RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo desta tese será disponibilizado somente a partir de 21/02/2026.

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

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

Bethina da Rocha Camargo

Botucatu
2024

Bethina da Rocha Camargo

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

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

Orientador: Prof. Dr. Rogério Antonio de Oliveira

Coorientadora: Profa. Dra. Patricia Pintor dos Reis

Botucatu
2024

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

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

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

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Rogério Antônio de Oliveira
Coorientador: Patrícia Pintor dos Reis
Capes: 10203001

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

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

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Botucatu

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

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

Prof. Dr. ROGERIO ANTONIO DE OLIVEIRA

gov.br Documento assinado digitalmente
ROGERIO ANTONIO DE OLIVEIRA
Data: 28/02/2024 09:45:25-0300
Verifique em https://validar.iti.gov.br

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

Agradecimentos

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

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

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

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

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

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

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

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

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

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

Mateus 6:6

Resumo

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

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

Abstract

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

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

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

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

Sumário

1 Introdução

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

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

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

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

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

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

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

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

1.1 Objetivos

Objetivo geral

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

Objetivos Específicos

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

1.2 Apresentação da tese

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

5 Conclusão

Este trabalho utilizou técnicas de bioinformática, incluindo metodologias estatísticas e ferramentas biológicas, para analisar grandes conjuntos de dados de miRNA-Seq e RNA-Seq, visando identificar potenciais biomarcadores associados ao LUAD e LUSC nos estadiamentos iniciais da doença que diferenciem T vs. N, que estejam relacionadas a sobrevivência dos pacientes e que diferenciam as doenças LUSC vs. LUAD. Todos esses procedimentos foram realizados com o intuito de conseguir compreender os mecanismos subjacentes à formação do câncer e colaborar para o desenvolvimento de estratégias diagnósticas mais precisas.

Vale destacar, que com o avanço da medicina moderna e a integração de NGS como ferramentas rotineiras de diagnóstico, prognóstico ou tratamento, há a necessidade de identificação precisa de moléculas clinicamente úteis, incluindo os RNAs e miRNAs, para LUAD e LUSC em estadiamento inicial.

Abaixo, estão destacados os resultados mais importantes deste trabalho e as conclusões pontuais em relação às análises I, II e III:

- **Análise I:** O teste de Wilcoxon-Mann-Whitney com correção de Bonferroni mostrou ser uma ferramenta mais simples e conservadora para a identificação de DE-miRNAs no banco de dados LUAD miRNA-seq quando comparado ao EdgeR e Deseq2, sendo assim, pode ser sugerido para detecção eficiente de DE-miRNAs, ao demonstrar altas taxas de acurácia na diferenciação de T vs. N em estadiamentos iniciais. Contudo, vale ressaltar que neste estudo é evidenciado que a escolha da metodologia pode influenciar nos resultados e consequentemente nas interpretações biológicas, sendo assim, sempre é importante estudar e analisar metodologias e ferramentas antes de utilizá-las.

- **Análises II e III:** Os Testes de Wilcoxon-Mann-Whitney, com correção de Bonferroni, realizados nas análises identificaram de 18% a 33% DE-miRNAs ou DEGs. As análises biológicas foram fundamentais para identificar pares de DE-miRNAs e DEGs com expressões opostas, correlações negativas e validados experimentalmente, foram identificados nas análises de 80 a 160 pares. Identificaram-se importantes vias biológicas relacionadas ao câncer, redes mostrando as ligações entre os pares e, por fim, a PPI confirmou a relação funcional e os fenômenos biológicos associados. As análises estatísticas em conjunto possibilitaram entender a importância desses DE-miRNAs e DEGs identificados para a discriminação

do tipo de tecido ou da doença e, também, para a sobrevivência dos pacientes. Assim, seguindo o fluxograma proposto, nas análises II e III, foram identificados importantes e potenciais RNAs e miRNAs para os pacientes em estadiamento inicial.

Por fim, foram identificadas valiosas informações a respeito de DE-miRNAs e DEGs que obtiveram comportamentos diferentes dos demais, podendo ser candidatos a potenciais biomarcadores de LUAD e LUSC em estadiamento inicial, e, além disso, apresentou árvores tanto para a classificação, quanto para a sobrevivência, que possuem potencial a serem mais investigadas e posteriormente utilizadas para diagnóstico e entendimento a respeito do tipo de tecido e subtipo da doença.

Referências

AGRESTI, A. A survey of exact inference for contingency tables. *Statistical science*, Institute of Mathematical Statistics, v. 7, n. 1, p. 131–153, 1992. 36

AGRESTI, A. *An introduction to categorical data analysis*. [S.l.]: John Wiley & Sons, 2018. 36

ALBERG, A. J.; BROCK, M. V.; SAMET, J. M. Epidemiology of lung cancer: looking to the future. *Journal of clinical oncology*, American Society of Clinical Oncology, v. 23, n. 14, p. 3175–3185, 2005. 7

ALBERTS, B. *et al. Molecular biology of the cell*. [S.l.]: Garland science, 2014. 12

AMBROS, V. The functions of animal micrornas. *Nature*, Nature Publishing Group UK London, v. 431, n. 7006, p. 350–355, 2004. 16

AMIT, Y.; GEMAN, D. Shape quantization and recognition with randomized trees. *Neural computation*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., v. 9, n. 7, p. 1545–1588, 1997. 30

ANUSEWICZ, D.; ORZECHOWSKA, M.; BEDNAREK, A. K. Lung squamous cell carcinoma and lung adenocarcinoma differential gene expression regulation through pathways of notch, hedgehog, wnt, and erbb signalling. *Scientific reports*, Springer, v. 10, n. 1, p. 1–15, 2020. 8, 9

ARDEKANI, A. M.; NAEINI, M. M. The role of micrornas in human diseases. *Avicenna journal of medical biotechnology*, Avicenna Research Institute, v. 2, n. 4, p. 161, 2010. 18

AZIZI, M. I. H. N.; OTHMAN, I.; NAIDU, R. The role of micrornas in lung cancer metabolism. *Cancers*, MDPI, v. 13, n. 7, p. 1716, 2021. 20

BARANAUSKAS, J. A.; MONARD, M. C. Reviewing some machine learning concepts and methods. 2000. 43

BARBA-ALIAGA, M.; ALEPUZ, P.; PÉREZ-ORTÍN, J. E. Eukaryotic rna polymerases: the many ways to transcribe a gene. *Frontiers in Molecular Biosciences*, Frontiers Media SA, v. 8, p. 663209, 2021. 17

BARBACID, M. Ras genes. *Annual review of biochemistry*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 56, n. 1, p. 779–827, 1987. 21

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, Wiley Online Library, v. 57, n. 1, p. 289–300, 1995. 24, 38, 40

BHAYANI, M. K.; CALIN, G. A.; LAI, S. Y. Functional relevance of mirna* sequences in human disease. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, Elsevier, v. 731, n. 1-2, p. 14–19, 2012. 20

BIAU, G.; SCORNET, E. A random forest guided tour. *Test*, Springer, v. 25, p. 197–227, 2016. 30, 59, 62

