UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
CENTRO DE BIOTECNOLOGIA DA UFRGS
CURSO DE GRADUAÇÃO EM BIOTECNOLOGIA

Abordagens de *Machine Learning* para análise de dados multiômicos de câncer de pulmão

Juliana Gabriela Passinato Coelho

Porto Alegre, 2024

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Centro de Biotecnologia da UFRGS
Departamento de Biologia Molecular e Biotecnologia

Juliana Gabriela Passinato Coelho

Abordagens de *Machine Learning* para análise de dados multiômicos de câncer de pulmão

Trabalho de Conclusão de Curso submetido à Universidade Federal do Rio Grande do Sul como parte dos requisitos necessários para a obtenção do Grau em Bacharel em Biotecnologia com ênfase em Bioinformática.

Orientador: Prof. Dr. Márcio Dorn
Coorientadora: Dra. Joice de Faria Poloni

Porto Alegre
2024

Juliana Gabriela Passinato Coelho

Abordagens de *Machine Learning* para análise de dados multiômicos de câncer de pulmão

Trabalho de Conclusão de Curso submetido à Universidade Federal do Rio Grande do Sul como parte dos requisitos necessários para a obtenção do Grau em Bacharel em Biotecnologia com ênfase em Bioinformática.

Prof. Dr. Márcio Dorn
Orientador

Dra. Joice de Faria Poloni
Coorientadora

Prof. Dr. Guido Lenz
Membro da Banca 1

Prof. Dr. Daniel Pens Gelain
Membro da Banca 2

Porto Alegre
2024

Resumo

Desde o Projeto Genoma Humano até hoje, a tecnologia e a capacidade computacional de processamento de dados biológicos se desenvolveu, possibilitando o processamento e facilitando o compartilhamento de uma grande quantidade de dados. Além do citado, as tecnologias de sequenciamento de DNA evoluíram e melhoram sua eficiência, e, como consequência, os custos foram reduzidos. Tendo isso em vista, ocorreu a necessidade de novos métodos cada vez mais eficientes para obtenção de dados biológicos e ferramentas para analisar as informações, agora abundantes. Deste modo, diversos conjuntos de dados biológicos estão disponíveis publicamente, podendo ser analisados por por múltiplos pesquisadores e diferentes abordagens. O câncer de pulmão é o tipo de câncer que mais incorre em morte, sendo que seu diagnóstico ocorre nos estágios mais avançados da doença em 75% dos casos, prejudicando o prognóstico. Os métodos de diagnóstico mais utilizados também não garantem acurácia para a identificação precoce, e, por vezes, levam pacientes a exposição a radiação e métodos invasivos desnecessariamente, além de terem altos custos. O câncer é uma doença complexa, envolvendo a desregulação de moléculas a níveis genômicos, transcriptômicos, proteicos e metabolômicos. Assim, há vantagem em analisar essa patologia de forma multiômica, ou seja, integrando as ômicas a fim de obter biomarcadores que consideram a complexidade do câncer de pulmão. Os biomarcadores são moléculas ou processos biológicos que são utilizados com propósito de diagnóstico, predição de risco, estadiamento, prognóstico, predição de resposta ao tratamento, seleção de tratamento, entre outros. Assim, a finalidade da análise realizada no presente trabalho, são os biomarcadores, ou seja, as características biológicas dentre o conjunto de dados que permite a predição da classificação de uma amostra entre a condição de amostra de câncer de pulmão ou tecido normal. Nesse sentido, um conjunto de dados RNA-seq proveniente de tecidos de câncer pulmão e de tecidos saudáveis adjacentes (dados pareados do mesmo indivíduo) foi submetido ao treinamento e teste com abordagens de machine learning. Usando abordagens de *machine learning* (*Random forest* e *Support Vector Machine*), os dados de transcritômica (expressão gênica) e genômica (SNPs) foram analisados independentemente e ambos os resultados foram considerados, buscando identificar processos, genes e mutações - biomarcadores - para propósito de diagnóstico do câncer de pulmão. Os genes selecionados e os processos bioquímicos associados a eles, na análise realizada com os dados de expressão gênica, mostraram-se em sua maioria associados ao câncer na literatura. De outra forma, as mutações selecionadas foram identificadas como ainda bastante desconhecidas no meio científico. Apesar disso, possíveis biomarcadores destacaram-se por estarem presentes na intersecção dos resultados para ambas as análises realizadas.

Palavras-chaves: Multiômica; Expressão gênica; SNPs; *Machine learning*; Biomarcadores; Adenocarcinoma de pulmão.

Abstract

From the Human Genome Project until today, the technology and computational capacity for processing biological data have developed, enabling the processing and facilitating the sharing of a large amount of data. In addition to the aforementioned, DNA sequencing technologies have evolved and improved their efficiency, leading to a reduction in costs. With this in mind, there arose the need for increasingly efficient methods to obtain biological data and tools to analyze the now abundant information. In this way, various sets of biological data are publicly available, capable of being analyzed by multiple researchers using different approaches. Lung cancer is the most deadly type of cancer, with 75% of cases being diagnosed in the advanced stages of the disease, impacting prognosis. The most commonly used diagnostic methods also do not ensure accuracy for early identification and, at times, subject patients to unnecessary radiation exposure and invasive procedures, in addition to having high costs. Cancer is a complex disease involving the dysregulation of molecules at genomic, transcriptomic, proteomic, and metabolomic levels. Therefore, there is an advantage in analyzing this pathology in a multiomic way, integrating omics to obtain biomarkers that consider the complexity of lung cancer. Biomarkers are molecules or biological processes used for purposes such as diagnosis, risk prediction, staging, prognosis, prediction of treatment response, treatment selection, among others. The purpose of the analysis in this study is biomarkers, i.e., the biological characteristics within the dataset that enable the prediction of the classification of a sample as either lung cancer or normal tissue. In this regard, an RNA-seq dataset from lung cancer tissues and adjacent healthy tissues (paired data from the same individual) was subjected to training and testing using machine learning approaches. Using machine learning techniques (Random Forest and Support Vector Machine), transcriptomic (gene expression) and genomic (SNPs) data were independently analyzed, and both results were considered to identify processes, genes, and mutations - biomarkers - for lung cancer diagnosis. The selected genes and the associated biochemical processes, in the analysis using gene expression data, were mostly found to be associated with cancer in the literature. On the other hand, the selected mutations were identified as largely unknown in the scientific community. Nevertheless, potential biomarkers stood out for being present at the intersection of the results for both analyses conducted. **Keywords**: Multiomics; Gene expression; SNPs; Machine learning; Biomarkers; Lung adenocarcinoma.

Sumário

1 Referencial Teórico

1.1 Sequenciamento de Nova Geração

A primeira geração de sequenciadores foi feita a partir da automatização de um processo chamado de sequenciamento de Sanger, que foi publicado em 1977 (PERVEZ et al., 2022). O destaque desta técnica é a inclusão de dideoxinucleotídeo, que faz com que o elongamento da fita pare quando utilizados pela polimerase. Sendo assim, existirá diversidade no tamanho das cópias, permitindo a compreensão de todo o fragmento. As moléculas são separadas por tamanho (com a resolução de uma base), durante a eletroforese capilar. Nela há um campo elétrico, logo, as moléculas de DNA (que tem carga negativa) são atraídas para o final do capilar, que é o eletrodo positivo. As menores moléculas chegam antes e são detectadas pela máquina com um laser, devido a cor fluorescente que é associada com cada dideoxinucleotídeo.

O procedimento descrito é considerado o padrão ouro para confirmação de sequências de DNA (sua acurácia é alta, podendo chegar a 99,99% (PERVEZ et al., 2022)) e também é amplamente utilizado no resequenciamento de alvos específicos em pesquisas e laboratórios clínicos (ALEKSEYEV et al., 2018), todavia, pode-se dizer que essa tecnologia de sequenciamento é cara e consome muito tempo.

Desde o início do PGH até hoje, a tecnologia e a capacidade computacional de processamento de dados se desenvolveu, possibilitando o processamento e facilitando o compartilhamento de uma grande quantidade de dados. Além disso, ocorreu a necessidade de desenvolvimento de novos métodos de análise cada vez mais eficientes para obtenção de dados biológicos e ferramentas para analisar as informações, agora abundantes. Tendo isso em vista, as tecnologias de sequenciamento de DNA evoluíram e melhoram sua eficiência, e, como consequência, os custos foram reduzidos (Figura 1).

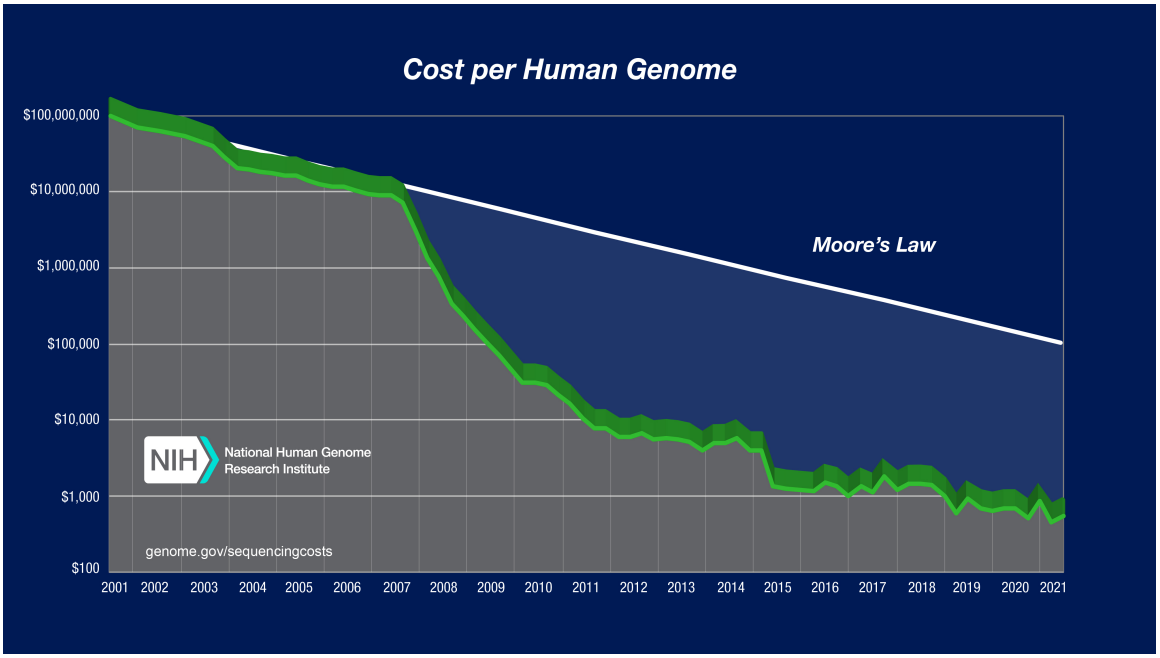

Figura 1 – O custo de sequenciamento de um genoma humano ao longo dos anos diminuiu e atualmente custa por volta de mil dólares. (Fonte: Wetterstrand (2021))

A tecnologia de sequenciamento de nova geração (NGS) são conhecidas como "de alto rendimento", do inglês "high-throughput", devido a sua capacidade de sequenciar diversos fragmentos de DNA em paralelo. Outras vantagens que o NGS apresenta sobre a primeira geração de sequenciamento são os custos menores e maior agilidade por apresentar ciclos de sequenciamento mais rápidos (PERVEZ et al., 2022). Para comparação, a velocidade de sequenciamento por Sanger (ABI 3730xl) é de aproximadamente 32kb/h, enquando um equipamento de NGS (1G genome analyzer from Illumina, Inc.) é de aproximadamente 13.800kb/h (MOROZOVA; MARRA, 2008).

Dentre os métodos de nova geração, existem diferentes abordagens de sequenciamento. Por exemplo, existe o sequenciamento pela síntese e o pela ligação, cujos métodos são pensados para gerar reads de 50 pares de bases (pb) a mais ao menos 1000 pb (GOODWIN; MCPHERSON; MCCOMBIE, 2016). Exsitem ainda outras tecnologias projetadas para gerar reads maiores, como é o caso do Oxford Nanopore MK 1 MinION, que chega a mais de 200 mil pares de base (GOODWIN; MCPHERSON; MCCOMBIE, 2016).

É possível utilizar o sequencimento de DNA para fins diversos. O primeiro uso dele foi o sequenciamento e a montagem de genomas parciais ou completos, que era o maior objetivo por trás da invenção das tecnologias de sequenciamento (SHENDURE et al., 2017). O primeiro genoma a ser sequenciado foi do bacteriófago MS2 em 1976, seguido pelo sequenciamento do genoma do bacteriófago ϕX174, com 4 e 10 genes, respectivamente (KOONIN, 2003). Entretanto, o primeiro genoma relativamente grande a ser sequenciado foi o do bacteriófago λ em 1982, com 48.502 pares de bases (pb) e aproximadamente 70 genes codificantes de proteína (KOONIN, 2003).

Depois da montagem de um genoma completo, como concretizado pelo PGH, outros genomas humanos puderam ser sequenciados para descoberta de variantes, muitas vezes provendo aplicações clínicas (SHENDURE et al., 2017). Atualmente, outros objetivos complexos também são buscados, como a metagenômica (SHENDURE et al., 2017).

A tecnologia de sequênciamento de NGS de alto rendimento revolucionou a área da transcritômica também, possibilitando a análise de RNA através de uma cópia complementar feita de DNA (KUKURBA; MONTGOMERY, 2015). Nesta área muitos estudos utilizam os dados para quantificar a expressão gênica a partir de todos os transcritos de uma amostra, o que não era possível com outros métodos, que mediam a expressão de apenas um transcrito, como o northern blot (KUKURBA; MONTGOMERY, 2015). O RNA-seq também pode ser aplicado para descoberta de estruturas gênicas, isoformas de *splicing* alternativo e expressão alelo-específica (KUKURBA; MONTGOMERY, 2015).

O uso de RNA-seq para a chamada de variantes não é utilizada com tanta frequência devido aos desafios existentes. Os transcritos maduros passam pelo processo de splicing e edição de RNA, que consistem de algumas alterações nucleotídicas que são inseridas ao nível de transcrito e que não existem no nível DNA (JEHL et al., 2021). A maior dificuldade para a chamada de variantes usando dados de RNA-seq é a diferença de cobertura entre os diferentes transcritos, devido a expressão gênica diferencial (JEHL et al., 2021). No entanto, um estudo mostrou que é possível atingir excelentes resultados com dados de RNA-seq, chegando a encontrar 91% dos SNPs encontrados por DNA-seq (JEHL et al., 2021). Eles utilizaram 767 RNA-seq de 382 aves de 11 populações e avaliaram as consequências dos 9.496.283 SNPs encontrados para uma das populações e a análise resultou em 25.344 transcritos que permitiram a predição de fortes impactos na função gênica (missenses deleterios, alteração de regiões de splicing, ganho de códon de parada e outros), correspondentes a 14.496 SNPs e 67,58 genes (JEHL et al., 2021).

1.2 Câncer

1.2.1 Características do Câncer

No câncer, as células precisam adquirir características específicas, se diferenciando de uma célula normal e adquirindo caráter patológico. Portanto, o câncer pode ser identificado por demonstrar uma série de atributos contidos nesse grupo de patologias. Conforme Hanahan e Weinberg (2011), Weinberg, Hanahan et al. (2000), as propriedades dessas células são:

1. Autosuficiência para sinalização celular de crescimento: As células tumorais podem sustentar seu crescimento e proliferação de três formas: produzindo seus próprios fatores de crescimento, requisitar fatores de crescimento dos tecidos de suporte associados aos tumores ou apresentar níveis aumentados de receptores destes fatores de crescimento.

2. Evitar supressores de crescimento: A proliferação celular resiste, apesar da sinalização de supressão existir. Desta forma, a célula se torna insensível aos supressores de crescimento.

3. Resistência à morte celular programada (apoptose): A apoptose é um processo natural que elimina células defeituosas ou indesejadas, e as células tumorais podem se tornar resistentes a este processo, evadindo dos mecanismos de indução à apoptose.

4. Obtenção da imortalidade replicativa: as células passam a se replicar ilimitadamente. Esse fato deve-se à ativação da capacidade de manutenção do comprimento dos telômeros.

