,2 (46,8 - 97,6)	2,2	100,0	41,7	1,95	1,04	0,0430	74,8 (60,8 - 88,8)	1,7	100,0	53,4
miR-1305	1,86	0,40	0,0084	89,6 (74,5 - 100,0)	1,1	100,0	83,3	1,13	0,00	0,0470	73,7 (61,4 - 86,0)	1,0	100,0	52,3
miR-508-3p	5,20	5,85	0,6200	58,3 (30,3 - 86,4)	5,1	100,0	50,0	4,46	4,42	0,7500	46,0 (28,0 - 64,1)	8,3	100,0	15,9
miR-514a-3p	3,71	4,91	0,8200	54,2 (26,5 - 81,8)	3,6	83,3	50,0	3,49	3,84	0,9200	48,7 (28,9 - 68,4)	2,2	100,0	18,2
miR-184	2,83	3,13	0,5500	59,7 (31,9 - 87,5)	3,6	83,3	41,7	2,48	3,44	0,6900	54,9 (38,3 - 71,5)	3,3	66,7	62,5
miR-338-5p	3,91	5,72	0,0670	77,8 (50,3 - 100,0)	4,5	83,3	75,0	3,64	6,41	0,0005	99,8 (99,3 - 100,0)	5,3	100,0	98,9
miR-139-3p	5,51	8,39	0,002	93,1 (81,5 - 100,0)	5,8	100,0	75,0	5,18	8,88	0,0001	98,9 (97,0 - 100,0)	7,1	100,0	95,5
miR-378c	4,91	6,04	0,0410	80,6 (57,8 - 100,0)	5,5	83,3	75,0	4,79	6,41	0,0009	90,7 (81,4 - 100,0)	5,5	100,0	73,9
miR-378d	1,38	2,44	0,1200	73,6 (48,8 - 98,4)	2,1	83,3	75,0	1,61	2,70	0,0043	85,0 (70,0 - 100,0)	2,2	83,3	75,0
miR-99a-5p	10,68	12,58	0,0013	94,4 (84,5 - 100,0)	11,8	100,0	83,3	10,38	13,34	0,0001	99,2 (97,8 - 100,0)	12,5	100,0	97,7
miR-29c-3p	13,06	15,08	0,0047	90,3 (75,2 - 100,0)	14,6	100,0	83,3	12,51	14,64	0,0010	90,5 (79,3 - 100,0)	14,2	83,3	88,6
miR-375	10,59	15,59	0,0004	97,2 (90,8 - 100,0)	13,4	100,0	91,7	11,03	14,24	0,0015	89,0 (81,1 - 96,9)	12,6	100,0	75,0
miR-144-3p	5,19	6,69	0,3400	65,3 (33,6 - 97,0)	5,9	83,3	66,7	4,63	7,15	0,0092	82,0 (54,5 - 100,0)	6,5	83,3	90,9
miR-486-5p	9,10	9,89	0,2900	66,7 (39,3 - 94,1)	9,3	83,3	94,1	8,67	10,90	0,0002	96,0 (92,0 - 100,0)	10,3	100,0	93,2
miR-6510-3p	5,75	8,02	0,0069	88,9 (71,1 - 100,0)	7,1	83,3	83,3	4,57	8,46	0,0002	95,3 (88,8 - 100,0)	8,2	83,3	97,7
miR-1-3p	5,38	6,45	0,1000	75,0 (50,1 - 99,9)	5,9	83,3	75,0	4,53	10,44	0,0023	87,5 (76,5 - 98,5)	6,9	100,0	73,9
miR-133a-3p	4,73	6,13	0,3400	65,3 (37,2 -93,4)	5,4	83,3	66,7	4,47	9,89	0,0032	86,2 (73,7 - 98,6)	6,1	100,0	70,5
miR-206	6,55	6,68	0,9600	51,4 (22,4 - 80,4)	5,9	83,3	41,7	5,32	10,98	0,0100	81,4 (66,5 - 96,4)	7,3	100,0	64,8
miR-133b	2,28	4,00	0,3300	65,3 (38,2 - 92,4)	2,9	83,3	58,3	2,17	7,059	0,0052	84,3 (71,9 - 65,9)	3,7	100,0	65,9

Nota: IC - Intervalo de confiança.