BIORENDER. *BioRender*. 2022. ix, 13, 17

BISCHL, B. *et al.* Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, p. e1484, 2021. 61

BLAGUS, R.; LUSA, L. Smote for high-dimensional class-imbalanced data. *BMC bioinformatics*, Springer, v. 14, p. 1–16, 2013. 46

BOERI, M. *et al.* Microrna signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 108, n. 9, p. 3713–3718, 2011. 150

BOFFETTA, P.; JÄRVHOLM, B.; BRENNAN, P.; NYRÉN, O. Incidence of lung cancer in a large cohort of non-smoking men from sweden. *International journal of cancer*, Wiley Online Library, v. 94, n. 4, p. 591–593, 2001. 6

BOGEDALE, K.; JAGANNATHAN, V.; GERBER, V.; UNGER, L. Differentially expressed micrornas, including a large microrna cluster on chromosome 24, are associated with equine sarcoid and squamous cell carcinoma. *Veterinary and comparative oncology*, Wiley Online Library, v. 17, n. 2, p. 155–164, 2019. 80

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, p. 123–140, 1996. 30

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001. 30, 58, 59

BREIMAN, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, Institute of Mathematical Statistics, v. 16, n. 3, p. 199–231, 2001. 41

BREIMAN, L.; CUTLER, A. *Random forests: Classification/clustering*. [S.l.]: Retrieved May, 2004. 59

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. Classification and regression trees. *CA: Wadsworth International Group*, n. 368, 1984. 29, 47, 55, 56, 60

CALIN, G. A. *et al.* Frequent deletions and down-regulation of micro-rna genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 99, n. 24, p. 15524–15529, 2002. 21

CAMARGO, B. da R. *et al.* A decision tree-based classifier compares three data analysis methods for the identification of mirnas associated with early-stage lung cancer. *REVISTA FOCO*, v. 16, n. 5, p. e2031–e2031, 2023. ix, 68, 69, 70, 72, 75, 76, 77, 78

CHANG, J. T.-H.; LEE, Y. M.; HUANG, R. S. The impact of the cancer genome atlas on lung cancer. *Translational Research*, Elsevier, v. 166, n. 6, p. 568–585, 2015. 22

CHANG, W.-H. *et al.* Jag1 is associated with poor survival through inducing metastasis in lung cancer. *PloS one*, Public Library of Science San Francisco, CA USA, v. 11, n. 3, p. e0150355, 2016. 150

CHARKIEWICZ, R. *et al.* Gene expression signature differentiates histology but not progression status of early-stage nsclc. *Translational oncology*, Elsevier, v. 10, n. 3, p. 450–458, 2017. 9

CHAUDHURI, P.; LOH, W.-Y. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, JSTOR, p. 561–576, 2002. 29

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. 46

CHEN, B. *et al.* A regulatory circuitry comprising tp53, mir-29 family, and setdb1 in non-small cell lung cancer. *Bioscience reports*, Portland Press Ltd., v. 38, n. 5, p. BSR20180678, 2018. 129

CHEN, H. *VennDiagram: Generate High-Resolution Venn and Euler Plots*. [S.l.], 2022. R package version 1.7.3. Disponível em: <https://CRAN.R-project.org/package=VennDiagram>. 40

CHEN, J. W.; DHAHBI, J. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Scientific reports*, Nature Publishing Group UK London, v. 11, n. 1, p. 13323, 2021. 30

CHEN, L. *et al.* Machine learning and network methods for biology and medicine. *Computational and Mathematical Methods in Medicine*, Hindawi, v. 2015, 2015. 104

CHEN, M. *et al.* Differentiated regulation of immune-response related genes between luad and lusc subtypes of lung cancers. *Oncotarget*, Impact Journals, LLC, v. 8, n. 1, p. 133, 2017. 1, 8

CHEN, Q. *et al.* Microrna-301a promotes growth and migration by repressing tgfbr2 in non-small cell lung cancer. *Int J Clin Exp Pathol*, v. 10, n. 2, p. 957–971, 2017. 103

CHEN, Q. *et al.* mir-210-3p promotes lung cancer development and progression by modulating usf1 and pcgf3. *OncoTargets and therapy*, Dove Press, v. 14, p. 3687, 2021. 129

CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. *Genomics*, Elsevier, v. 99, n. 6, p. 323–329, 2012. 27, 28, 30

CHI, X.-J. *et al.* Identification of high expression profiles of mir-31-5p and its vital role in lung squamous cell carcinoma: a survey based on qrt-pcr and bioinformatics analysis. *Translational Cancer Research*, AME Publications, v. 8, n. 3, p. 788, 2019. 130

CIMMINO, A. *et al.* mir-15 and mir-16 induce apoptosis by targeting bcl2. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 102, n. 39, p. 13944–13949, 2005. 22

CINEGAGLIA, N. C. *et al.* Integrative transcriptome analysis identifies deregulated microrna-transcription factor networks in lung adenocarcinoma. *Oncotarget*, Impact Journals, LLC, v. 7, n. 20, p. 28920, 2016. 21

COLOSIMO, E. A.; GIOLO, S. R. Análise de sobrevivência. *São Paulo: Abe-Projeto Fisher*, 2006. 51, 52

CORCHETE, L. A. *et al.* Systematic comparison and assessment of rna-seq procedures for gene expression quantitative analysis. *Scientific reports*, Nature Publishing Group UK London, v. 10, n. 1, p. 19737, 2020. 80

CORRALES, L. *et al.* Lung cancer in never smokers: The role of different risk factors other than tobacco smoking. *Critical reviews in oncology/hematology*, Elsevier, v. 148, p. 102895, 2020. 6

COURAUD, S. *et al.* Lung cancer in never smokers–a review. *European journal of cancer*, Elsevier, v. 48, n. 9, p. 1299–1311, 2012. 6

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. 53, 54

COX, D. R. Partial likelihood. *Biometrika*, Oxford University Press, v. 62, n. 2, p. 269–276, 1975. 54

CRUZ, C. S. D.; TANOUE, L. T.; MATTHAY, R. A. Lung cancer: epidemiology, etiology, and prevention. *Clinics in chest medicine*, Elsevier, v. 32, n. 4, p. 605–644, 2011. 6

DAVID, C. J.; MASSAGUÉ, J. Contextual determinants of tgfβ action in development, immunity and cancer. *Nature reviews Molecular cell biology*, Nature Publishing Group UK London, v. 19, n. 7, p. 419–435, 2018. 103

DENISON, D. G.; MALLICK, B. K.; SMITH, A. F. A bayesian cart algorithm. *Biometrika*, Oxford University Press, v. 85, n. 2, p. 363–377, 1998. 29

DETTERBECK, F. C.; BOFFA, D. J.; KIM, A. W.; TANOUE, L. T. The eighth edition lung cancer stage classification. *Chest*, Elsevier, v. 151, n. 1, p. 193–203, 2017. 10