5. Indução e acesso a vascularização: O câncer pode induzir a formação de novos vasos sanguíneos (angiogênese), que facilitam o acesso à nutrição e oxigênio.

6. Invasão e metástase: Habilidade de invadir tecidos distantes do sítio neoplásico primário, se estabelecer e proliferar.

7. Reprogramação do metabolismo: O metabolismo é alterado para processar a maior parte da glicose via glicólise, em vez de utilizar o processo mitocondrial. Indícios sugerem que utilizar a via da glicólise favorece a proliferação celular rápida.

8. Evasão do sistema imune: O sistema imune atua identificando e eliminando células identificadas como anormais, inclusive as de câncer. Entretanto, as células tumorais desenvolvem diversos mecanismos para evadir da resposta imune.

1.2.2 Desenvolvimento do Câncer

O câncer pode se desenvolver em qualquer idade e em qualquer célula do corpo, entretanto, todos os tipos de câncer adquirem gradualmente um acúmulo de mutações (WEINBERG; WEINBERG, 2006). Assim, o risco de morrer de câncer de cólon é quase 1000 vezes maior para um homem de 70 anos do que para um menino de 10 anos (WEINBERG; WEINBERG, 2006). Portanto, a formação do câncer é um processo que normalmente leva décadas (WEINBERG; WEINBERG, 2006). O processo da tumorigênese envolve alterações em quatro tipos de genes: ativação de oncogenes, inativação de genes supressores de tumor, evasão dos genes de apoptose e genes de reparação do DNA defeituosos (MALARKEY; HOENERHOFF; MARONPOT, 2013). Também há grande importância de que as mutações alterem o ciclo celular, impedindo os mecanismos de reparo do DNA (MALARKEY; HOENERHOFF; MARONPOT, 2013).

O desenvolvimento do câncer é desencadeado por múltiplas mudanças ao nível molecular, portanto é um processo que ocorre em múltiplas etapas (ARJMAND et al., 2022; COOPER, 2018). A carcinogênese começa com mutações genéticas ou alterações epigenéticas e essa etapa é chamada de iniciação. Os fatores necessários para a iniciação são divididos em intrínsecos, que são relativos a mutações de ordem aleatória no DNA durante a replicação, e os extrínsecos, que são o metabolismo, sistema imune, funcionamento dos hormônios, dieta, estilo de vida, exposição a químicos e/ou radiação (ARJMAND et al., 2022).

Alguns fatores extrínsecos não são mutagênicos e atuam, interagindo com vias de sinalização de receptores das células ou crescimento celular (promovendo a proliferação das células) e diferenciação e/ou apoptose e atuam primariamente alterando a expressão gênica (MENDELSOHN et al., 2014; COOPER, 2018). Sendo assim, essa fase é chamada de promoção e nela ocorrem mais mutações, devido a proliferação, durante a replicação do DNA (COOPER, 2018).

Esta fase é reversível, exceto se a célula adquire mutações suficientes para manter seu crescimento autônomo (MENDELSOHN et al., 2014). Um exemplo de agente promotor para o pulmão é o 2 e o 3-terc-butil-4-hidroxianisol e também certos componentes da fumaça de cigarro (MENDELSOHN et al., 2014). Os outros carcinógenos, chamados de genotóxicos, estão associados a iniciação, causando danos ao DNA (COOPER, 2018). Eles também estão presentes na fumaça de cigarro, por exemplo, moléculas como Benzo(a)pireno e Dimetilnitrosamina (COOPER, 2018).

A última etapa do câncer é chamada de progressão, o genoma se torna instável, levando a mais mutações (COOPER, 2018). As células que adquirem mutações que conferem capacidade de se multiplicar mais rápido se tornam as células dominates, por vantagem seletiva (COOPER, 2018). Assim, a massa de células torna-se grande o suficiente

para ser detectada, podendo ser um tumor benigno ou preneoplástico (MALARKEY; HOENERHOFF; MARONPOT, 2013).

Tumores benignos não invadem os tecidos ao redor e também não metastizam (RUDDON, 2007). É necessária uma regulação negativa da adesão celular, adquirir mobilidade e habilidade de invadir outros tecidos para ocorrer a metástase e esse processo é chamado de transição epitelial mesenquimal (TEM) (MENDELSOHN et al., 2014). A maior parte dos tumores passam pelo processo de TEM durante a progressão, tanto que canceres derivados do epitélio apresentam o TEM como processo determinante (RIBATTI; TAMMA; ANNESE, 2020). As células tumorais epiteliais perdem a polaridade, a adesão entre células e ganham propriedades invasivas e migratórias. O processo é modulado por vias bioquímicas complexas, envolvendo microRNAs, epigenética, reguladores pós-translacionais e eventos de *splicing* alternativos.

Demonstrou-se que via TGF-β/Smads é a mais forte na indução do TEM, regulando positivamente fatores de transcrição importantes para o processo (XU; LAMOUILLE; DERYNCK, 2009). No CPNPC, a perda da expressão de E-caderina é correlacionada com a metástase (KASE et al., 2000). Na figura 2, os marcadores do processo estão sumarizados.

EMT Markers
Increased proteins
N-cadherin
Vimentin
Fibronectin
Snail 1 (Snail)
Snail 2 (Slug)
Twist
FOX C2
SOX 10
MMP-2, MMP-3, MMP-9
N-cadherin
Decreased proteins
E-cadherin
Desmoplakin
Cytokeratin
Occludin
Functional markers
Increased migration
Increased invasion
Increased scattering
Elongation of cell shape
Resistance to anoikis

Figura 2 – Alguns marcadores do processo de transição epitelial mesenquimal (RIBATTI; TAMMA; ANNESE, 2020).

Como explicado anteriormente, os cânceres tem diversas características em comum, mas também existem especificidades que permitem a divisão em subgrupos de patologias, de acordo com seu tecido de origem, composição molecular, taxa de progressão e responsividade à terapia (MORDENTE et al., 2015; ARJMAND et al., 2022).

1.2.3 Câncer de Pulmão

O câncer de pulmão é o segundo câncer mais diagnosticado no mundo (FERLAY et al., 2021), atrás apenas de câncer de próstata (para os homens) e câncer de mama (para as mulheres) (THANDRA et al., 2021). Em 2018, estima-se que ocorreram por volta de 2 milhões de novos casos e 1.76 milhões de óbitos (contabilizando 18,4% de todas as mortes por câncer) (THANDRA et al., 2021). É o tipo de câncer com mais mortes relacionadas no mundo tanto para homens como para mulheres (THANDRA et al., 2021).

Em 37 nações, ele é o câncer com maior incidência, incluíndo países como Russia, China, Oriente Médio, Sudeste Asiático e países do Leste Europeu (THANDRA et al., 2021). Medidas dos sistema público de países com alta renda como países do leste europeu e Estados Unidos, onde a taxa de fumantes adultos é por volta de 20% (JEMAL et al., 2010), para diminuir a taxa de fumantes têm ajudado a diminuir a incidência de câncer de pulmão, entretanto, os diagnósticos da doença em país pobres continua a aumentar (THAI et al., 2021).

Conforme dados da Globocan (2018), o Brasil apresenta incidência de câncer de pulmão na faixa de 10 a 20% (todas as idades e sexos), assim como os outros países da América do Sul, que não passam de 20%. Países do Norte Europeu, Estados unidos e China estão na faixa de incidência de 30 a 40%, sendo que a máxima taxa é atingida pela Hungria em homens (77,4%) e a máxima taxa para ambos os sexos ocorre na Polinésia (52,2%). No entanto, para o Brasil, a taxa de mortalidade se aproxima bastante da taxa de incidência, sendo 12 e 13% (12 e 13 a cada 100.000 habitantes), respectivamente. Portanto ficando com uma taxa de mortalidade menor que outros países mais desenvolvidos como França, Estados Unidos e Alemanha, porém a diferença entre incidência e mortalidade é muito pequena indicando que a maior parte das pessoas com câncer de pulmão vêm à óbito no Brasil. O fato pode estar relacionado a desigualdade de acesso ao sistema de saúde, levando ao disgnóstico e tratamento tardio (THANDRA et al., 2021).

Segundo Antunes et al. (2008) e Chatenoud et al. (2010), é muito provável que no Brasil exista subnotificação de mortes, dificultando a comparação com outros países. O primeiro estudo, que considerou dados de 1995 a 2003, observou que a taxa de mortalidade estava relacionada positivamente com áreas mais ricas de São Paulo, com um provável fator de influência por estilo de vida das classes mais altas associado ao consumo de cigarro. No segundo estudo, que compreende dados de 1980 a 2004, as taxas de mortalidade por câncer de pulmão se mostraram mais baixas no Brasil comparado ao resto da América, não mostrando mudanças drásticas nas duas décadas observadas.

Aproximadamente 60 a 80% dos casos de câncer de pulmão ocorrem em pacientes que apresentam o habito de fumar (SCHABATH; COTE, 2019). Apesar disso, em torno de apenas 15% dos fumantes desenvolvem câncer de pulmão, apontando para a existência

de outros fatores de risco (SCHABATH; COTE, 2019). As mulheres apresentam risco aumentado para o desenvolvimento de câncer de pulmão devido à influências hormonais (células tumorais apresentam receptores de estrogênio, que atuam promovendo a proliferação celular e o hormônio também é capaz de formar adutos de DNA), reparação de DNA reduzida e, quando fumantes, apresentam níveis mais elevados de adutos no DNA em comparação aos homens (AKHTAR; BANSAL, 2017; RIVERA, 2009). Além disso, a idade avançada também é um indicativo do risco de câncer de pulmão (TOUMAZIS et al., 2020). Os casos de câncer de pulmão têm aumentado desde 1985 em 51%, sendo um aumento de 44% em homens e 76% para mulheres (AKHTAR; BANSAL, 2017). Conforme o artigo de Akhtar e Bansal (2017), são fatores de risco para o câncer de pulmão:

- A exposição à fumaça de cigarro é a causa principal do câncer de pulmão, pois ela contém aproximadamente 3500 substâncias carcinogênicas. No organismo essas substâncias passam por transformações por meio de enzimas e os produtos destes processos podem formar adutos no DNA que interagem com genes cruciais como KRAS e p53. Em adição, os radicais livres presentes na fumaça de cigarro podem formar espécies altamente reativas na pulmão e causar ruptura nas moléculas de DNA.

- Alguns fatores genéticos aumentam o risco de câncer. Variantes em genes envolvidos com o metabolismo, reparo por excisão de nucleotídeos e ciclo celular podem afetar bastante o risco de câncer de pulmão, já que essas vias tem um papel muito importante na doença.

- Doenças pulmonares prévias acarretam em inflamação e por isso, em danos aos tecidos, aumentando a divisão celular e taxa de mutações. A inflamação pode atuar como iniciador ou promotor no câncer com sinalização anti-apoptótica que pode levar a angiogênese. As doenças também podem obstruir o fluxo de ar levando as células a uma condição de hipóxia que ativa a glicólise e inibe a apoptose.

- Infecções virais estão bastante associadas com o risco de câncer, tais como o Papilomavírus (HPV), vírus da imunodeficiência humana (HIV) e vírus da herpes (EBV).

- Substâncias como arsênico, asbestos, berílio, cádmio, clorometil, éteres, cromio, níquel, radônio, silica e cloreto de vinila são classificados como carcinogênios. Arsênico e asbestos são conhecidas por causar espécies reativas de oxigênio e nitrogênio, além de o segundo causar também respostas inflamatórias, dano ao DNA e mutações. O decaímento do radônio é ionizante, causando danos cromossomais e mutações. A silica causando resposta inflamatória, o que aumenta o risco de lesões pré-neoplásticas.

- Estrogênio, um hormônio sexual feminino foi associado a carcinogênese induzida por benzopireno (componente da fumaça do cigarro) por meio de estresse oxidativo. A expressão e atividade de receptores de estrogênio se mostraram aumentadas em células de câncer de pulmão em relação a células epiteliais normais. Em câncer de pulmão o estrogênio ativa vias de sinalização de proliferação celular, angiogênese e progressão do tumor.
- A obesidade e alguns fatores da dieta estão associados com maior risco de câncer de pulmão, como o consumo de álcool, baixa concetração de betacaroteno e vitamina C.
- A poluição do ar também aumenta o risco de câncer de pulmão, aumentando a frequência de dados no DNA, mutações somáticas e germinativas e aberrações cromossomais.

Existem várias formas de classificar os estágios do câncer de pulmão (AMIN et al., 2017). Conforme o National Cancer Institute (NCI), o modelo mais utilizado é o TNM. O número após cada uma das letras tem um significado, sendo que T indica o tamanho do tumor principal, o N indica se linfonodos regionais apresentam invasão das células cancerígenas e o M indica se o tumor se espalhou para outras partes do corpo ou não. Essa forma de classificação fica ainda mais completa, podendo ter mais letras e números para cada caso específico, por isso é uma classificação bastante complexa (AMIN et al., 2017). As tabelas 1 e 2 mostram esse modelo de forma simplificada, conforme a oitava edição do manual de classificação do câncer, do *American Joint Committee on Cancer.*

Tamanho do Tumor (cm)	
TX	Tumor não pode ser avaliado, ou tumor comprovado mas não aparece nas imagens ou broncoscopia
T0	Sem evidência de tumor primário
Tis	Tumor in situ (quando o tumor está no seu local original - tumor não invasivo)
T1	$\leqslant 3$ (compreende as classificações T1, T1mi, T1a, T1c)
T2	>3 e $\leqslant 5$ (compreende as classificações T2, T2a, T2b)
T3	>5 e $\geqslant 7$ ou tumor de qualquer tamanho que invada órgãos ou tecidos específicos descritos no manual
T4	>7 ou tumor de qualquer tamanho que invada órgãos ou tecidos específicos descritos no manual

Tabela 1 – Descreve de forma simplificada o "T"do modelo TNM

Gânglios Linfáticos		Metástase	
NX	Linfonodos da região não podem ser avaliados	M0	Ausente
N0	Sem metástase regional nos linfonodos	M1 (m1a, m1b,m1c)	Presente
N1	linfonodo peribrônquicos ipsilaterais e/ou hilares ipsilaterais e intrapulmonares		
N2	linfonodos mediastinais ipsilaterais e/ou subcarinais		
N3	linfonodos contralaterais mediastinais ou hilares, ipsilaterais ou contralaterais escalenos ou linfonodos supraclaviculares		

Tabela 2 – Descreve de forma simplificada o "N"e o "M"do modelo TNM

Outra forma de classificação é a que se utiliza de números romanos seguidos por letras, que tem uma correspondência com o sistema TNM (Figura 3). Ademais, existem 5 palavras/expressões para descrever onde está o câncer, representada por (*In situ*) - células anormais estão presentes; (Localizado) - limitado ao local onde se iniciou; (Regional) se espalhou para linfonodos, tecidos ou órgãos; (Distante) se espalhou para partes distantes do corpo (NIH, 2022); (Desconhecido) Sem informações suficentes para classificar.

Para alguns cânceres são obtidos diagnósticos no estágio inicial da doença com maior frequência, porque exames de rotina são capazes de expor a condição. A premissa anterior expões uma realidade mais próxima do ideal, que é diferente do que ocorre nos casos de câncer de pulmão, onde o diagnóstico é obtido nos estágios III ou IV (que são os mais avançados da doença) em mais de 75% dos casos e esse fato prejudica muito o prognóstico, pois o tratamento tardio reduz as chances de sobrevivência (NOORELDEEN; BACH, 2021). A Figura 4 de Schabath e Cote (2019), ilustra o problema.