DEVROYE, L.; GYÖRFI, L.; LUGOSI, G. *A probabilistic theory of pattern recognition*. [S.l.]: Springer Science & Business Media, 2013. v. 31. 59

DIAS, M. *et al.* Lung cancer in never-smokers–what are the differences? *Acta oncologica*, Taylor & Francis, v. 56, n. 7, p. 931–935, 2017. 7

DÍAZ-URIARTE, R.; ANDRÉS, S. Alvarez de. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, Springer, v. 7, p. 1–13, 2006. 58

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., v. 10, n. 7, p. 1895–1923, 1998. 41

DIETTERICH, T. G. Ensemble methods in machine learning. In: SPRINGER. *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*. [S.l.], 2000. p. 1–15. 30

DONSON, A. M. *et al.* Mgmt promoter methylation correlates with survival benefit and sensitivity to temozolomide in pediatric glioblastoma. *Pediatric blood & cancer*, Wiley Online Library, v. 48, n. 4, p. 403–407, 2007. 14

DUBIN, S.; GRIFFIN, D. Lung cancer in non-smokers. *Missouri medicine*, Missouri State Medical Association, v. 117, n. 4, p. 375, 2020. 1, 6, 7

EFRON, B. Bootstrap methods: another look at the jackknife. In: *Breakthroughs in statistics: Methodology and distribution*. [S.l.]: Springer, 1992. p. 569–593. 58

ELGELDAWI, E.; SAYED, A.; GALAL, A. R.; ZAKI, A. M. Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In: MULTIDISCIPLINARY DIGITAL PUBLISHING INSTITUTE. *Informatics*. [S.l.], 2021. v. 8, n. 4, p. 79. 61

ESTELLER, M. *et al.* Inactivation of the dna repair gene o 6-methylguanine-dna methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. *Cancer research*, AACR, v. 59, n. 4, p. 793–797, 1999. 14

ETTINGER, D. S. Overview and state of the art in the management of lung cancer. *Oncology (Williston Park, NY)*, v. 18, n. 7 Suppl 4, p. 3–9, 2004. 7

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. Inteligência artificial: uma abordagem de aprendizado de máquina. 2011. ix, 27, 28, 41, 42, 43, 44, 47, 50, 62

FARUKI, H. *et al.* Lung adenocarcinoma and squamous cell carcinoma gene expression subtypes demonstrate significant differences in tumor immune landscape. *Journal of Thoracic Oncology*, Elsevier, v. 12, n. 6, p. 943–953, 2017. 9

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. 64

FENG, Y.-H.; TSAO, C.-J. Emerging role of microrna-21 in cancer. *Biomedical reports*, Spandidos Publications, v. 5, n. 4, p. 395–402, 2016. 22

FISHER, R. A. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the royal statistical society*, JSTOR, v. 85, n. 1, p. 87–94, 1922. 36

FISHER, R. A. *Statistical methods for research workers 12 th ed.* [S.l.]: Oliver & Body, 1954. 36

FONSECA, J. da. *Indução de Árvores de decisão*. Tese (Doutorado) — Dissertação de Mestrado, Universidade Nova de Lisboa, Lisboa, 1994. 48

FOSS, K. M. *et al.* mir-1254 and mir-574-5p: serum-based microrna biomarkers for early-stage non-small cell lung cancer. *Journal of thoracic oncology*, Elsevier, v. 6, n. 3, p. 482–488, 2011. 22

FOTOUHI, S.; ASADI, S.; KATTAN, M. W. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of biomedical informatics*, Elsevier, v. 90, p. 103089, 2019. 45

FRANKISH, A. *et al.* Gencode 2021. *Nucleic acids research*, Oxford University Press, v. 49, n. D1, p. D916–D923, 2021. 34

FRIEDMAN, J. H.; KOHAVI, R.; YUN, Y. Lazy decision trees. In: *AAAI/IAAI, Vol. 1*. [S.l.: s.n.], 1996. p. 717–724. 48

GAO, M.; KONG, W.; HUANG, Z.; XIE, Z. Identification of key genes related to lung squamous cell carcinoma using bioinformatics analysis. *International journal of molecular sciences*, MDPI, v. 21, n. 8, p. 2994, 2020. 8

GAUTHIER, M.; AGNIEL, D.; THIÉBAUT, R.; HEJBLUM, B. P. dearseq: a variance component score test for rna-seq differential analysis that effectively controls the false discovery rate. *NAR genomics and bioinformatics*, Oxford University Press, v. 2, n. 4, p. lqaa093, 2020. 25, 80

GAZDAR, A. Activating and resistance mutations of egfr in non-small-cell lung cancer: role in clinical response to egfr tyrosine kinase inhibitors. *Oncogene*, Nature Publishing Group, v. 28, n. 1, p. S24–S31, 2009. 12

GENG, Q. *et al.* Five micrornas in plasma as novel biomarkers for screening of early-stage non-small cell lung cancer. *Respiratory research*, BioMed Central, v. 15, n. 1, p. 1–9, 2014. 79

GENTLES, A. J. *et al.* Integrating tumor and stromal gene expression signatures with clinical indices for survival stratification of early-stage non–small cell lung cancer. *JNCI: Journal of the National Cancer Institute*, Oxford University Press, v. 107, n. 10, 2015. 9

GIOLO, S. R. *Introdução à análise de dados categóricos com aplicações*. [S.l.]: Editora Blucher, 2017. 35, 36, 37, 54

GIRBIG, M.; MISIASZEK, A. D.; MÜLLER, C. W. Structural insights into nuclear transcription by eukaryotic dna-dependent rna polymerases. *Nature Reviews Molecular Cell Biology*, Nature Publishing Group UK London, v. 23, n. 9, p. 603–622, 2022. 17

GOLDSTRAW, P. *et al.* The iaslc lung cancer staging project: proposals for revision of the tnm stage groupings in the forthcoming (eighth) edition of the tnm classification for lung cancer. *Journal of Thoracic Oncology*, Elsevier, v. 11, n. 1, p. 39–51, 2016. 11

GORLOVA, O. Y. *et al.* Aggregation of cancer among relatives of never-smoking lung cancer patients. *International journal of cancer*, Wiley Online Library, v. 121, n. 1, p. 111–118, 2007. 7

GRIDELLI, C. *et al.* Non-small-cell lung cancer. *Nature reviews Disease primers*, Nature Publishing Group, v. 1, n. 1, p. 1–16, 2015. 8

HAMFJORD, J. *et al.* Differential expression of mirnas in colorectal cancer: comparison of paired tumor tissue and adjacent normal mucosa using high-throughput sequencing. *PloS one*, Public Library of Science San Francisco, USA, v. 7, n. 4, p. e34150, 2012. 80

HAMNER, B.; FRASCO, M. *Metrics: Evaluation Metrics for Machine Learning*. [S.l.], 2018. R package version 0.1.4. Disponível em: <https://CRAN.R-project.org/package=Metrics>. 65

Harrell Jr, F. E. *Hmisc: Harrell Miscellaneous*. [S.l.], 2023. R package version 5.0-1. Disponível em: <https://CRAN.R-project.org/package=Hmisc>. 40