When T is...	And N is...	And M is...	Then the stage group is...
TX	N0	M0	Occult carcinoma
Tis	N0	M0	0
T1mi	N0	M0	IA1
T1a	N0	M0	IA1
T1a	N1	M0	IIB
T1a	N2	M0	IIIA
T1a	N3	M0	IIIB
T1b	N0	M0	IA2
T1b	N1	M0	IIB
T1b	N2	M0	IIIA
T1b	N3	M0	IIIB
T1c	N0	M0	IA3
T1c	N1	M0	IIB
T1c	N2	M0	IIIA
T1c	N3	M0	IIIB
T2a	N0	M0	IB
T2a	N1	M0	IIB
T2a	N2	M0	IIIA
T2a	N3	M0	IIIB
T2b	N0	M0	IIA
T2b	N1	M0	IIB
T2b	N2	M0	IIIA
T2b	N3	M0	IIIB
T3	N0	M0	IIB
T3	N1	M0	IIIA
T3	N2	M0	IIIB
T3	N3	M0	IIIC
T4	N0	M0	IIIA
T4	N1	M0	IIIA
T4	N2	M0	IIIB
T4	N3	M0	IIIC
Any T	Any N	M1	IV
Any T	Any N	M1a	IVA
Any T	Any N	M1b	IVA
Any T	Any N	M1c	IVB

Figura 3 – Correspondência entre o sistema de classificação TNM e o agrupamento conforme o estágio do câncer e seu prognóstico (oitava edição do manual de classificação do câncer do *American Joint Committee on Cancer)* (AMIN et al., 2017)

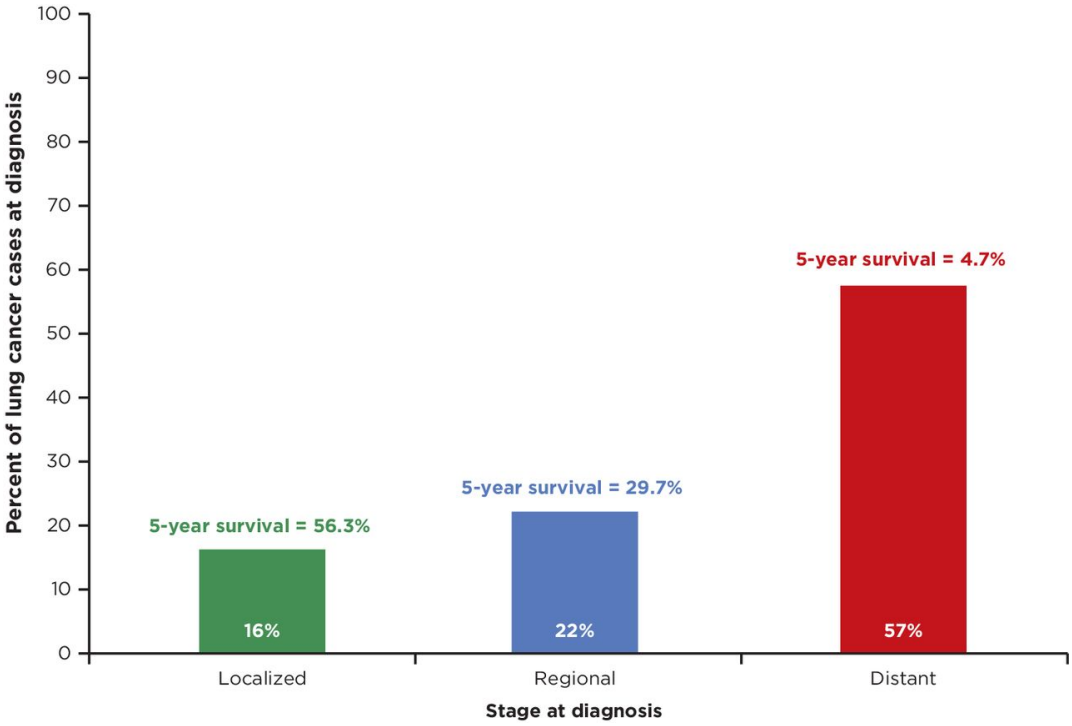

Figura 4 – Chance de sobrevivência nos 5 anos seguintes conforme o estágio de diagnóstico e porcentagem de diagnósticos feitos em cada estágio (Dados dos Estados Unidos da América). Os casos em que o estagiamento não foi possível classificiar contribuem para 4% dos diagnósticos e tem 8,2% de chances de sobrevivência pelos 5 anos decorrentes.

O câncer de pulmão é classificado em dois grandes grupos (conforme mostra a Figura 5): os cânceres de pulmão de pequenas células (CPPC) e os cânceres de pulmão de não pequenas células (CPNPC). O primeiro é responsável por 15%-20% dos casos (NOORELDEEN; BACH, 2021). O segundo que acomete o percentual restante dos casos, se subdivide em adenocarcinoma, carcinoma de célula grande e carcinoma de célula escamosa e estes ocorrem em 40%, 15% e 25% dos casos de CPNPC, respectivamente (SCHABATH; COTE, 2019).

Segundo Rodriguez-Canales, Parra-Cuentas e Wistuba (2016), o adenocarcinoma caracteriza-se por ser um tumor epitelial malígno com diferenciação glandular que pode produzir mucina, ou fator de transcrição da tireoide 1 ou expressão de marcadores de pneumócitos, tal como napsina A. Além disso, o adenocarcinoma normalmente localiza-se nas partes periféricas do pulmão. Já o carcinoma de célula escamosa normalmente apresenta uma localização central, como o bronco principal ou lobar e caracteriza-se por ser um tumor epitelial malígno que apresenta queratinização ou marcadores imunohistoquímicos de diferenciação de célula escamosa (p40, p63, e citoqueratinas 5/6). O carcinoma de célula grande não cumpre os requisitos para ser classificado como os dois descritos acima e nem como CPPC, e é isso que o caracteriza.

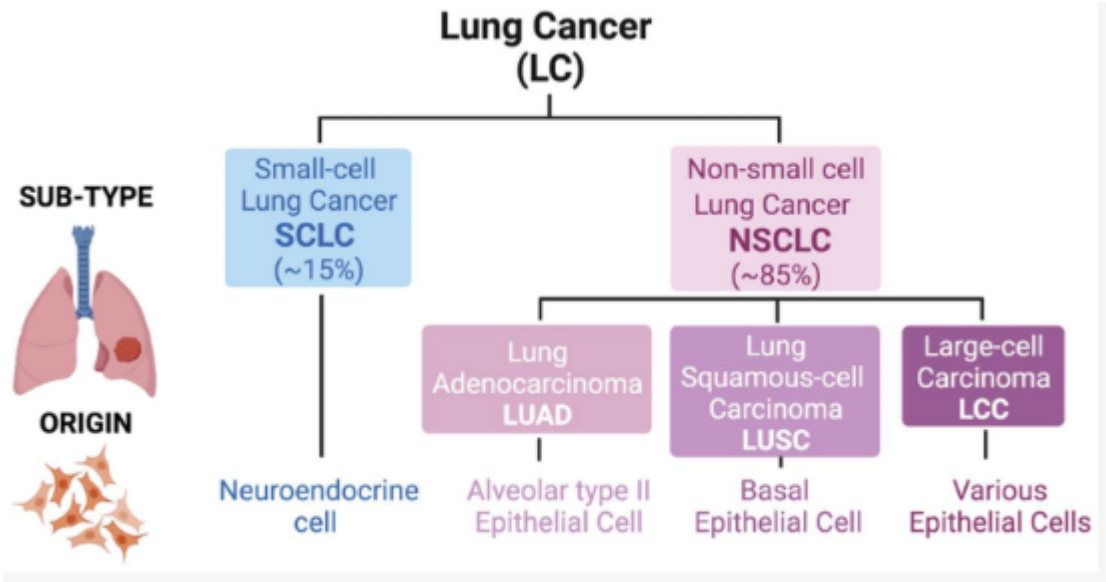

Figura 5 – Figura mostrando a divisão entre subtipos do câncer de pulmão (junto com a porcentagem que cabe a cada um) e o tipo celular de origem em cada subtipo (SÁNCHEZ-ORTEGA; CARRERA; GARRIDO, 2021).

Conforme a oitava edição do manual de classificação do câncer do *American Joint Committee on Cancer*, para o diagnóstico de câncer de pulmão existem vários tipos de exames de imagem e exames invasivos que devem ser feitos seguindo uma ordem específica, começando dos menos invasivos para os mais invasivos. O primeiro passo é examinação do paciente e coleta do histórico médico, seguido de raio-x do peito e exame de sangue. Com o raio-x espera-se revelar a quantidade e propagação de tumores, tamanho, localização, entre outras coisas mais específicas. A tomografia computadorizada (TC) após o raio-x tem o propósito de refinar as informações encontradas. A tomografia por emissão de pósitrons é útil para avaliar a progressão metástica, exceto no cérebro (onde é usada a ressonância magnética) e é indicada para aqueles que tem a intenção de começar a tratar o câncer e não encontraram nenhuma anormalidade ou metástase na TC (no sentido de eliminar a hipótese de metástase). Os exames citados provém informações muito úteis, entretanto, o diagnóstico não é dado por eles. Para a classificação TNM é necessária a confirmação microscópica da malignidade do tumor e confirmação do tipo histopatológico.

Um estudo, realizado pelo *The National Lung Screening Trial* (NLST), iniciado em 2002 e finalizado em 2009, acompanhou a triagem de pessoas que tinham histórico de consumo de cigarros de pelo menos 30-maços-ano (o que significa que a pessoa fumou o equivalente a 1 maço de 20 cigarros por dia durante 30 anos). A triagem foi feita a partir de exames de raios-X do tórax e da tomografia computadorizada de baixa dose (TCBD), com 53.454 participantes da faixa etária entre 55 e 74 anos. Os resultados do estudo indicam que o TCBD detecta mais canceres e nódulos no pulmão em relação ao raio-X e mostrou uma redução na taxa de mortalidade de 20% . Entretanto, ambos mostraram altas taxas de falsos positivos de 96,4% e 94,4% (TEAM, 2011).

A alta taxa de falsos positivos, muitas vezes, pode levar o paciente a mais exames, sendo que estes podem ser invasivos e cirúrgicos, adicionando riscos e complicações desnecessárias (KNIGHT et al., 2017). Soma-se a isso os riscos associados aos métodos de detecção que envolvem radiação, mesmo que baixos. Por exemplo, os índices que radiação de uma TCBD são muito mais altos que os recomendados (nos Estados Unidos) para uma mamografia em triagens (NANAVATY; ALVAREZ; ALBERTS, 2014).

Uma triagem em Milão que acompanhou 5203 pessoas assintomáticas mas com alto risco de desenvolverem câncer de pulmão, por 10 anos consecutivos. A TCBD era realizada anualmente e posterior PET scan era realizado para casos com descobertas suspeitas de câncer. O estudo observou que havia um risco adicional de 0.05% de desenvolvimento de câncer por causa da radiação. Quando comparado com cânceres de pulmão detectados por TC na mesma época, esse número é de 1 induzido por radiação a cada 100 detectados (RAMPINELLI et al., 2017).

Devido aos exames de imagem não proverem o dignóstico, outros exames são necessários, como a análise citológica do escarro, método não-invasivo onde se buscam células cancerosas para a realização do diagnóstico. Essa técnica geralmente falha em identificar adenocarcinomas pequenos ($\leqslant 2cm$), pois através do uso do escarro é mais fácil de identificar tipos de câncer cuja principal localidade de desenvolvimento sejam mais centrais no pulmão, como o carcinoma de pequenas células e de célula escamosa (NOORELDEEN; BACH, 2021). Esta abordagem também não é indicada para triagens pois sua capacidade de detecção de cânceres em estágio inicial é de apenas 20 - 30% (NOORELDEEN; BACH, 2021). Conforme um review com 29.145 pacientes os valores diagnósticos dessa técnica foram de 0,66 de sensibilidade, 0,99 de especificidade, falso positivo 8% e falso negativo 10% (AMIN et al., 2017).

Outro método de imagem usado é a broncoscopia, que é feita a partir de um aparelho com câmera que é inserido no aparelho respiratório do paciente. As imagens são analisadas na busca por lesões pré-malignas cujo tamanho é diminuto (menores que 1 mm) e por isso, é um método de difícil diagnóstico (NOORELDEEN; BACH, 2021). Conforme Amin et al. (2017), ele apresenta sensibilidade de 78% a 88% dependendo da localização do tumor. Por fim, o padrão ouro para a confirmação de câncer de pulmão é a biópsia de tecido, uma abordagem invasiva feita a partir de coleta dos mesmos, mas que tem sensibilidade de 0,9, especificidade de 0,97, falso positivo de 1% e falso negativo de 22% Amin et al. (2017). Segundo Nooreldeen e Bach (2021), pesquisas devem focar seus esforços em encontrar biomarcadores para o diagnóstico precoce do câncer de pulmão, de forma a aliviar o desconforto dos pacientes e os custos para os sistemas de saúde, já que os métodos e tecnologias atuais são caros.

1.2.4 CPNPC e Biomarcadores

Segundo o NIH, um biomarcador é uma característica que é medida objetivamente como um indicador de processos biológicos normais ou patogênicos ou respostas a intervenções terapêuticas. Biomarcadores podem ser anticorpos, DNA circulante, metilações no DNA, RNA de células do epitélio das vias aéreas, entre outros (SEIJO et al., 2019).

Conforme (KULASINGAM; DIAMANDIS, 2008; DUFFY, 2012), para que um biomarcador seja considerado adequado para a prática na clínica oncológica, ele deve:

- Ser produzido apenas por tumores;
- Ter correlação em níveis quantitativos com a carga tumoral (número total de células de câncer) e esse marcador deve ocorrer e ser passível de detecção entre período assintomático e o diagnóstico clínico;
- Presente em quantidades mensuráveis no estágio pré-clínico;
- Indetectável em indivíduos sem a doença ou com a doença em forma benigna;
- Mensurável em pequenas amostras e que elas demandem pouca preparação (para ter um teste confiável e custo-benefício);
- Apresentar alta especificidade e sensibilidade.

Os própositos dos biomarcadores no câncer são diversos, podendo ter capacidades de serem usados para (a) diagnóstico (screening, diagnóstico e diagnóstico precoce); (b) predição e estadiamento (predição, estadiamento e estratificação por risco); (c) prognóstico (recorrência, metástase e de prognóstico); (d) tratamento (seleção de tratamento por meio de biomarcadores, predição da resposta à terapia) (HOSEOK; CHO, 2015).

Já existem biomarcadores preditivos para terapias com alvos moleculares para o CPNPC em estágio avançado ou metástico. Buscas de mutações que ativem o gene *epidermal growth factor receptor* (EGFR) (aumentando a sinalização realizada pela proteína e conferindo sensibilidade aos medicamentos) são mandatórias antes da administração de inibidores anti-EGFR como erlotinib, gefitinib, afatinib, or osimertinib (DUFFY; O'BYRNE, 2018). Para predizer a resposta a crizotinib, testar se há rearranjos em ALK e ROS1 é necessário (DUFFY; O'BYRNE, 2018). Além de testar esses genes, também é indicado pelas direrizes atuais, que sejam testados BRAF Val600Glu, rearranjos em RET e mutações no exon 14 de MET, quando ocorrer diagnóstico em estágio avançado de adenocarcinoma de pulmão (KALEMKERIAN et al., 2018; LINDEMAN et al., 2018).

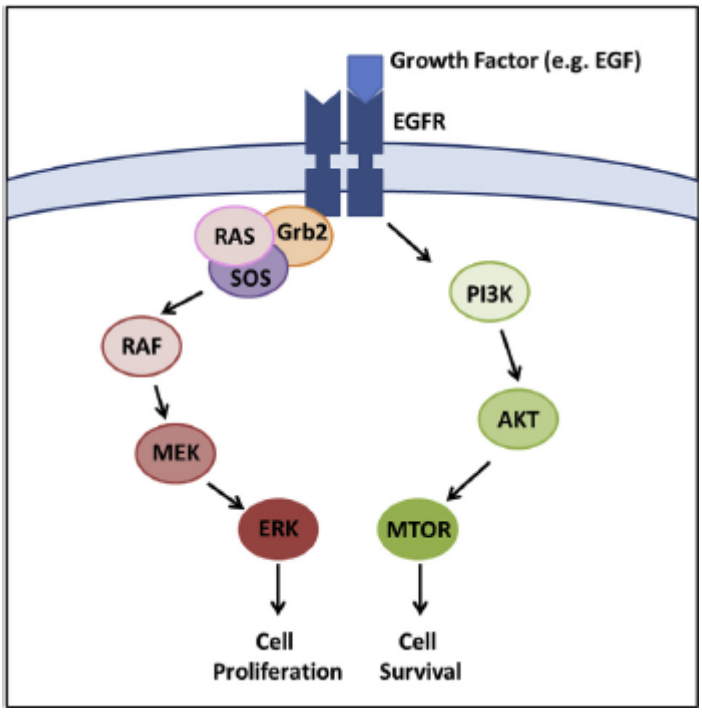

Figura 6 – Vias de sinalização de EGFR que promovem a proliferação e a sobrevivência celular (DUFFY; O'BYRNE, 2018)

Segundo Herbst, Morgensztern e Boshoff (2018) as mutações mais comuns encontradas no CPNPC estão presentes nos genes KRAS, EGFR e nos supressores de tumores TP53, KEAP1, STK11 e NF1. Na Figura 7 são mostradas alterações em componentes em vias que são importantes no adenocarcinoma de pulmão como a via de sinalização receptora de tirosina quinase, mTOR, resposta ao estresse oxidativo, proliferação e progressão do ciclo celular (HERBST; MORGENSZTERN; BOSHOFF, 2018). A frequência das alterações é a soma de mutações somáticas, deleções homozigotas, amplificações, e desregulação da expressão gênica.

Para traçar o perfil molecular do câncer de pulmão podem ser usados diferentes materiais biológicos, de forma não-invasiva: usando DNA livre circulante encontrado no plasma sanguíneo (sensibilidade $> 95\%$ para EGFR, KRAS ou BRAF) ou células circulantes de câncer (sensibilidade para KRAS e EGFR de 78% e 92% respectivamente) (CALVAYRAC et al., 2017). Existem vários métodos para isolar as células de tumor circulantes, como Cellsearch e *ISET (isolation by size of epithelial tumour cells)*.

Um dos desafios para o tratamento personalizado dos pacientes conforme os biomarcadores indicadores de resposta a drogas é a evolução do câncer, que passa a ter outro tipo de perfil mutacional devido à seleção realizada pelo medicamento (SCATENA, 2015; TANNOCK; HICKMAN, 2016). Além disso, o poder estatístico do estudo deve ser avaliado (por exemplo, mensurando se há pacientes suficientes de um subtipo de câncer mais raro) e os biomarcadores devem passar por validação e testes funcionais (VARGAS; HARRIS, 2016).

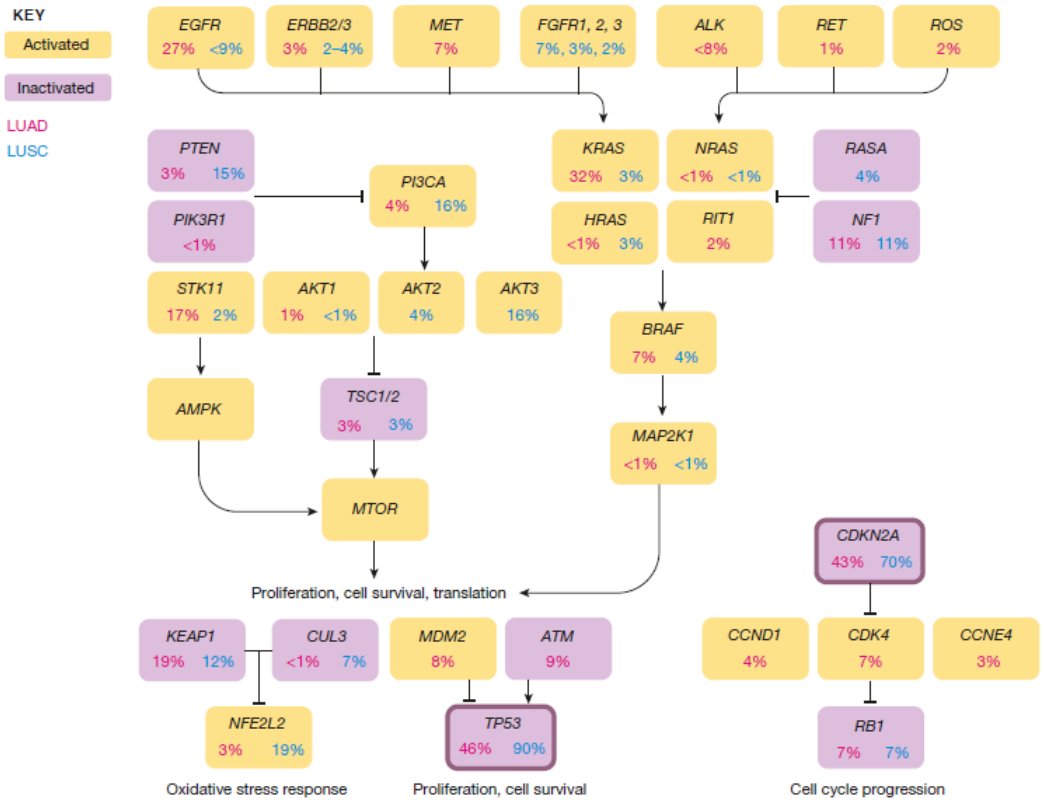

Figura 7 – Figura ilustrando os principais genes onde ocorrem alterações moleculares no adenocarcinoma (rosa) e carcinoma de célula escamosa (azul). Em laranja são os genes ativados e em lilás os genes inativados. As alterações que ocorrem nos genes compreendem mutações somáticas, deleções homozigotas, amplificações, e desregulação da expressão gênica (HERBST; MORGENSZTERN; BOSHOFF, 2018).

Existe também uma heterogenidade dentro até mesmo de uma amostra, que pode gerar medidas inconsistentes (VARGAS; HARRIS, 2016). Assim, existem estudos que buscam caracterizar as diferentes populações e sinais provenientes de um mesmo câncer de um indivíduo, além de descobrir o grau de heterogenidade (WU et al., 2021). O estudo de Wu et al. (2021), por exemplo, encontrou maior heterogenidade entre amostras de carcinoma de célula escamosa do que em adenocarcinoma (pulmão).

Apesar de existirem várias pesquisas buscando por biomarcadores para o câncer, com frequências os resultados não são promissores, gerando biomarcadores com baixa sensibilidade e especificidade. Conforme Hoseok e Cho (2015), este fato pode ser devido a utilização de apenas um marcador, sendo este um indicativo para investir em usar biomarcadores de diferentes moléculas juntos.

Isso porque, existe grande diferença de perfil molecular entre indivíduos com a mesma doença, por exemplo, a mutação em EGFR é mais frequente em asiáticos, enquanto mutações em K-Ras são menos frequentes (ambas em adenocarcinoma de pulmão) (ZHOU; CHRISTIANI, 2011). Além disso, também existe a heterogenidade do câncer em um

mesmo indivíduo. Nesse sentido, a utilização de uma combinação de marcadores teria maior facilidade em destacar o fenótipo. No estudo de Jang et al. (2021) por exemplo, uma combinação de microRNAs teve melhor performance na detecção prematura de câncer de mama do que apenas um.

1.3 Dados Ômicos e a Multiômica

1.3.1 Expressão Gênica

Por meio da habilidade de regulação da expressão gênica, as células podem se diferenciar para diferentes tecidos e órgãos com funções específicas, ou ainda, se adaptar às mudanças no ambiente e responder a determinados estímulos ambientais (Lesk, 2012). É possível identificar e quantificar os genes que estão sendo transcritos em larga escala por meio das técnicas de RNA-seq ou Microarray. A primeira permite a obtenção de informações mais amplas, pois diferente do Microarray que analisa uma quantidade limitada de transcritos presentes no chip, o RNA-seq quantifica todos os transcritos presentes em quantidade suficiente para ser identificado (Pevsner, 2015). Assim, o RNA-seq é capaz de detectar novos transcritos e isoformas de transcritos, bem como permite a análise de eventos de *splicing* alternativo (Pevsner, 2015). Usualmente, os dados de expressão gênica são utilizados com o propósito de comparar duas condições biológicas diferentes (Pevsner, 2015), como por exemplo, expressão em tecido tumoral e não tumoral.

1.3.2 SNPs

A maior parte das doenças é multifatorial e um resultado da interação entre genética e fatores ambientais que acontecem ao longo da vida (TALSETH-PALMER; SCOTT, 2011), por exemplo, existem variantes que são ferramentas para a identificação de resistência a insulina, entretanto, ter uma medida de circunferência da cintura alta predispõe o indivíduo ao risco da diabetes tipo 2 também (SCOTT et al., 2014). A variabilidade genética é a diferença gênomica que existe entre indivíduos de uma mesma população ou espécie e ela vem somando evidências de sua contribuição em doenças complexas como o câncer, doenças cardiovasculares e diabetes (TALSETH-PALMER; SCOTT, 2011). Existem varias formas de variação genética, como variantes de nucleotídeo único (SNVs), repetição em tandem de DNA, indels, alelos nulos, entre outros.

Indels são inserções ou deleções menores que 50 nucleotídeos, sendo que a maior parte deles são de até 3 nucleotídeos (TALSETH-PALMER; SCOTT, 2011). Já as repetições em tandem de DNA variam de repetições de dois ou três nucleotídeos até milhares (TALSETH-PALMER; SCOTT, 2011). As SNVs são as variantes de apenas um nucleotídeo quando comparadas com um genoma de referência e são as mais comuns: um genoma apresenta na faixa de 3,5 milhões desse tipo de variante (MARIAN, 2020).

As SNVs são consideradas SNPs (do inglês, polimorfismos de nucleotídeo único) quando atingem pelo menos 1% da população (TALSETH-PALMER; SCOTT, 2011). As SNVs podem ter diferentes consequências dependendo de onde elas estão localizadas no genoma, isso porque diferentes locais têm diferentes funções que o SNV pode estar

afetando. Conforme Marian (2020), os SNVs podem ocorrer em diferentes regiões do genoma e portando de genes:

- SNVs não-sinônimos: São aqueles que afetam a sequência de aminoácidos de uma proteína, adicionam um códon de parada ou removem um códon de parada. Esse tipo de mutação pode levar à perda de função da proteína. Tanto a adição e a perda do códon de parada levam à ativação das vias de degradação dos transcritos em questão. Já a mudança de um aminoácido na proteína pode resultar em mudanças na polaridade, carga, hidrofobicidade, hélices e localização de cada resíduo na proteína final, podendo também afetar sua funcionalidade;
- SNVs sinônimas: Como o código genético é degenerado, a mutação de um nucleotídeo não altera a sequência de nucleotídeos de uma proteína. Entretanto, ela pode alterar a eficiência de tradução e de transcrição, *splicing*, a estabilidade do RNA mensageiro e a estrutura terciária de RNAs (afetando sua interação com outros RNAs);
- SNVs de *splicing*: Esse tipo de variante pode alterar sítios de *splicing*, gerar novos sítios ou ativar um sítio críptico. Podendo levar ao salto de um exon, retenção de íntron e mudança do quadro de leitura.
- SNVs intrônicos: Não é esperado que esse tipo de variante cause muitos efeitos, a não ser que eles ativem sítios crípticos de *splicing*, afetem os *enhancers* ou os RNAs não codificantes que são transcritos a partir daquela região.
- SNVs indels: É quando ocorre a deleção ou a inserção de 1 nucleotídeo e podem afetar a ligação de elementos regulatórios ou de transcrição quando ocorrem nas regiões regulatórias 5'. Quando ocorrem na região regulatória 3' pode gerar instabilidade no mRNA. Entretanto, a maioria dos indels está localizada nas regiões intergênicas e descobrir seus efeitos é um desafio.

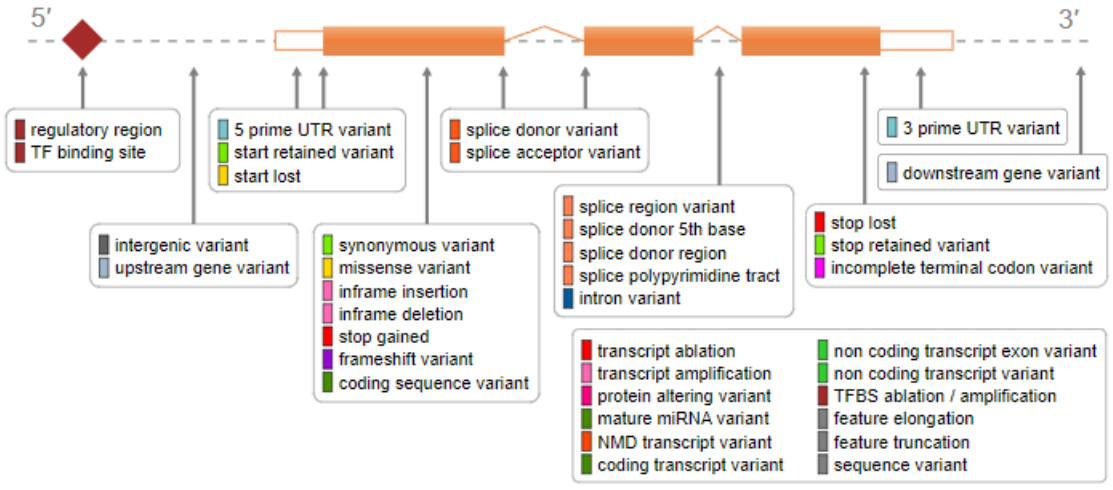

Figura 8 – Regiões onde ocorrem as SNVs (Ensembl)

1.3.3 Multiômica

O estudo de todo um conjunto de moléculas na biologia molecular, como o genoma, transcritoma, epigenoma, proteoma e metaboloma, por meio da quantificação e caracterização, é dado o nome de ômica. Estes diferentes níveis de regulação celular atuam conjuntamente através de uma rede de mecanismos celulares, composta por elementos interdependentes, cuja comunicação e regulação é responsável por definir o destino celular.

Mesmo que as ômicas separadamente já permitam a geração de muito conhecimento acerca de doenças e biomarcadores importantes, segundo (CAI et al., 2022), análises considerando uma única ômica não causaram a revolução no tratamento de doenças complexas. Por exemplo, no estudo realizado analisando pesquisas genômicas clínicas no tratamento de câncer, apenas 3 a 13% do pacientes receberam tratamentos personalizados conforme seus dados genômicos Tannock e Hickman (2016). O sucesso em entender a arquitetura genética e genômica das doenças complexas tem sido modesto e parcialmente esse problema pode ser explicado pela pouca exploração da interação entre as ômicas (RITCHIE et al., 2015). Com isso, é possível que as análises que combinem múltiplas informações ômicas (análises multiômicas) possam lançar luz aos padrões complexos que caracterizam as patologias (CAI et al., 2022).

Com a multiômica, podemos entender como a variação e a interação entre várias ômicas contribui para a fisiologia e doença em organismos complexos. A integração das diferentes ômicas pode diminuir a quantidade de falsos positivos, já que os resultados procurados serão aqueles nos quais múltiplos tipos de dados ômicos se complementam para o entendimento de um resultado ou apontam para os mesmos genes e mesmas redes de interação de moléculas (RITCHIE et al., 2015).

Como exemplo pode-se extrair destes estudos o potencial de chegar a novos biomarcadores que possibilitam novos alvos terpêuticos e personalização de tratamento de pacientes, descoberta de novas redes, descobrir as alterações que levam a um doença, construir um modelo e fazer predições de fenótipos (como acontece na área da saúde para diagnóstico) (MISRA et al., 2019).

Para tratar computacionalmente dados multiômicos são utilizados métodos baseados em aprendizado de máquina (do inglês *machine learning*, ML), sendo assim, análises de larga-escala multiômicas têm ajudado a desvendar a complexa regulação sistêmica associada com os diferentes fenótipos, o que seria impossível na análise de uma única ômica.

1.4 Machine Learning

O padrão nas ciências da saúde era utilizar métodos estatísticos para analisar os dados e interpretar. Com o advento da geração de dados abundantes provenientes do sequenciamento de nova geração, a redução de custo da capacidade computacional e o sucesso das abordagens de *machine learning* em diversas áreas, o uso da última tornou-se popular nas ciências da saúde também (REEL et al., 2021). A diferença entre as duas abordagens é que enquanto o ML está focando em fazer predições com acurácia e permite maior flexibilidade e escalabilidade, os métodos estatísticos tem por objetivo inferir relações entre as variáveis (RAJULA et al., 2020).

ML apresenta mais flexibilidade por não necessitar de tantas pressuposições antes da análise, tais como a distribuição de erros, aditividade dos parâmetros no preditor linear e riscos proporcionais (RAJULA et al., 2020). Em adição, ML também é mais indicada para dados multiômicos por apresentam poucas observações e muitos preditores (ocorre na genômica, transcriptômica, proteômica e metabolômica), além de ser mais apropriada devido a maior capacidade de considerar as relações complexas entre as variáveis que contribuem para o fenótipo ou efeito (RAJULA et al., 2020; FABRIS et al., 2017; IJ, 2018).