HART, A. Mann-whitney test is not just a test of medians: differences in spread can be important. *Bmj*, British Medical Journal Publishing Group, v. 323, n. 7309, p. 391–393, 2001. 37

HE, J.-H. *et al.* Analyzing the lncrna, mirna, and mrna regulatory network in prostate cancer with bioinformatics software. *Journal of Computational Biology*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 25, n. 2, p. 146–157, 2018. 80

HENNESSEY, P. T. *et al.* Serum microrna biomarkers for detection of non-small cell lung cancer. *PloS one*, Public Library of Science San Francisco, USA, v. 7, n. 2, p. e32307, 2012. 22

HERBST, R. S.; JR, P. A. B. Targeting the epidermal growth factor receptor in non-small cell lung cancer. *Clinical Cancer Research*, AACR, v. 9, n. 16, p. 5813–5824, 2003. 9

HIRONO, T. *et al.* Microrna-130b functions as an oncomirna in non-small cell lung cancer by targeting tissue inhibitor of metalloproteinase-2. *Scientific Reports*, Nature Publishing Group UK London, v. 9, n. 1, p. 6956, 2019. 103

HO, T. K. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, Ieee, v. 20, n. 8, p. 832–844, 1998. 30

HOGEWEG, P. The roots of bioinformatics in theoretical biology. *PLoS computational biology*, Public Library of Science San Francisco, USA, v. 7, n. 3, p. e1002021, 2011. 23

HOLM, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, JSTOR, p. 65–70, 1979. 38

HOOD, L.; ROWEN, L. The human genome project: big science transforms biology and medicine. *Genome medicine*, Springer, v. 5, p. 1–8, 2013. 23

HOTHORN, T.; HORNIK, K.; ZEILEIS, A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, Taylor & Francis, v. 15, n. 3, p. 651–674, 2006. 29

HOTHORN, T.; ZEILEIS, A. partykit: A modular toolkit for recursive partytioning in r. *The Journal of Machine Learning Research*, JMLR. org, v. 16, n. 1, p. 3905–3909, 2015. 48, 50

HOU, J. *et al.* Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PloS one*, Public Library of Science San Francisco, USA, v. 5, n. 4, p. e10312, 2010. 9

HU, J. *et al.* Mir-497-5p down-regulates cdca4 to restrains lung squamous cell carcinoma progression. *Journal of Cardiothoracic Surgery*, Springer, v. 16, p. 1–8, 2021. 130

HU, Y. *et al.* Identification of key differentially expressed micrornas in cancer patients through pan-cancer analysis. *Computers in biology and medicine*, Elsevier, v. 103, p. 183–197, 2018. 80

HUANG, C.-Y. *et al.* A review on the effects of current chemotherapy drugs and natural agents in treating non–small cell lung cancer. *Biomedicine*, China Medical University, v. 7, n. 4, 2017. 11

HUANG, L. *et al.* Let-7c-5p represses cisplatin resistance of lung adenocarcinoma cells by targeting cdc25a. *Applied Biochemistry and Biotechnology*, Springer, v. 195, n. 3, p. 1644–1655, 2023. 104

HUYNH, P.-H.; NGUYEN, V. H.; DO, T.-N. Improvements in the large p, small n classification issue. *SN Computer Science*, Springer, v. 1, p. 1–19, 2020. 30

INAMURA, K.; ISHIKAWA, Y. Microrna in lung cancer: Novel biomarkers and potential tools for treatment. *Journal of Clinical Medicine*, v. 5, n. 3, 2016. ISSN 2077-0383. Disponível em: <https://www.mdpi.com/2077-0383/5/3/36>. ix, 19, 20

INAMURA, K. *et al.* let-7 microrna expression is reduced in bronchioloalveolar carcinoma, a non-invasive carcinoma, and is not correlated with prognosis. *Lung cancer*, Elsevier, v. 58, n. 3, p. 392–396, 2007. 22

INCA. *Instituto Nacional de Câncer José Alencar Gomes da Silva. Estimativa 2025: Incidência de Câncer no Brasil. Rio de Janeiro*. 2022. Https://www.gov.br/inca/pt-br/assuntos/noticias/2022/inca-estima-704-mil-casos-de-cancer-por-ano-no-brasil-ate-2025. Acesso em 25 de janeiro de 2023. 5

IORIO, M. V.; CROCE, C. M. Microrna dysregulation in cancer: diagnostics, monitoring and therapeutics. a comprehensive review. *EMBO molecular medicine*, WILEY-VCH Verlag Weinheim, v. 4, n. 3, p. 143–159, 2012. 17

ISHWARAN, H. *et al.* High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, Taylor & Francis, v. 105, n. 489, p. 205–217, 2010. 30

ISLAM, K. *et al.* Comorbidity and survival in lung cancer patientscomorbidity and survival in lung cancer patients. *Cancer Epidemiology, Biomarkers & Prevention*, AACR, v. 24, n. 7, p. 1079–1085, 2015. 6

IWAMOTO, T. *et al.* Distinct gene expression profiles between primary breast cancers and brain metastases from pair-matched samples. *Scientific reports*, Nature Publishing Group UK London, v. 9, n. 1, p. 13343, 2019. 25

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.]: Rafael Izbicki, 2020. 43

JANSSEN-HEIJNEN, M. L. *et al.* Prevalence of co-morbidity in lung cancer patients and its relationship with treatment: a population-based study. *Lung cancer*, Elsevier, v. 21, n. 2, p. 105–113, 1998. 6

JOHNSON, S. M. *et al.* Ras is regulated by the let-7 microrna family. *Cell*, Elsevier, v. 120, n. 5, p. 635–647, 2005. 22

KADARA, H. *et al.* Early diagnosis and screening for lung cancer. *Cold Spring Harbor perspectives in medicine*, Cold Spring Harbor Laboratory Press, v. 11, n. 9, p. a037994, 2021. 1

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, Oxford University Press, v. 28, n. 1, p. 27–30, 2000. 39, 40

KANTHAJE, S.; BAIKUNJE, N.; KANDAL, I.; RATNACARAM, C. K. Repertoires of microrna-30 family as gate-keepers in lung cancer. *Frontiers in Bioscience-Scholar*, IMR Press, v. 13, n. 2, p. 141–156, 2021. 104, 129

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. 52

KASS, G. V. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 29, n. 2, p. 119–127, 1980. 29

KAUR, B.; KUMAR, S.; KAUSHIK, B. K. Recent advancements in optical biosensors for cancer detection. *Biosensors and Bioelectronics*, Elsevier, v. 197, p. 113805, 2022. 7

KAWAGUCHI, T. *et al.* Japanese ethnicity compared with caucasian ethnicity and never-smoking status are independent favorable prognostic factors for overall survival in non-small cell lung cancer: a collaborative epidemiologic study of the national hospital organization study group for lung cancer (nhsglc) in japan and a southern california regional cancer registry databases. *Journal of Thoracic Oncology*, Elsevier, v. 5, n. 7, p. 1001–1010, 2010. 6