A inteligência artificial é derivada da ciência da computação e matemática, buscando imitar a inteligência natural a partir de uma máquina, com a intenção de aprender e imitar tarefas humanas (ARJMAND et al., 2022; CIPOLLA-FICARRA; QUIROGA; FICARRA, 2021). ML é um tipo de inteligência artificial que consiste em abordagens algoritmicas para resolver problemas sem ter uma programação específica para eles (NGIAM; KHOR, 2019; IJ, 2018), possibilitando compilar dados e adquirir experiência (identificar padrões e regularidades) a partir dos mesmos e realizar predições (MOHAMED, 2017; ALPAYDIN, 2014).

	Classe Predita		
Classe Verdadeira	Positivo	Negativo	Total
Positivo	Verdadeiro positivo	Falso negativo	p
Negativo	Falso positivo	Verdadeiro negativo	n
Total	p’	n’	N

Tabela 3 – Matriz de Confusão

Abaixo, são fórmulas que podem ser extraídas da matriz de confusão:

$$Acurácia = \frac{tn + tp}{tn + fn + fp + tp}$$

$$Taxa\ de\ erro = 1 - Acurácia = \frac{fn + fp}{tn + fn + fp + tp}$$

$$Sensibilidade = \frac{tp}{tp + fn}$$

$$Especificidade = \frac{tn}{fp + tn}$$

Apesar de *machine learning* ser uma abordagem mais flexível de análise de dados, é indicado o treinamento e teste de diferentes modelos para avaliar o desempenho dos mesmos para o dataset em questão (GREENER et al., 2022). Para tanto, o dataset é divido em um conjunto para induzir o modelo preditivo e outra para testá-lo. Há diferentes medidas para avaliar esse desempenho do classificador, dentre elas a matriz de confusão, que engloba os verdadeiros e falsos positivos e os verdadeiros e falsos negativos, cujos valores podem ser utilizados para calcular ainda a sensibilidade (taxa de positivos verdadeiros), especificidade (taxa de negativos verdadeiros), acurácia (taxa de acertos dentre as classificações) e taxa de erro (1 - acurácia) (tabela 3). Outro exemplo, é a Curva Característica de Operação do Receptor (curva ROC) representa o *trade-off* entre a sensibilidade e a especificidade, sendo que quanto mais próxima do canto superior estiver a curva, melhor (MOHAMED, 2017).

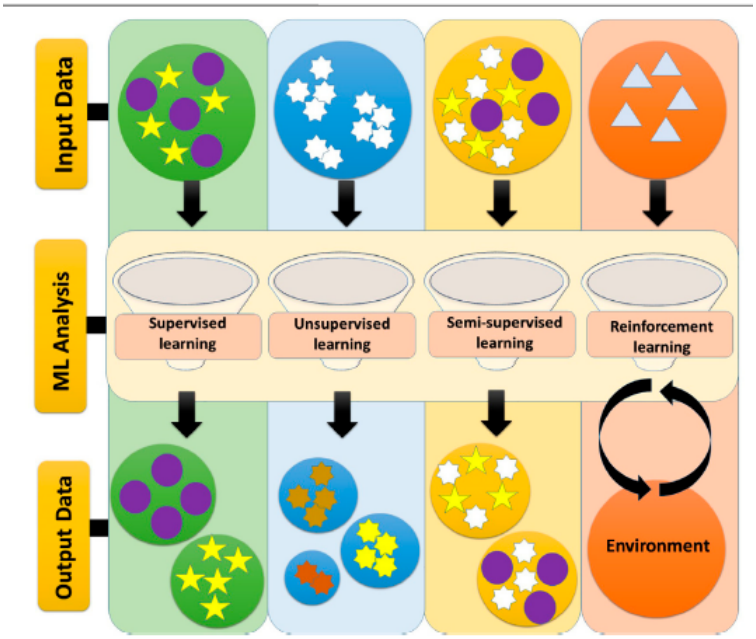

Figura 9 – Abordagens de *machine learning* (ARJMAND et al., 2022)

Conforme Arjmand et al. (2022), dentre os métodos de classificação, existe o aprendizado supervisionado, semi-supervisionado, não-supervisionado e por reforço (9):

- No aprendizado não-supervisionado, os dados não apresentam um rótulo correspondente, assim, o modelo é construído a partir do aprendizado do algoritmo, sem receber informações prévias no input. Espera-se que o modelo encontre padrões dentre os dados para dividi-los em grupos com atributos semelhantes;
- Para usar os métodos supervisionados é necessário que cada amostra esteja associada a um rótulo, que representa um grupo ou um valor. Sendo assim, o algoritmo de aprendizado supervisionado receberá dados com os quais irá treinar e seus respectivos alvos e com isso é esperado que ele aprenda o padrão para caracterizar novos inputs;
- O método semi-supervisionado é uma mistura das duas abordagens, sendo que a maior parte do dataset é, normalmente, não rotulado;
- Por fim, existe o aprendizado por reforço, onde o agente de aprendizagem toma ações e interage com um ambiente continuamente, recebendo recompensas quando atinge objetivos.

A figura de Greener et al. (2022) (Figura 10) ilustra como pode ser feita a escolha de um método de ML, bem como mostra os passos de treinamento do modelo. Três das abordagens de ML supervisionada são brevemente explicadas abaixo:

- Decision Tree é um dos métodos supervisionados mais utilizados e é bastante flexível pois não exige que os dados sejam paramétricos (JIANG; GRADUS; ROSELLINI, 2020). Um dos benefícios da decision tree é a possibilidade de interpretação visual do modelo criado e uma das limitações é que existe a possibilidade de árvores individuais causem overfit dos dados (JIANG; GRADUS; ROSELLINI, 2020).
- Random forests são um conjunto de decision trees, sendo que as predições serão baseadas nos resultados das várias árvores, como se fosse uma votação, assim diminuindo as chances de overfit dos dados e possibilitando a extração do ordenamento das features mais imporantes para a predição (CAI et al., 2022; JIANG; GRADUS; ROSELLINI, 2020).
- O objetivo do Support vector machine (SVM) é encontrar um limite para máxima separação, que é chamado de hiperplano, entre duas classes em um espaço multidimensional (dentre muitas variáveis) (JIANG; GRADUS; ROSELLINI, 2020). Um de seus pontos fortes é performar bem mesmo quando o número de preditores (genes) é maior do que o numero de observações (amostras) (JIANG; GRADUS; ROSELLINI, 2020). Por outro lado, não é possível conhecer a razão da acurácia do modelo, já que a forma como os preditores são combinados para gerar o hiperplano não é fornecida (JIANG; GRADUS; ROSELLINI, 2020).

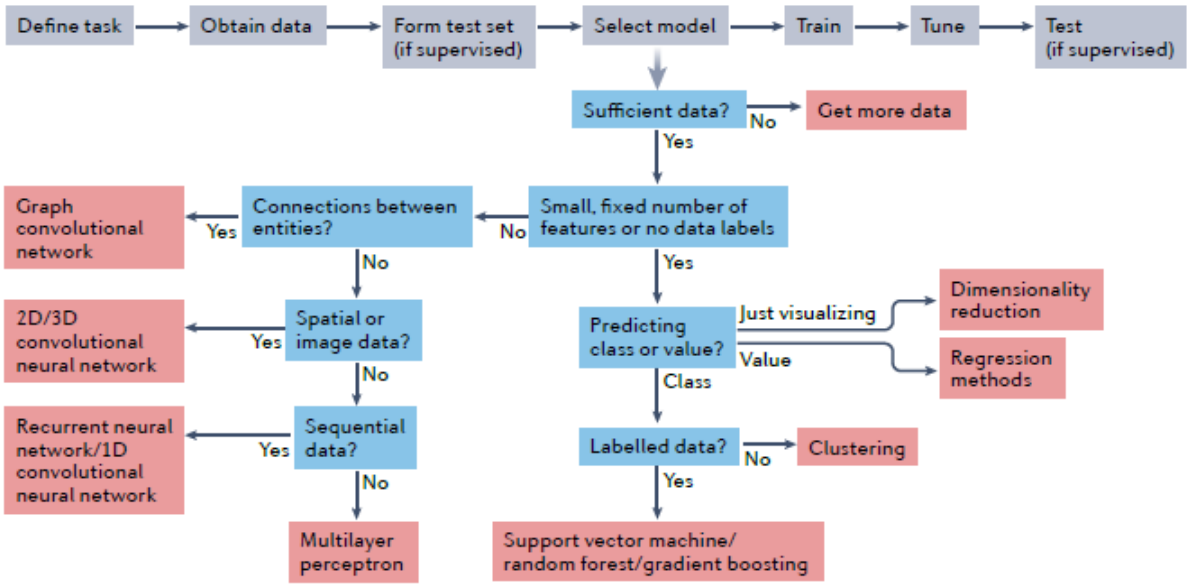

Figura 10 – No topo da figura há o passo a passo geral para treinamento do modelo escolhido. Em colorido há uma árvore de decisões mostrando os caminhos para a escolha de cada método de ML (GREENER et al., 2022).

Além de escolher um método de ML, alguns desafios podem aparecer ao longo do caminho. Dados genômicos podem apresentar algumas características particulares, como ter número de genes para análise muito maior que o número de amostras (p » n) e pode ocorrer desbalanceamento do número de amostras dentro de cada classe ou poucas amostras, podendo levar ao *overfitting* (XU et al., 2020) (quando o modelo construído não obtém sucesso em generalizar o que foi "aprendido"para outros *datasets* (YING, 2019)). A seleção de *features* é a solução para alguns dos desafios, se propondo a diminuir a dimensonalidade dos dados, para promover um modelo preditivo mais robusto, utilizando as *features* mais relevantes (XU et al., 2020). Assim, um passo a passo comum na área da multiômica é começar pela seleção de *features*, para após treinar o modelo e realizar predições quando novos dados estiverem disponíveis (Figura 11).

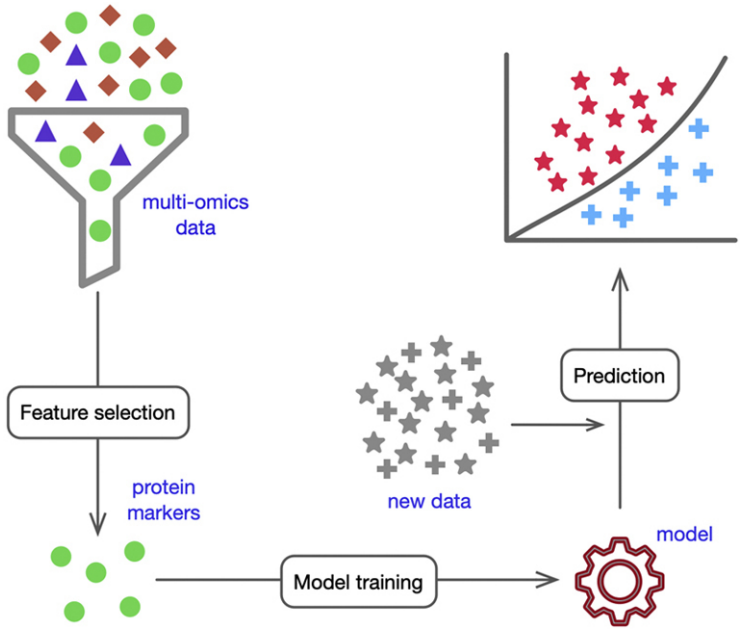

Figura 11 – Figura ilutrando o passo a passo utilizado com os dados multiômicos - feature selection, treinamento do modelo, predição (SHI et al., 2021)

2 Justificativa

Conforme (CAI et al., 2022), os tratamentos resultantes de análises de ômica única não causaram a revolução esperada, por exemplo, no estudo realizado analisando pesquisas genômicas clínicas no tratamento de câncer, apenas 3 a 13% do pacientes receberam tratamentos personalizados conforme seus dados genômicos Tannock e Hickman (2016). As drogas com alvos moleculares aumentaram a sobrevida e qualidade de vida. No entanto, elas ainda apresentam falhas por não levarem em conta a complexidade da câncer. Um exemplo é a heterogeneidade contida na patologia, permitindo que o câncer persista apesar do tratamento devido a existência de outras alterações moleculares que mantém as características da célula tumoral. Soma-se o problema dos falsos-positivos dos exames de imagem, que muitas vezes levam o paciente a passar por procedimentos mais invasivos em busca de diagnóstico. Tais procedimentos envolvem altos custos, além de riscos devido à radiação (exames de imagem) e desgaste ao paciente. Nesse sentido, os biomarcadores tumorais podem atuar conjuntamente com tais exames de imagem para o auxílio no diagnóstico e prognóstico da doença a fim de contribuir com a detecção da doença de forma mais prematura e assertiva. Isso pode ser de grande vantagem quando consideramos a complexidade do câncer, que envolve a desregulação da atividade de diversas moléculas biológicas à níveis genômicos, transcritômicos, proteicos, e metabolômicos. Assim, a análise multiômica com aplicação de técnicas de ML se mostra promissora para auxiliar na resolução deste impasse. A multiômica tem por princípio a análise de diferentes tipos de dados de forma conjunta, integrando as moléculas de diferentes camadas da biologia, portanto é esperado uma quantidade menor de falsos-positivos devido à múltiplas fontes de evidência apontando para genes e vias metabólicas de forma integrada, o que leva à maiores chances de compreensão dos processos que associam as características genotípicas ao fenótipo que se busca desvendar (RITCHIE et al., 2015). Assim, é possível que a compreensão acerca do câncer de pulmão seja enriquecida, abrindo-se possibilidades de estudo de novas rotas de interação entre genes que afetam significativamente esta patologia.

3 Objetivos

3.1 Objetivo geral

Este trabalho visa identificar biomarcadores com capacidade de predição de câncer de pulmão a partir de dados multiômicos através da integração de informações de variantes gênicas (polimorfismos de nucleotídeo único, SNPs) e de expressão gênica.

3.2 Objetivos específicos

- Revisar a literatura para a identificação de abordagens de ML consideradas padrão ouro para análise de dados de expressão gênica;
- Revisar a literatura para a identificação de abordagens de ML consideradas padrão ouro para análise de dados de variantes genéticas;
- Identificar genes com capacidade preditora da condição da amostra (amostra tumoral ou não) a partir dos dados de expressão gênica;
- Identificar genes com capacidade preditora da condição da amostra (amostra tumoral ou não) a partir dos SNPs;
- Montar uma rede de interação proteína-proteína com os genes com capacidade preditora encontrados a partir da análise da expressão gênica;
- Explorar as ontologias gênicas encontradas para os módulos da rede de interação proteína-proteína;
- Explorar as variantes resultantes da análise de ML;
- Integrar os resultados da análise de variantes genéticas e de expressão gênica;
- Revisar a literatura para verificar se os genes resultantes das análises já estão associados ao câncer e interpretar os resultados obtidos.

4 Procedimentos metodológicos

4.1 Obtenção dos dados

Para a realização das análises, foram utilizados dados de RNA-seq de amostras de seres humanos coreanos que apresentavam tumores malígnos e o nome do projeto é "Lung Cancer Sequencing Project Exome sequencing of lung adenocarcinomas and their normal counterparts", disponíveis no GEO (*Genome Expression Omnibus*) Dataset[1] sob o código de acesso GSE40419[2]. Este dataset original consiste de 87 amostras de adenocarcinomas de pulmão e 77 amostras de tecidos normais adjacentes. Entretanto, utilizamos para análise neste trabalho apenas 140 amostras que são pareadas entre tecido tumoral e não-tumoral do mesmo paciente. Os dados estão distribuídos conforme as tabelas 4 e 5.

4.2 Pré-processamento de reads

A análise de qualidade dos reads foi avaliada pelo programa FastQC (ANDREWS et al., 2012), seguida pela remoção dos adaptadores e bases de baixa qualidade através do programa Trimmomatic 0.32 (BOLGER; LOHSE; USADEL, 2014). A qualidade foi avaliada considerando-se o phred score > 30 (precisão de base-calling de 99,9%), duplicidade de bases GC, tamanhos de reads (mínimo 75nt) e distribuição das bases pelo tamanho dos reads.