KENT, O.; MENDELL, J. A small piece in the cancer puzzle: micrornas as tumor suppressors and oncogenes. *Oncogene*, Nature Publishing Group, v. 25, n. 46, p. 6188–6196, 2006. 20

KERN, F. *et al.* mieaa 2.0: integrating multi-species microrna enrichment analysis and workflow management systems. *Nucleic acids research*, Oxford University Press, v. 48, n. W1, p. W521–W528, 2020. 39

KIM, H.; LOH, W.-Y. Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, Taylor & Francis, v. 96, n. 454, p. 589–604, 2001. 29

KIM, V. N.; NAM, J.-W. Genomics of microrna. *TRENDS in Genetics*, Elsevier, v. 22, n. 3, p. 165–173, 2006. 17, 18, 20

KINGSFORD, C.; SALZBERG, S. L. What are decision trees? *Nature biotechnology*, Nature Publishing Group US New York, v. 26, n. 9, p. 1011–1013, 2008. 30

KLECZKO, E. K.; KWAK, J. W.; SCHENK, E. L.; NEMENOFF, R. A. Targeting the complement pathway as a therapeutic strategy in lung cancer. *Frontiers in immunology*, Frontiers Media SA, v. 10, p. 954, 2019. 9

KOBAYASHI, S. *et al.* Egfr mutation and resistance of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, Mass Medical Soc, v. 352, n. 8, p. 786–792, 2005. 12

KOTHANDAN, R. Handling class imbalance problem in mirna dataset associated with cancer. *Bioinformation*, Biomedical Informatics Publishing Group, v. 11, n. 1, p. 6, 2015. 46

KOZOMARA, A.; BIRGAOANU, M.; GRIFFITHS-JONES, S. mirbase: from microrna sequences to function. *Nucleic acids research*, Oxford University Press, v. 47, n. D1, p. D155–D162, 2019. 20

KRIS, M. G. *et al.* Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. *Jama*, American Medical Association, v. 311, n. 19, p. 1998–2006, 2014. 21

KUBAT, M.; BRATKO, I.; MICHALSKI, R. S. A review of machine learning methods. *Machine learning and data mining: methods and applications*, Citeseer, p. 3–69, 1998. 47

KUBAT, M.; KUBAT, J. *An introduction to machine learning*. [S.l.]: Springer, 2017. v. 2. 45

KUHN, M. *caret: Classification and Regression Training*. [S.l.], 2022. R package version 6.0-93. Disponível em: <https://CRAN.R-project.org/package=caret>. 64

KUHN, M.; VAUGHAN, D.; HVITFELDT, E. *yardstick: Tidy Characterizations of Model Performance*. [S.l.], 2022. R package version 1.1.0. Disponível em: <https://CRAN.R-project.org/package=yardstick>. 65

KUO, W.-T. *et al.* Bioinformatic interrogation of 5p-arm and 3p-arm specific mirna expression using tcga datasets. *Journal of clinical medicine*, MDPI, v. 4, n. 9, p. 1798–1814, 2015. 34

LANDER, E.; LINTON, L.; BIRREN, B. Initial sequencing and analysis of the human genome [published correction appears in nature. 2001; 411 (6838): 720]. *Nature*, v. 409, n. 6822, p. 860–921, 2001. 12

LANE, D. P.; CRAWFORD, L. V. T antigen is bound to a host protein in sy40-transformed cells. *Nature*, Nature Publishing Group UK London, v. 278, n. 5701, p. 261–263, 1979. 21

LEBANONY, D. *et al.* Diagnostic assay based on hsa-mir-205 expression distinguishes squamous from nonsquamous non–small-cell lung carcinoma. *Journal of clinical oncology*, American Society of Clinical Oncology, v. 27, n. 12, p. 2030–2037, 2009. 128, 149

LEE, R. C.; FEINBAUM, R. L.; AMBROS, V. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, v. 75, n. 5, p. 843–854, 1993. ISSN 0092-8674. Disponível em: <https://www.sciencedirect.com/science/article/pii/009286749390529Y>. 18

LEE, Y.-h.; BANG, H.; KIM, D. J. How to establish clinical prediction models. *Endocrinology and Metabolism*, Korean Endocrine Society, v. 31, n. 1, p. 38–44, 2016. 41

LI, L. *et al.* Microrna biomarker hsa-mir-195-5p for detecting the risk of lung cancer. *International Journal of Genomics*, Hindawi, v. 2020, 2020. 104, 148

LI, Y. *et al.* Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome biology*, Springer, v. 23, n. 1, p. 79, 2022. 24, 25, 79

LIANG, R. *et al.* Snhg6 functions as a competing endogenous rna to regulate e2f7 expression by sponging mir-26a-5p in lung adenocarcinoma. *Biomedicine & Pharmacotherapy*, Elsevier, v. 107, p. 1434–1446, 2018. 148

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002. Disponível em: <https://CRAN.R-project.org/doc/Rnews/>. 61, 62

LIM, E. L. *et al.* Microrna expression-based model indicates event-free survival in pediatric acute myeloid leukemia. *Journal of Clinical Oncology*, American Society of Clinical Oncology, v. 35, n. 35, p. 3964, 2017. 25

LIM, H. J.; CROWE, P.; YANG, J.-L. Current clinical regulation of pi3k/pten/akt/mtor signalling in treatment of human cancer. *Journal of cancer research and clinical oncology*, Springer, v. 141, p. 671–689, 2015. 129

LIN, T.-C. *et al.* Microrna-184 deregulated by the microrna-21 promotes tumor malignancy and poor outcomes in non-small cell lung cancer via targeting cdc25a and c-myc. *Annals of surgical oncology*, Springer, v. 22, p. 1532–1539, 2015. 105

LIU, M.; ZHOU, K.; CAO, Y. Microrna-944 affects cell growth by targeting epha7 in non-small cell lung cancer. *International journal of molecular sciences*, MDPI, v. 17, n. 10, p. 1493, 2016. 149

LOH, W.-Y. Fifty years of classification and regression trees. *International Statistical Review*, Wiley Online Library, v. 82, n. 3, p. 329–348, 2014. 28

LOH, W.-Y.; SHIH, Y.-S. Split selection methods for classification trees. *Statistica sinica*, JSTOR, p. 815–840, 1997. 29

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, BioMed Central, v. 15, n. 12, p. 1–21, 2014. 24

LU, T. X.; ROTHENBERG, M. E. Microrna. *Journal of allergy and clinical immunology*, Elsevier, v. 141, n. 4, p. 1202–1207, 2018. 18

LU, Y. *et al.* Microrna profiling and prediction of recurrence/relapse-free survival in stage i lung cancer. *Carcinogenesis*, Oxford University Press, v. 33, n. 5, p. 1046–1054, 2012. 150

LÜCHTENBORG, M. *et al.* The effect of comorbidity on stage-specific survival in resected non-small cell lung cancer patients. *European journal of cancer*, Elsevier, v. 48, n. 18, p. 3386–3395, 2012. 6