4.3 Quantificação da expressão gênica

Os reads foram alinhados ao genoma de referência de Homo sapiens GRCh38 através do programa Spliced Transcripts Alignment to a Reference (STAR) 2.6.0a (DOBIN et al., 2012). Os dados alinhados serão avaliados pelo programa RNA-Seq by Expectation Maximization (RSEM) (LI; DEWEY, 2011), que realizará a estimativa de abundância dos reads. Posteriormente, a matriz de estimativa de abundância da expressão foi normalizada pelo deseq usando a função vst, que é adequada para tratamento de inputs de ML.

4.4 Análise do perfil mutacional e predição de variantes

Os dados provenientes da etapa de pré-processamento foram analizados para a identificação de variantes gênicas de acordo com o *best practices* pipeline do programa

[1] https://www.ncbi.nlm.nih.gov/gds
[2] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40419

The Genome Analysis Toolkit (GATK) (DEPRISTO et al., 2011) para dados de RNA-seq. Para isso, os dados foram alinhados através do método 2-pass no programa STAR (DOBIN et al., 2012) para maior sensiblidade na detecção de junções de splicing. Posteriormente, o programa Picard (PICARD. . . , 2019) foi utilizado para a manipulação dos dados, seguido de etapas de filtragem pelo uso da ferramenta Split'N'Trim e Base Recalibration (BQSR), disponíveis no GATK. Para a identificação de variantes foi utilizado o programa Mutect2 (BENJAMIN et al., 2019). As mutações foram filtradas pelo VariantFiltration e apenas os SNPs foram selecionados utilizando o comando SelectVariants. A anotação e predição de efeito foi realizada pelo programa VEP.

4.5 Machine learning para análise de dados de SNPs

A análise de machine learning com os dados dos SNPs em formato VCF foram feitas com a ferramenta VariantSpark (O'BRIEN et al., 2015), que utiliza Random forest e é otimizada para análise com variantes. O VariantSpark constrói um modelo para estimar a importância das variáveis em seu contexto biológico.

4.6 Machine learning para análise de dados de expressão gênica

Para as análises de machine learning com dados de expressão gênica, o pacote Scikit-learn para linguagem de programação python foi utilizado. Dentro desse pacote, um classificador foi treinado com os dados utilizando o método de machine learning random forest e as 100 features mais importantes para a classificação segundo esse método foram extraídas, sendo essa etapa chamada de feature selection.

As features selecionadas são então classificadas utilizando o método Support Vector Machine (SVM). Durante o treinamento do modelo com os dados, o método utilizado para validação cruzada foi o Leave-One-Out (LOOCV), o que significa que o modelo é treinado com as amostras e uma é deixada de fora para que depois seja usada para testar a capacidade de predição do modelo. Isso acontece até que todas as amostras tenham sido deixadas de fora uma vez. Assim, a matriz de confusão foi obtida, assim como outras métricas a partir dela.

O F1-score foi a métrica avaliada para decidir a quantidade de genes (entre 1 e 100 genes) a serem analisados nas próximas etapas. Para possibilitar a visualização dos genes selecionados e sua expressão gênica, bem como a capacidade de separação das condições para tais genes, foi utilizado no R o pacote pheatmap para geração do mapa de calor.

4.7 Biologia de Sistemas

As features que juntas apresentaram o melhor F1-score foram usadas na criação da rede interatômica. Neste sentido, 68 genes foram usados como input no banco de metabusca stringDB versao 12.0 ((SZKLARCZYK et al., 2023)) usando os parametros a seguir: pontuação mínima de interação = 0.4, fontes sobre as interações ativas = experimentos, bancos de dados e co-expressão, quantidade de interatores na primeira camada = 50"

Posteriormente, o programa Cytoscape 3.10.0 (SHANNON et al., 2003) foi utilizado para a criação do design e análises das redes geradas pelo stringDB, utilizando para isso os aplicativos MCODE 2.0.3 (BADER; HOGUE, 2003), BiNGO 3.0.3 (MAERE; HEYMANS; KUIPER, 2005), e Centiscape 2.2 (SCARDONI et al., 2014). Com o Centiscape fizemos as análises de centralidade, marcando as opções "degree"e "betweeness". Os nós que obtiveram uma pontuação acima da média da rede foram considerados *hubs* (para o atributo *degree*) e bottlenecks (para o atributo *betweenness*). O MCODE permitiu a descoberta de módulos (clusters) e o BiNGO foi usado para a identificação de ontologias gênicas (GO) super-representadas nos módulos. Para MCODE e Centiscape foram usados os parâmetros padrão. No BiNGO os seguintes parâmetros foram usados: teste estatístico = teste hipergeométrico, correções de testes múltiplos = Benjamini-Hochberg, nível de significância = 0,05, categorias = superrepresentadas após correção, arquivo de ontologias = *GO_Biological_Process*. Os processos biológicos mais relevantes foram selecionados em detrimento dos mais genéricos e estão disponíveis em tabelas ao longo dos resultados. O Biomart foi usado para buscar informações adicionais dos genes nas redes.

Gender	SmokingStatus	Stage	Count
female	current smoker	3A	1
	never smoker	1A	13
		1B	9
		2A	2
		3B	2
		4	2
		NA	1
	smoker	3A	1
		1B	1
		1A	1
male	NA	3A	1
	NA	1B	1
	NA	NA	1
	current smoker	3A	2
		2B	2
		2A	1
		1B	1
		1A	1
	never smoker	1B	2
		1A	1
	smoker	1A	10
		1B	6
		2B	4
		4	2
		3A	2

Tabela 4 – Distribuição de dados dos 70 pacientes conforme gênero, status de fumante e estágio do câncer

	AgeDiagnosis
count	70.000000
mean	63.685714
std	9.547050
min	38.000000
25%	58.250000
50%	65.000000
75%	69.000000
max	82.000000

Tabela 5 – Distribuição das idades dos 70 pacientes

5 Resultados e Discussão

5.1 Análise de ML para dados transcritômicos

Após a seleção das 100 *features* mais importantes pelo treinamento de um modelo *Random Forest*, a métrica F1 *score* dada pela validação cruzada no modelo de SVM foi avaliada para cada conjunto de *features* de 1 a 100 (cada novo conjunto era adicionado 1 *feature*, até chegar ao total, 100).

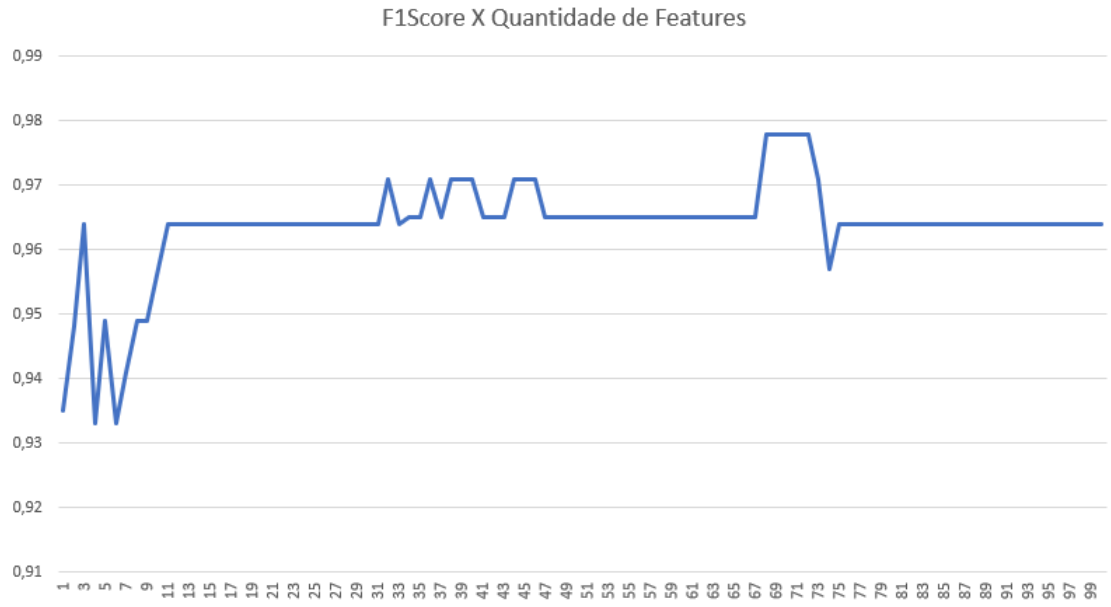

Figura 12 – F1 score em função da quantidade de features

Na Figura 12 é mostrado o gráfico com a métrica F1 para a respectiva quantidade de features avaliada. O trecho que contém os pontos de 68 a 72 features apresentou um F1 score de 0,978. Portanto, decidimos trabalhar com os primeiros 68 genes da análise, pois com eles a predição por meio de Machine Learning atingiu o seu valor máximo. Mapas de calor também foram produzidos para avaliar a como a expressão dos genes se configuravam nas amostras tumorais e normais (Figura 13).

No mapa de calor das 68 features (Figura 13) é observada a clusterização hierárquica das amostras e da expressão dos genes. A separação entre as duas condições (câncer em turquesa e normal em vermelho) é bem definida, entretanto, algumas amostras permeiam as amostras da outra condição. As amostras que estão em meio a maioria de amostras de outra condição têm a expressão dos genes apresentados no gráfico mais parecidas com a outra condição.

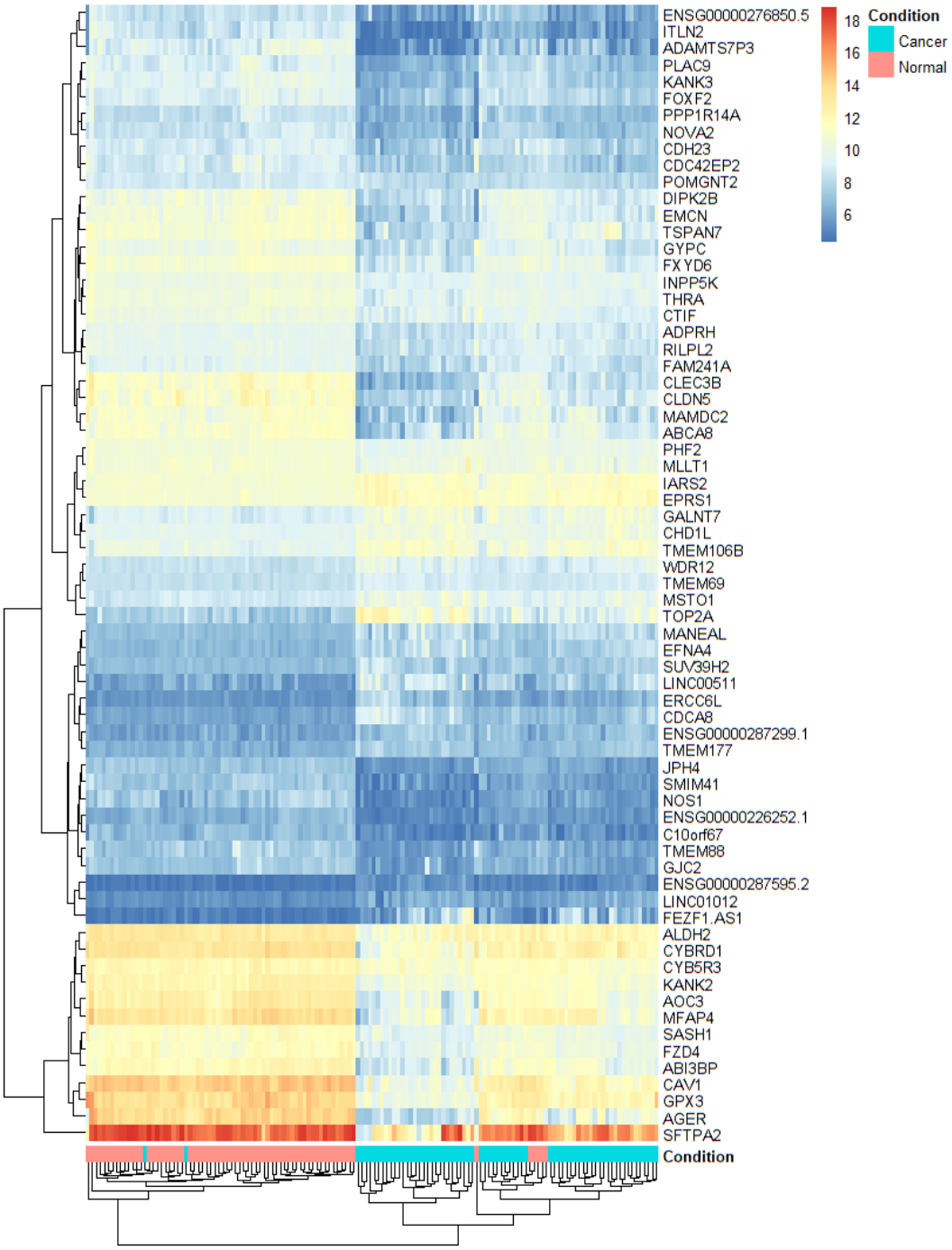

Figura 13 – Mapa de calor com 68 features (genes) que apresentaram o maior F1 score juntos. A opção de clusterização foi utilizada tanto nas amostras (vertical) quanto nos genes (horizontal), assim, amostras classificadas como tumorais estão antecipadas, pela cor turquesa e amostras classificadas como tecido normal pela cor vermelha. Genes com expressão aumentada estão em tons avermelhados, enquanto os que tem a expressão diminuída estão em tons azulados.

Para a criação da rede de interação proteín-proteína no StringDB, todos os 68 genes foram usados como *input*. Alguns genes não apresentaram interações com o restante, assim 24 genes foram removidos. A rede final (Figura 14) ficou com 44 genes do *input* original (Tabela 6) mais 293 outros genes. Unindo os genes existem 1626 pontes.

Com a criação dessa rede foi possível obter as interações de quais participam ou quais são os circuitos de sinalização dos genes selecionados. Essas vias de interação podem estar relacionadas com as causas do câncer de pulmão ou mediando uma resposta à doença. Além disso, a partir da análise da rede, será possível compreender o papel dos genes selecionados por ML e potencialmente prever as principais vias metabólicas que são afetadas na patologia.

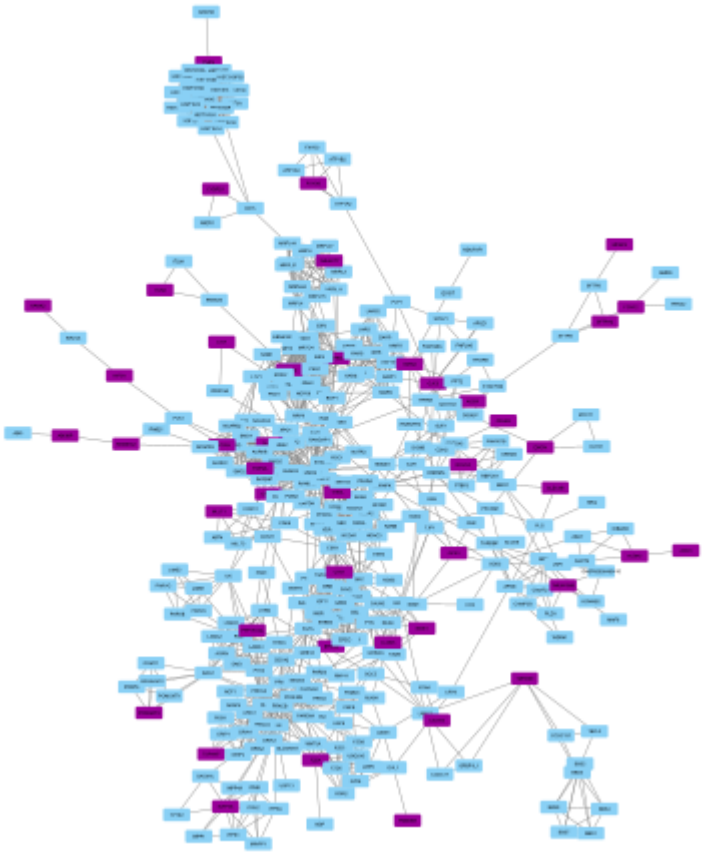

Figura 14 – Rede final completa com 44 inputs de ML em roxo e mais 293 de interações encontradas pelo stringDB, com 1626 pontes entre os dados. Na rede interatômica, a cor roxa foi utilizada para diferenciar àqueles genes provenientes da análise de ML e os azuis foram inseridos na etapa de utilização do STRINGDB.

A partir da rede formada, foram realizadas diferentes análises topológicas a fim de identificar medidas de centralidade, como o grau de nó e o *betweenness*. O grau de nó se define pela quantidade de nós conectados a ele. Se ele tiver um grau maior que a média dos graus de todos os nós, ele é chamado de hub. Já o *betweenness* é relacionado à quantidade de caminhos curtos que passam pelo nó, nesse sentido, um valor acima

da média de *betweenness* indica um nó *bottleneck*, ou seja, é um nó que serve de ponte para diferentes processos. Quando usadas essas características a fim de selecionar nós, chamamos o resultado de hubs-bottlenecks (HBs). Eles são nós com número conexões acima da média e que mais fazem parte de caminhos mais curtos entre diferentes nós. Assim, são os nós mais importantes da rede, pela capacidade de comunicação de vários processos biológicos, além de apresentarem muitas conexões com outras proteínas.

Gene name	NCBI gene description	Chromosome	Gene start (bp)
ABI3BP	ABI family member 3 binding protein	3	100749156
AGER	advanced glycosylation end-product specific receptor	6	3485949
ALDH2	aldehyde dehydrogenase 2 family member	12	111766887
AOC3	amine oxidase copper containing 3	17	42851184
CAV1	caveolin 1	7	116524994
CDCA8	cell division cycle associated 8	1	37692481
CDH23	cadherin related 23	10	71396920
CHD1L	chromodomain helicase DNA binding protein 1 like	1	147242654
CLDN5	claudin 5	22	19523024
CLEC3B	C-type lectin domain family 3 member B	3	45001548
CTIF	cap binding complex dependent translation initiation factor	18	48539031
CYBRD1	cytochrome b reductase 1	2	171522247
EFNA4	ephrin A4	1	155063737
EMCN	endomucin	4	100395341
EPRS1	glutamyl-prolyl-tRNA synthetase 1	1	219968600
ERCC6L	ERCC excision repair 6 like, spindle assembly checkpoint helicase	X	72204657
FAM241A	family with sequence similarity 241 member A	4	112145454
FOXF2	forkhead box F2	6	1389576
FXYD6	FXYD domain containing ion transport regulator 6	11	117836976
FZD4	frizzled class receptor 4	11	86945679
GJC2	gap junction protein gamma 2	1	228149930
GPX3	glutathione peroxidase 3	5	151020591
IARS2	isoleucyl-tRNA synthetase 2, mitochondrial	1	220094132
INPP5K	inositol polyphosphate-5-phosphatase K	17	1494577
ITLN2	intelectin 2	1	160945025
KANK2	KN motif and ankyrin repeat domains 2	19	11164270
KIF2A	kinesin family member 2A	5	62306162
MAMDC2	MAM domain containing 2	9	70043848
MFAP4	microfibril associated protein 4	17	19383442
MLLT1	MLLT1 super elongation complex subunit	19	6210381
NOS1	nitric oxide synthase 1	12	117208142
NOVA2	NOVA alternative splicing regulator 2	19	45933734
PHF2	PHD finger protein 2	9	93576584
POMGNT2	protein O-linked mannose N-acetylglucosaminyltransferase 2 (beta 1,4-)	3	43079229
PPP1R14A	protein phosphatase 1 regulatory inhibitor subunit 14A	19	38251237
SFTPA2	surfactant protein A2	10	79555852
THRA	thyroid hormone receptor alpha	17	40058290
TMEM106B	transmembrane protein 106B	7	12211270
TMEM177	transmembrane protein 177	2	119679167
TMEM69	transmembrane protein 69	1	45688181
TMEM88	transmembrane protein 88	17	7855066
TOP2A	DNA topoisomerase II alpha	17	40388525
TSPAN7	tetraspanin 7	X	38561542
WDR12	WD repeat domain 12	2	202874261

Tabela 6 – Informações adicionais sobre os genes em roxo da rede de interação proteína-proteína, sendo estes resultado da análise por meio de ML. As colunas, da esquerda para a direita exibem o nome do gene, a descrição dele provinda do NCBI, o cromossomo em que o gene está localizado, e a posição em que ele começa.

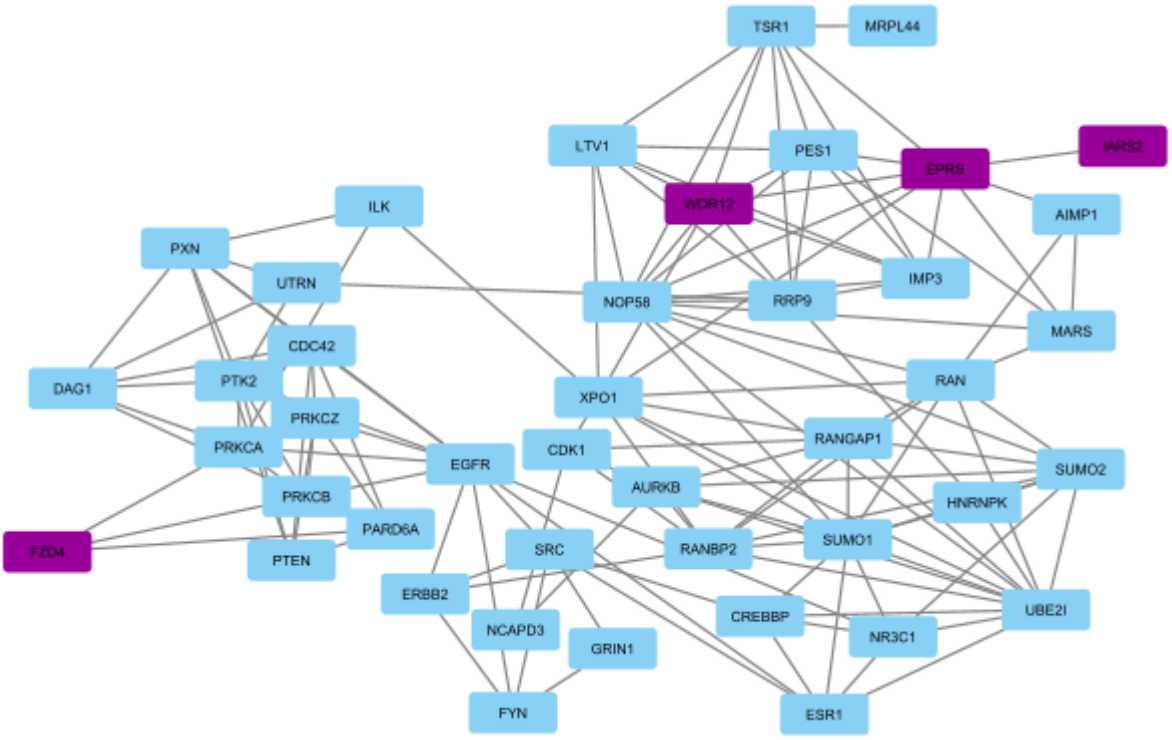

Figura 15 – Genes considerados Hubs-bottlenecks

Assim como os HBs, os principais módulos (clusteres) da rede também foram analisados. Eles são um conjunto de genes altamente conectados dentro da rede, permitindo visualizar as vias bioquímicas importantes para a patalogia. Os processos biológicos mais relevantes para o grupo de genes de cada cluster (tabelas 7 à 11) foram analisados para que fosse possível ter uma ideia de qual é a função daquela via e se eles pode explicar qual sua importância dentro do contexto do adenocarcinoma de pulmão, assim indicando a relevância de cada gene como biomarcador.

Na rede de ontologias gênicas do cluster 1 há diversos processos biológicos que estão relacionados a organização do DNA por meio da cromatina. Isso ocorre pelo fato de este cluster ser composto por genes de histonas e pelo gene PHF2. Este último, é uma demetilase de lisinas que demetila proteínas que são histonas e outras que não são histonas (HORTON et al., 2011; WEN et al., 2010; BABA et al., 2011). Sabe-se que uma das funções dessa proteína é demetilar histonas e por conseguinte, ativar a transcrição de genes (BABA et al., 2011).

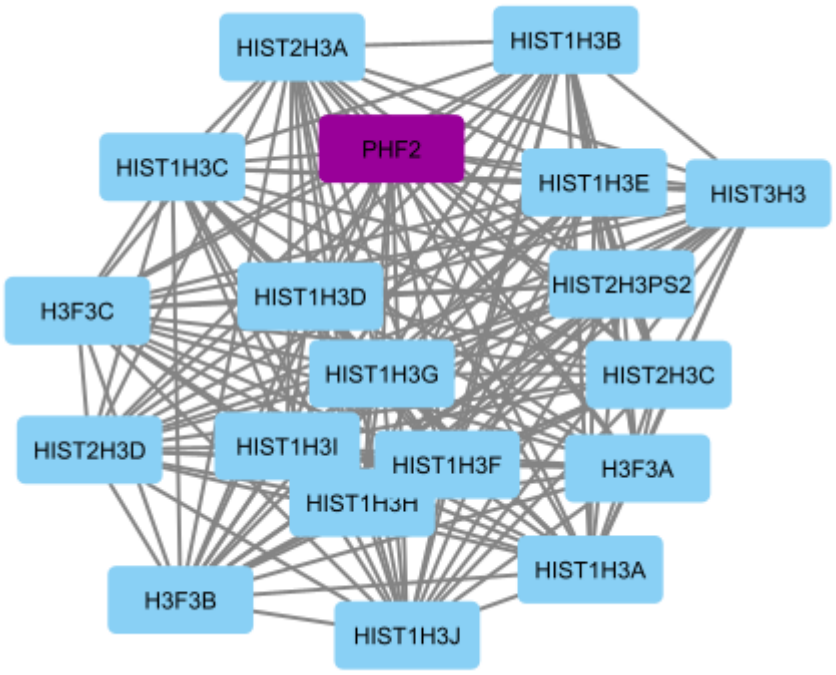

Figura 16 – Cluster 1

GO-ID	corr p-value	x	X	Description
6334	8.46E-36	17	18	nucleosome assembly
31497	9.27E-36	17	18	chromatin assembly
65004	1.41E-35	17	18	protein-DNA complex assembly
34728	1.41E-35	17	18	nucleosome organization
6323	6.02E-34	17	18	DNA packaging
6333	1.72E-33	17	18	chromatin assembly or disassembly
71103	6.69E-33	17	18	DNA conformation change
6325	8.77E-28	18	18	chromatin organization
51276	4.73E-26	18	18	chromosome organization

Tabela 7 – Ontologias gênicas do Cluster 1 - tabela apresenta GO-ID, valor de p, valor de p corrigido, x (o número de nós do cluster anotado para esta ontologia), X (número total de genes no cluster) e descrição do processo biológico

Foi previamente relatado que a depleção ou baixa expressão do gene PHF2 afetava o mecanismo de correção de DNA por recombinação homóloga (VEGA et al., 2020). A baixa expressão de PHF2 nas condições de indução de erros por radiação ionizante causava a diminuição da atuação de BRCA1 e a depleção de PHF2 também causava diminuição da acumulação de CtIP nas lesões de DNA e consequentemente afetando Rad51 resultando em eficiência reduzida na recombinação homóloga (forma de reparação de DNA). Isso ocorre porque PHF2 controla a transcrição desses genes removendo os marcadores de repressão na região promotora (LEE et al., 2015).

Também é conhecido que a deleção do gene causa instabilidade do genoma, afeta o

crescimento e o ciclo celular, resposta inflamatória e a diferenciação celular (YANG et al., 2018; STENDER et al., 2012; PAPPA et al., 2019). PHF2 já foi encontrado deletado ou hipermetilado em cânceres de mama, e mutante em cânceres gátricos e de cólon (SINHA et al., 2008; LEE et al., 2017). Olhando para a expressão desse gene (ENSG00000197724) no mapa de calor dos 68 genes, ele apresenta menor intensidade de expressão nas amostras de câncer (Figura 13).

PHF2 aumenta a resposta supressora de tumor de outro gene, o p53, por demetilar H3K9-Me2 nos locais alvo do p53 (LEE et al., 2015). Um processo (EMT - Epithelial to Mesenchymal Transition) que é bastante influenciado pelo controle da configuração da cromatina e modificação de histonas é quando as células epiteliais perdem a adesão entre si e adquirem propiedades de migração e invasão que são importantes para ocorrer a metástase. PHF2 tem o papel de ser o mediador entre o papel da AMPK de demetilase, diminuindo a quantidade de modificações do tipo H3K9me2 que reprimem a expressão de genes cruciais para o epitélio (DONG et al., 2023). Assim, PHF2 é um gene importante para evitar a metástase (DONG et al., 2023).

Nesse sentido, é observado que no cluster 1 há processos biológicos importantes para a regulação da transcrição devido a organização da cromatina. O gene do cluster que é um dos inputs de ML (PHF2) já conhecido por estar envolvido no EMT, possibilitando a transcrição de genes importantes para manter as características do epitélio, além de participar do processo de reaparação do DNA. Além disso, outros genes já conhecidos também por seu papel no câncer, como BRCA1, Rad51, p53 e AMPK atuam em vias bioquímicas que PHF2 também atua.

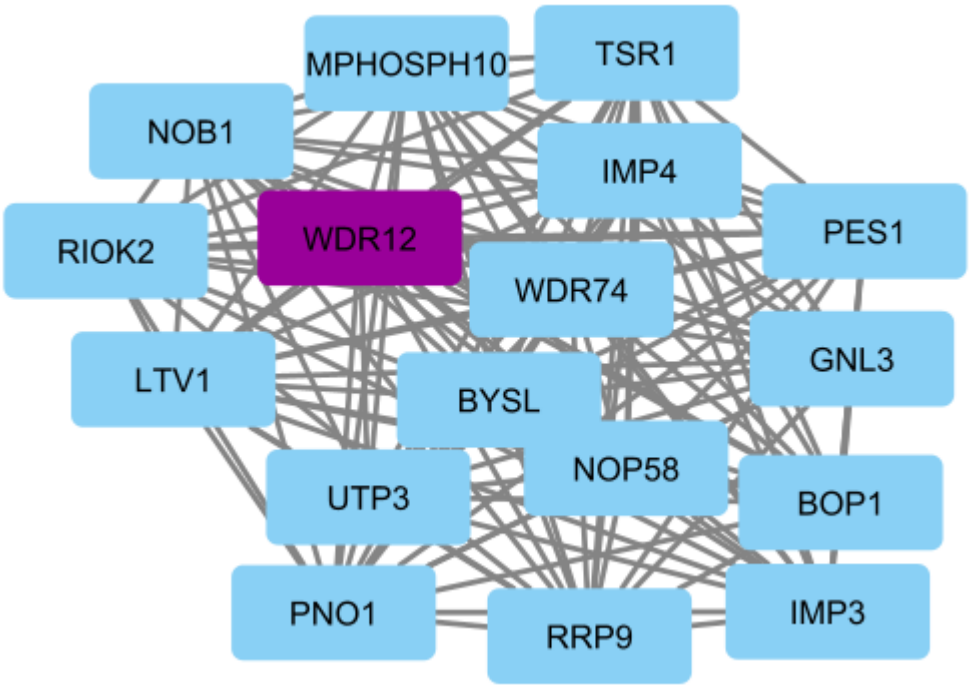

Figura 17 – Cluster 2

GO-ID	corr p-value	x	X	Description
42254	$3{,}89\times10^{-16}$	10	14	ribosome biogenesis
22613	$1{,}15\times10^{-14}$	10	14	ribonucleoprotein complex biogenesis
6364	$4{,}42\times10^{-13}$	8	14	rRNA processing
463	$1{,}28\times10^{-08}$	3	14	maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
470	$1{,}28\times10^{-08}$	3	14	maturation of LSU-rRNA
460	$5{,}62\times10^{-05}$	2	14	maturation of 5.8S rRNA
466	$5{,}62\times10^{-05}$	2	14	maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
10467	$6{,}47\times10^{-05}$	8	14	gene expression
42273	$4{,}17\times10^{-04}$	2	14	ribosomal large subunit biogenesis
6608	$6{,}70\times10^{-03}$	1	14	snRNP protein import into nucleus
8283	$4{,}69\times10^{-02}$	3	14	cell proliferation

Tabela 8 – Ontologias gênicas do Cluster 2 - tabela apresenta GO-ID, valor de p, valor de p corrigido, x (o número de nós do cluster anotado para esta ontologia), X (número total de genes no cluster) e descrição dos processos biológicos

Na Figura 17 está ilustrado o cluster 2. Nesse caso, as ontologias estão associadas com a biogênese do ribossomo.