LUNARDON, N.; MENARDI, G.; TORELLI, N. ROSE: a Package for Binary Imbalanced Learning. *R Journal*, v. 6, n. 1, p. 82–92, 2014. 64

LYNCH, T. J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, Mass Medical Soc, v. 350, n. 21, p. 2129–2139, 2004. 12

MA, J. *et al.* Bioinformatic analysis reveals an exosomal mirna-mrna network in colorectal cancer. *BMC medical genomics*, BioMed Central, v. 14, n. 1, p. 1–18, 2021. 80

MAK, D. W.; LI, S.; MINCHOM, A. Challenging the recalcitrant disease—developing molecularly driven treatments for small cell lung cancer. *European Journal of Cancer*, Elsevier, v. 119, p. 132–150, 2019. 11

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, JSTOR, p. 50–60, 1947. 25, 37

MANTEL, N. *et al.* Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, v. 50, n. 3, p. 163–170, 1966. 52

MARIAN, A. Sequencing your genome: what does it mean? *Methodist DeBakey cardiovascular journal*, Methodist DeBakey Heart & Vascular Center, v. 10, n. 1, p. 3, 2014. 14

MEISTER, G.; TUSCHL, T. Mechanisms of gene silencing by double-stranded rna. *Nature*, Nature Publishing Group, v. 431, n. 7006, p. 343–349, 2004. 17

MENCK, C. F.; SLUYS, M. *Genética molecular básica: dos genes ao genoma*. [S.l.]: Editora Guanabara Koogan: Rio de Janeiro, Brazil, 2017. 15

MESSENGER, R.; MANDELL, L. A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American statistical association*, Taylor & Francis, v. 67, n. 340, p. 768–772, 1972. 29

MICHIE, D.; SPIEGELHALTER, D. J.; TAYLOR, C. C.; CAMPBELL, J. *Machine learning, neural and statistical classification*. [S.l.]: Ellis Horwood, 1995. 43

MILLER, K. D. *et al.* Cancer statistics for the us hispanic/latino population, 2021. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 71, n. 6, p. 466–487, 2021. 4

MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-hill New York, 1997. 27

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003. 62

MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, Taylor & Francis, v. 58, n. 302, p. 415–434, 1963. 29

MUELLER, J. P.; MASSARON, L. *Machine learning for dummies*. [S.l.]: John Wiley & Sons, 2021. 43

MYERS, D. J.; WALLEN, J. M. Lung adenocarcinoma. In: *StatPearls [Internet]*. [S.l.]: StatPearls Publishing, 2022. 8

NESBITT, J. C. *et al.* Survival in early-stage non-small cell lung cancer. *The Annals of thoracic surgery*, Elsevier, v. 60, n. 2, p. 466–472, 1995. 11

NEVE, J. D.; THAS, O.; OTTOY, J.-P.; CLEMENT, L. An extension of the wilcoxon-mann-whitney test for analyzing rt-qpcr data. *Statistical Applications in Genetics and Molecular Biology*, De Gruyter, v. 12, n. 3, p. 333–346, 2013. 24

NYBERG, F. *et al.* A european validation study of smoking and environmental tobacco smoke exposure in nonsmoking lung cancer cases and controls. *Cancer Causes & Control*, Springer, v. 9, p. 173–182, 1998. 7

ÖBERG, M. *et al.* Worldwide burden of disease from exposure to second-hand smoke: a retrospective analysis of data from 192 countries. *The lancet*, Elsevier, v. 377, n. 9760, p. 139–146, 2011. 1, 6

O'BRIEN, J.; HAYDER, H.; ZAYED, Y.; PENG, C. Overview of microrna biogenesis, mechanisms of actions, and circulation. *Frontiers in endocrinology*, Frontiers Media SA, v. 9, p. 402, 2018. 17

OLIVEROS, J. C. Venny. an interactive tool for comparing lists with venn diagrams. *http://bioinfogp. cnb. csic. es/tools/venny/index. html*, 2007. 39

OMS, O. M. da S. *Estimativas Globais de Saúde 2020: Mortes por Causa, Idade, Sexo, por País e por Região, 2000-2019.* 2020. Https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death. Acesso em 12 de dezembro de 2022. 3

PAEZ, J. G. *et al.* Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, American Association for the Advancement of Science, v. 304, n. 5676, p. 1497–1500, 2004. 14

PARKIN, D. M.; BRAY, F.; FERLAY, J.; PISANI, P. Global cancer statistics, 2002. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 55, n. 2, p. 74–108, 2005. 7

PEARSON, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Taylor & Francis, v. 50, n. 302, p. 157–175, 1900. 35

PENG, Y.; CROCE, C. M. The role of micrornas in human cancer. *Signal transduction and targeted therapy*, Nature Publishing Group, v. 1, n. 1, p. 1–9, 2016. 24

PERNEGER, T. V. What's wrong with bonferroni adjustments. *Bmj*, British Medical Journal Publishing Group, v. 316, n. 7139, p. 1236–1238, 1998. 38

PETO, R. *et al.* Smoking, smoking cessation, and lung cancer in the uk since 1950: combination of national statistics with two case-control studies. *Bmj*, British Medical Journal Publishing Group, v. 321, n. 7257, p. 323–329, 2000. 6

PFEFFER, S. R.; YANG, C. H.; PFEFFER, L. M. The role of mir-21 in cancer. *Drug development research*, Wiley Online Library, v. 76, n. 6, p. 270–277, 2015. 22

PIKOR, L. A.; RAMNARINE, V. R.; LAM, S.; LAM, W. L. Genetic alterations defining nsclc subtypes and their therapeutic implications. *Lung cancer*, Elsevier, v. 82, n. 2, p. 179–189, 2013. 1, 8

PRATI, R. C. *Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos*. Tese (Doutorado) — Universidade de São Paulo, 2006. 65

PRATI, R. C.; FLACH, P. A. Roccer: An algorithm for rule learning based on roc analysis. In: *Ijcai*. [S.l.: s.n.], 2005. p. 823–828. 64

PROBST, P.; BOULESTEIX, A.-L.; BISCHL, B. Tunability: Importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, JMLR. org, v. 20, n. 1, p. 1934–1965, 2019. 62

QUINLAN, J. C4. 5: Programs for machine learning, san mateo 1993. *Google Scholar Google Scholar Digital Library Digital Library*, p. 1–302, 1993. 29

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, v. 1, p. 81–106, 1986. 29

QUINLAN, J. R. Decision trees and multivalued attributes. *Machine intelligence 11*, Oxford University Press, p. 305–318, 1988. 55

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>. 31, 32

REIS, P. P. *et al.* Circulating mir-16-5p, mir-92a-3p, and mir-451a in plasma from lung cancer patients: potential application in early detection and a regulatory role in tumorigenesis pathways. *Cancers*, Multidisciplinary Digital Publishing Institute, v. 12, n. 8, p. 2071, 2020. 22

RELLI, V.; TREROTOLA, M.; GUERRA, E.; ALBERTI, S. Abandoning the notion of non-small cell lung cancer. *Trends in Molecular Medicine*, Elsevier, v. 25, n. 7, p. 585–594, 2019. 8

RESS, A. L. *et al.* Mir-96-5p influences cellular growth and is associated with poor survival in colorectal cancer patients. *Molecular carcinogenesis*, Wiley Online Library, v. 54, n. 11, p. 1442–1450, 2015. 104

RHODES, D. R.; CHINNAIYAN, A. M. Integrative analysis of the cancer transcriptome. *Nature genetics*, Nature Publishing Group US New York, v. 37, n. Suppl 6, p. S31–S37, 2005. 1

RITCHIE, M. E. *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, Oxford Academic, v. 43, n. 7, p. e47–e47, 2015. 24

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, Oxford University Press, v. 26, n. 1, p. 139–140, 2010. 24

ROSSUM, G. V.; JR, D.; L, F. The python language reference manual (version 2.5). *Network Theory Ltd*, 2006. 31

RUANO-RAVIÑA, A. *et al.* Lung cancer symptoms at diagnosis: results of a nationwide registry study. *ESMO open*, Elsevier, v. 5, n. 6, p. e001021, 2020. 11

RUDIN, C. M.; BRAMBILLA, E.; FAIVRE-FINN, C.; SAGE, J. Small-cell lung cancer. *Nature Reviews Disease Primers*, Nature Publishing Group UK London, v. 7, n. 1, p. 3, 2021. 8

SACHIDANANDAM, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, v. 409, n. 6822, p. 928–933, 2001. 13

SANDLER, D. P. *et al.* Indoor radon and lung cancer risk in connecticut and utah. *Journal of Toxicology and Environmental Health, Part A*, Taylor & Francis, v. 69, n. 7-8, p. 633–654, 2006. 7

SANTOS, G. da C.; SHEPHERD, F. A.; TSAO, M. S. Egfr mutations and lung cancer. *Annual Review of Pathology: Mechanisms of Disease*, Annual Reviews, v. 6, p. 49–69, 2011. 14

SARDO, F. L. *et al.* Yap/taz and ezh2 synergize to impair tumor suppressor activity of tgfbr2 in non-small cell lung cancer. *Cancer Letters*, Elsevier, v. 500, p. 51–63, 2021. 103

SARUMI, O. A.; LEUNG, C. K. Adaptive machine learning algorithm and analytics of big genomic data for gene prediction. *Tracking and preventing diseases with artificial intelligence*, Springer, p. 103–123, 2022. 23

SCHABATH, M. B.; COTE, M. L. Cancer progress and priorities: lung cancer. *Cancer epidemiology, biomarkers & prevention*, AACR, v. 28, n. 10, p. 1563–1579, 2019. 1

SCHMID, K.; KUWERT, T.; DREXLER, H. Radon in indoor spaces: an underestimated risk factor for lung cancer in environmental medicine. *Deutsches Arzteblatt international*, Deutscher Arzte-Verlag GmbH, v. 107, n. 11, p. 181, 2010. 7

SCHURCH, N. J. *et al.* How many biological replicates are needed in an rna-seq experiment and which differential expression tool should you use? *Rna*, Cold Spring Harbor Lab, v. 22, n. 6, p. 839–851, 2016. 80

SEIJO, L. M. *et al.* Biomarkers in lung cancer screening: achievements, promises, and challenges. *Journal of Thoracic Oncology*, Elsevier, v. 14, n. 3, p. 343–357, 2019. 20

SEYEDNASROLLAH, F.; LAIHO, A.; ELO, L. L. Comparison of software packages for detecting differential expression in rna-seq studies. *Briefings in bioinformatics*, Oxford University Press, v. 16, n. 1, p. 59–70, 2015. 80

SHANNON, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, Cold Spring Harbor Lab, v. 13, n. 11, p. 2498–2504, 2003. 40

SHAW, A. T.; SOLOMON, B. Targeting anaplastic lymphoma kinase in lung canceralk in lung cancer. *Clinical Cancer Research*, AACR, v. 17, n. 8, p. 2081–2086, 2011. 14

SHEN, J. *et al.* Novel insights into mir-944 in cancer. *Cancers*, MDPI, v. 14, n. 17, p. 4232, 2022. 149

SHENDURE, J. *et al.* Dna sequencing at 40: past, present and future. *Nature*, Nature Publishing Group UK London, v. 550, n. 7676, p. 345–353, 2017. 13

SHERAFATIAN, M.; ARJMAND, F. Decision tree-based classifiers for lung cancer diagnosis and subtyping using tcga mirna expression data. *Oncology letters*, Spandidos Publications, v. 18, n. 2, p. 2125–2131, 2019. 30

SIEGEL, R.; NAISHADHAM, D.; JEMAL, A. Cancer statistics, 2013. *CA: a cancer journal for clinicians*, v. 63, n. 1, p. 11–30, 2013. 8

SOUZA, V. G. *et al.* Identifying new contributors to brain metastasis in lung adenocarcinoma: A transcriptomic meta-analysis. *Cancers*, MDPI, v. 15, n. 18, p. 4526, 2023. 27

SPACKMAN, K. A. Signal detection theory: Valuable tools for evaluating inductive learning. In: ELSEVIER. *Proceedings of the sixth international workshop on Machine learning*. [S.l.], 1989. p. 160–163. 64

STRIMBU, K.; TAVEL, J. A. What are biomarkers? *Current Opinion in HIV and AIDS*, NIH Public Access, v. 5, n. 6, p. 463, 2010. 20

SUN, M. *et al.* Integrated analysis identifies microrna-195 as a suppressor of hippo-yap pathway in colorectal cancer. *Journal of hematology & oncology*, Springer, v. 10, p. 1–16, 2017. 104

SUN, S.; SCHILLER, J. H.; GAZDAR, A. F. Lung cancer in never smokers—a different disease. *Nature reviews cancer*, Nature Publishing Group UK London, v. 7, n. 10, p. 778–790, 2007. 1, 6

SUNG, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 71, n. 3, p. 209–249, 2021. 1, 4

ŚWITLIK, W. *et al.* mir-30a-5p together with mir-210-3p as a promising biomarker for non-small cell lung cancer: A preliminary study. *Cancer Biomarkers*, IOS Press, v. 21, n. 2, p. 479–488, 2018. 104

SZYMANSKI, M.; ERDMANN, V. A.; BARCISZEWSKI, J. Noncoding regulatory rnas database. *Nucleic acids research*, Oxford University Press, v. 31, n. 1, p. 429–431, 2003. 17

TAFT, R. J. *et al.* Nuclear-localized tiny rnas are associated with transcription initiation and splice sites in metazoans. *Nature structural & molecular biology*, Nature Publishing Group, v. 17, n. 8, p. 1030–1034, 2010. 17

TAKAMIZAWA, J. *et al.* Reduced expression of the let-7 micrornas in human lung cancers in association with shortened postoperative survival. *Cancer research*, AACR, v. 64, n. 11, p. 3753–3756, 2004. 22

TAN, M.; WU, J.; CAI, Y. Suppression of wnt signaling by the mir-29 family is mediated by demethylation of wif-1 in non-small-cell lung cancer. *Biochemical and biophysical research communications*, Elsevier, v. 438, n. 4, p. 673–679, 2013. 129

TARAZONA, S. *et al.* Noiseq: a rna-seq differential expression method robust for sequencing depth biases. *EMBnet. journal*, v. 17, n. B, p. 18–19, 2011. 25

THANDRA, K. C. *et al.* Epidemiology of lung cancer. *Contemporary Oncology/Współczesna Onkologia*, Termedia, v. 25, n. 1, p. 45–52, 2021. 6

THERNEAU, T. *A package for survival analysis in R (R package version 3.5-0)*. [S.l.]: Springer: New York, NY, USA, 2023. 50

THERNEAU, T.; ATKINSON, B. *rpart: Recursive Partitioning and Regression Trees*. [S.l.], 2022. R package version 4.1.19. Disponível em: <https://CRAN.R-project.org/package=rpart>. 48, 50, 56, 62

THERNEAU, T. M.; ATKINSON, E. J. *et al. An introduction to recursive partitioning using the RPART routines*. [S.l.], 1997. 56

THUN, M. J. *et al.* Lung cancer death rates in lifelong nonsmokers. *Journal of the National Cancer Institute*, Oxford University Press, v. 98, n. 10, p. 691–699, 2006. 7

TIAN, S. Classification and survival prediction for early-stage lung adenocarcinoma and squamous cell carcinoma patients. *Oncology letters*, Spandidos Publications, v. 14, n. 5, p. 5464–5470, 2017. 9

TIMOFEEV, R. Classification and regression trees (cart) theory and applications. *Humboldt University, Berlin*, v. 54, 2004. 48, 49, 50

TOUSSAINT, G. Bibliography on estimation of misclassification. *IEEE Transactions on information Theory*, IEEE, v. 20, n. 4, p. 472–479, 1974. 43

TRAVIS, W. D. Lung cancer pathology: current concepts. *Clinics in chest medicine*, Elsevier, v. 41, n. 1, p. 67–85, 2020. 9

TRAVIS, W. D.; BRAMBILLA, E.; RIELY, G. J. New pathologic classification of lung cancer: relevance for clinical practice and clinical trials. *J Clin Oncol*, v. 31, n. 8, p. 992–1001, 2013. 7

VLACHOS, I. S. *et al.* Diana-mirpath v3. 0: deciphering microrna function with experimental support. *Nucleic acids research*, Oxford University Press, v. 43, n. W1, p. W460–W466, 2015. 39

VYKOUKAL, J. *et al.* Contributions of circulating micrornas for early detection of lung cancer. *Cancers*, MDPI, v. 14, n. 17, p. 4221, 2022. 22

WANG, W. *et al.* Early detection of non-small cell lung cancer by using a 12-microrna panel and a nomogram for assistant diagnosis. *Frontiers in Oncology*, Frontiers Media SA, v. 10, p. 855, 2020. 22

WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, Nature Publishing Group UK London, v. 10, n. 1, p. 57–63, 2009. 15

WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, Nature Publishing Group UK London, v. 171, n. 4356, p. 737–738, 1953. 12

WEISSTEIN, E. W. Bonferroni correction. *https://mathworld. wolfram. com/*, Wolfram Research, Inc., 2004. 38

WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: <https://ggplot2.tidyverse.org>. 48

WICKHAM, H.; FRANçOIS, R.; HENRY, L.; MüLLER, K. *dplyr: A Grammar of Data Manipulation*. [S.l.], 2022. R package version 1.0.9. Disponível em: <https://CRAN.R-project.org/package=dplyr>. 64

WILCOXON, F. *Individual comparisons by ranking methods. Biom. Bull., 1, 80–83*. 1945. 37

WILKERSON, M. D. *et al.* Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. *PloS one*, Public Library of Science San Francisco, USA, v. 7, n. 5, p. e36530, 2012. 9

WILLIAMS, G. J. *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer, 2011. (Use R!). Disponível em: <https://rd.springer.com/book/10.1007/978-1-4419-9890-3>. 48

WITTEN, I. H.; FRANK, E.; MARK, A. *Hall, and Christopher J Pal. Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. 43

WOOSTER, R. *et al.* Identification of the breast cancer susceptibility gene brca2. *Nature*, Nature Publishing Group UK London, v. 378, n. 6559, p. 789–792, 1995. 21

WU, K.-L. *et al.* The roles of microrna in lung cancer. *International journal of molecular sciences*, MDPI, v. 20, n. 7, p. 1611, 2019. 20

WU, W. *et al.* Lncrna dleu2 accelerates the tumorigenesis and invasion of non–small cell lung cancer by sponging mir-30a-5p. *Journal of cellular and molecular medicine*, Wiley Online Library, v. 24, n. 1, p. 441–450, 2020. 105

XING, P.-Y. *et al.* What are the clinical symptoms and physical signs for non-small cell lung cancer before diagnosis is made? a nation-wide multicenter 10-year retrospective study in china. *Cancer medicine*, Wiley Online Library, v. 8, n. 8, p. 4055–4069, 2019. 8

YANG, X. *et al.* Temporal trends of the lung cancer mortality attributable to smoking from 1990 to 2017: a global, regional and national analysis. *Lung Cancer*, Elsevier, v. 152, p. 49–57, 2021. 1

YANG, Y. *et al.* Machine learning application in personalised lung cancer recurrence and survivability prediction. *Computational and Structural Biotechnology Journal*, Elsevier, v. 20, p. 1811–1820, 2022. 30

YU, I. T. *et al.* Dose-response relationship between cooking fumes exposures and lung cancer among chinese nonsmoking women. *Cancer research*, AACR, v. 66, n. 9, p. 4961–4967, 2006. 7

ZAMAY, T. N. *et al.* Current and prospective protein biomarkers of lung cancer. *Cancers*, MDPI, v. 9, n. 11, p. 155, 2017. 4, 12

ZAPPA, C.; MOUSA, S. A. Non-small cell lung cancer: current treatment and future advances. *Translational lung cancer research*, AME Publications, v. 5, n. 3, p. 288, 2016. 8

ZEILEIS, A.; HOTHORN, T.; HORNIK, K. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 17, n. 2, p. 492–514, 2008. 29

ZHANG, H. *et al.* Plasma mir-145, mir-20a, mir-21 and mir-223 as novel biomarkers for screening early-stage non-small cell lung cancer. *Oncology letters*, Spandidos Publications, v. 13, n. 2, p. 669–676, 2017. 79

ZHANG, Y.; YANG, Q.; WANG, S. Micrornas: a new key in lung cancer. *Cancer chemotherapy and pharmacology*, Springer, v. 74, p. 1105–1111, 2014. 11

ZONG, F.-Y. *et al.* The rna-binding protein qki suppresses cancer-associated aberrant splicing. *PLoS genetics*, Public Library of Science San Francisco, USA, v. 10, n. 4, p. e1004289, 2014. 104

6 Material Suplementar