O papel da biogênese aumentada de ribossomos e consequente produção de proteínas na sustentação do crescimento dos tumores é bem descrita na literatura (PELLETIER; THOMAS; VOLAREVIĆ, 2018; PECORARO et al., 2021). Uma das características do câncer é apresentar número e forma alteradas de nucléolos, e a célula tumoral permite uma aumentada biogênese de ribossomos (PECORARO et al., 2021). Estudos sugerem seu papel também na tumorigênese, devido ao fato de que altas taxas de produção de proteínas estão associadas à baixa fidelidade da tradução ou mudança no padrão do mRNA traduzido (PELLETIER; THOMAS; VOLAREVIĆ, 2018).

WDR12 forma uma complexo estável com as proteínas PES1 e BOP1 (presentes no cluster), chamado de PeBoW que é essencial para a biogênese de ribossomos e proliferação celular (HOLZEL et al., 2005; EID et al., 2023; YIN et al., 2018). No câncer de glioma foi descoberto que a deleção desse gene inibia o crescimento do tumor e aumentava o tempo de sobrevivência (MI et al., 2021), contudo, em outro estudo a deleção dele promoveu a apoptose (LI et al., 2020).

Altos níveis de expressão de WDR12 foi encontrada em vários tipos de cânceres (adenocarcinoma de estômago, timo, glioma, linfoma de células B e adenocarcinoma de pulmão) e no adenocarcinoma de pulmão a expressão aumentada de WDR12 também se mostrou relacionada negativamente com a sobrevivência (EID et al., 2023). No mapa de calor, podemos ver que na condição de câncer o gene WDR12 está ligeiramente mais

GO-ID	X	N	corr p-value	Description
6412	21	20	1.84×10^{-31}	translation
10467	21	21	4.01×10^{-21}	gene expression
6418	21	9	7.04×10^{-17}	tRNA aminoacylation for protein translation
43039	21	9	7.04×10^{-17}	tRNA aminoacylation
43517	21	1	1.83×10^{-02}	positive regulation of DNA damage response, signal transduction by p53 class mediator
43516	21	1	4.88×10^{-02}	regulation of DNA damage response, signal transduction by p53 class mediator

Tabela 9 – Ontologias gênicas do Cluster 3 - tabela apresenta GO-ID, valor de p, valor de p corrigido, x (o número de nós do cluster anotado para esta ontologia), X (número total de genes no cluster) e descrição dos processos biológico

expresso que no condição normal (Figura 13). Além disso, o WDR12 é um dos hubs-bottlenecks da rede, juntamente com PES1, SUMO1 e NOP58 sendo que os dois últimos estão positivamente relacionados com a expressão de WDR12 e participam da regulação do ciclo celular e replicação do DNA (EID et al., 2023).

Assim, o hub-bottleneck WDR12 é muito importante para o câncer sendo que ele e outros genes do cluster estão envolvidos na biogênese de ribossomos e, consequentemente, na proliferação celular, essencial para a patologia.

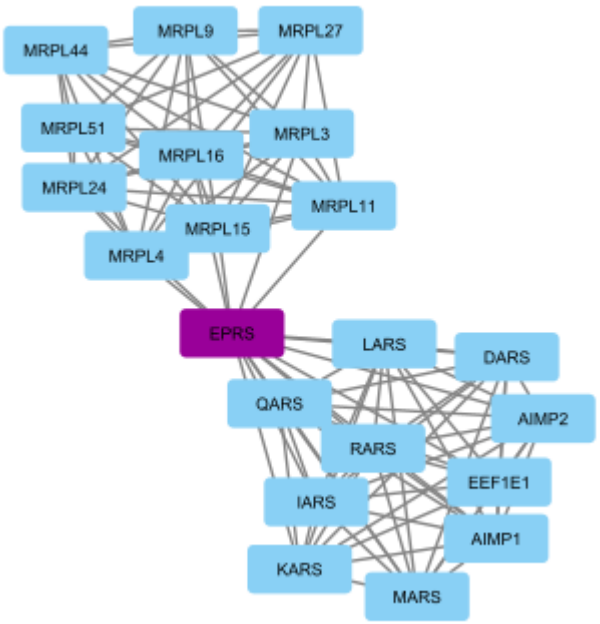

Figura 18 – Cluster 3

Na ontologia gênica do cluster 3 (Tabela 9) é possível observar que novamente

os processos estão envolvidos na síntese proteica, nesse caso, da aminoacilação do RNA transportador (tRNA). A glutamil-prolil-tRNA sintetase (EPRS) e outros genes presentes nesse cluster que são terminados em ARS pertencem a família das aminoacil tRNA sintetases (ARS), que tem função na tradução das sequências nucleotídicas, fazendo com que o tRNA se ligue ao seu aminoácido específico (aminoacilação do tRNA). Uma pequena parte da ontologia aponta para a regulação positiva da resposta ao dano do DNA, com a transdução de sinal pelo mediador da classe p53: o EEF1E1 (AIMP3).

As proteínas na parte inferior do cluster, juntamente com EPRS, formam um complexo de proteínas formado por 8 enzimas (ARS) e 3 proteínas multifuncionais que intergem com ARS (AIMPs 1 a 3 - EEF1E1 é AIMP3) (BIAN et al., 2021; ZHOU et al., 2020). As AIMPs já tem funções conhecidas como supressoras de tumores quando sozinhas, e algumas ARS controlam rotas de interação que são associadas ao câncer (KIM et al., 2019). Na figura de Kim et al. (2019) são ilustradas quais são essas rotas e suas finalidades, como morte celular, proliferação, angiogênese e migração.

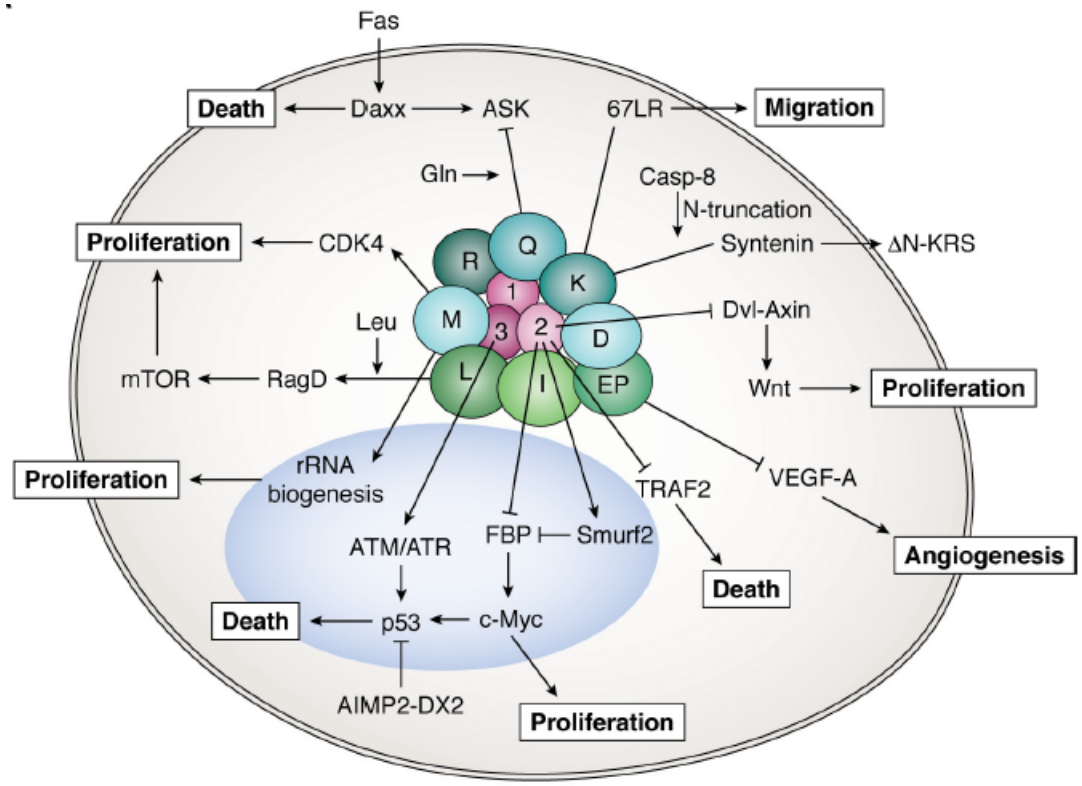

Figura 19 – Rotas de interação relacionadas ao câncer mediadas pelas ARSs e AIMPs formadoras do complexo MSC. No centro, os círculos em verde com letras são as ARSs e os círculos em rosa com números são as AIMPs 1, 2 e 3. (KIM et al., 2019).

Os genes na parte superior do cluster 3 são nucleares e são responsáveis por codificar alguns dos ribossomos mitocondriais (mitorribossomo). Quando há uma disfunção nos mitorribossomos e o processo de tradução na mitocondria é diminuído, pode ocorrer morte

celular no pulmão e doenças (KARIM; KOSMIDER; BAHMED, 2022). Ademais, descobriu-se que um dos mitorribossomos (MRPL41) estabiliza p53 e na ausência dessa estabiliza p27 (Kip1), contribuindo para controle do crescimento celular e apoptose (YOO et al., 2005). Não só essa, mas diversas proteínas mitorribossomais tem papéis reguladores da apoptose, o que leva os cientistas a hipotetizar sua função na sinalização apoptótica (KOC et al., 2001; GREBER; BAN, 2016; SHARMA et al., 2003; KIM; MAITI; BARRIENTOS, 2017).

EPRS é uma dos genes de input para gerar a rede e além disso também é um dos hubs-bottlenecks. Ele, assim como outras algumas ARS mostraram potenciais interações com a parte anterior da rota MTOR, como MAPK e PI3K (KIM et al., 2019). Um estudo comprovou a importância de EPRS também na regulação do ciclo celular e resposta ao estrogênio, importantes para o progresso tumoral em câncer de mama, assim apresentando maior número de cópias e expressão (KATSYV et al., 2016).

Outra ARS foi observada dentre os hubs bottlenecks, a IARS2. Observou-se que a deleção de IARS2 promoveu a apoptose dependente de mitocondria, portanto é sugerido que ela promova a proliferação celular e impeça a apoptose pela via se sinalização AKT/MTOR (DI et al., 2019). A via AKT/MTOR é um regulador central da proliferação celular, apoptose, ciclo celular, metabolismo e angiogênese. Sua ativação está associada à tumorigênese, resistência tumoral, invasão e metástase (DI et al., 2019).

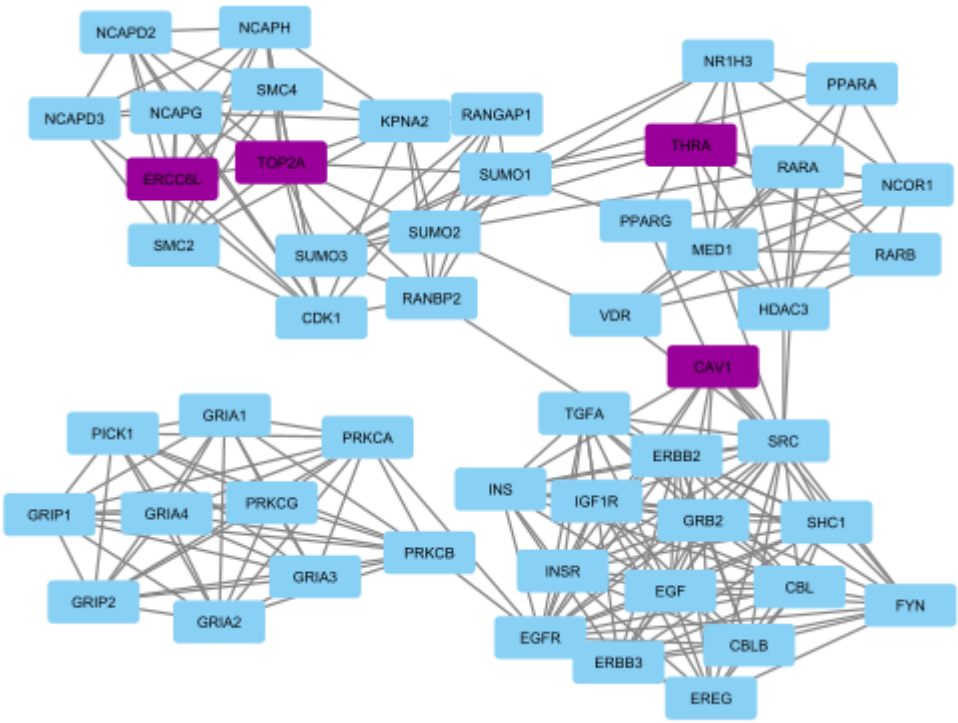

Figura 20 – Cluster 4

No cluster 4, os nós em roxo (que são inputs diretos da análise de expressão gênica

por ML, foram todos descritos na literatura como associados a algum tipo de câncer ou a algum processo associado ao câncer.

O gene TOP2A é uma topoisomerase, mais especificamente a topoisomerase II α, que é crucial para a replicação do DNA, a transcrição, a replicação dos cromossomos e a estabilidade do genoma (WANG et al., 2022; TIAN et al., 2021). Em um estudo realizado a partir do banco de dados TGCA, com dados de RNA-seq de amostras de tecido normal e câncer, descobriu-se que o TOP2A é super-expresso em diversos tipos de câncer e que a super-expressão está associada com um prognóstico ruim e resistência a drogas utilizadas no tratamento do câncer(WANG et al., 2022). No câncer pancreático descobriu-se que TOP2A interage com β-catenin, ativando oncogenes e contribuindo para a transição epitélio-mesenquimal que afeta positivamente a capacidade de metástase (PEI; YIN; LIU, 2018).

Complementação cruzada de reparo por excisão tipo grupo 6 (ERCC6L), também conhecida como PICH é um gene que é super-expresso no adenocarcinoma de pulmão comparado com os tecidos normais (HOU et al., 2022; HUANG et al., 2022). Os níveis de ERCC6L estão positivamente associados com a infiltração de células imunes, incluindo Th1, que estão ligadas ao prognóstico favorável de pacientes com câncer. Em contraste, a expressão de ERCC6L está negativamente correlacionada com a infiltração de células imunes que matam tumores, como as células Th2 e células NK (HOU et al., 2022). ERCC6L foi encontrado super-expresso em tecidos com câncer (mRNA e proteína) em câncer gástrico e nessas condições também ativa a transição epitélio-mesenquimal pela via NF-κB (CHEN; LIU; CAO, 2021). Ele também foi descrito como associado ao crescimento celular nos cânceres coloretal (XIE et al., 2019), de mama (LIU et al., 2018a) e hepatocelular (HOU et al., 2022).

O hormônio da tireóide é um regulador do crescimento, diferenciação e homeostáse de tecidos e algumas de suas atividades são mediadas por seus receptores nucleares (TR) THRA e THRB (GONZÁLEZ-SANCHO et al., 2003). Dados indicam a relação entre TR e p53, Rb, ciclina D e outros reguladores do ciclo celular e oncogenes (GONZÁLEZ-SANCHO et al., 2003). De acordo com um estudo, THRA e THRB podem afetar o desenvolvimento do glioma através da regulação, pelo menos parcialmente, das vias de sinalização da proteína quinase ativada por mitógeno (MAPK)/ERK e fosfoinositídeo 3-quinase (PI3K)/Akt (ZHANG et al., 2021). O estudo também descobriu que o T3 afetou a apoptose e o ciclo celular das células de glioma através da regulação das expressões THRA e THRB (ZHANG et al., 2021). Além disso, o T3 levou à parada das fases G1 e G2 das células, o que proporcionou tempo extra para a célula reparar o dano, reduzindo assim a ocorrência de mutações e evitando a formação de tumores (ZHANG et al., 2021).

CAV1 é uma proteína estrutural encontrada em vesículas específicas na superfície celular e ela está presente em uma ampla gama de tipos celulares porém em diferentes

níveis de expressão (FU et al., 2017; WANG et al., 2017; LIU et al., 2018b). Além disso, sua expressão está associada com tumores de vários tipos (SHERIF; SULTAN, 2013; LIU et al., 2018b). Suas funções fisiológicas são diversas, pois são descritos nos processos de proliferação, diferenciação, transdução de sinal e apoptose. Em câncer de pulmão foi descrito que a alta expressão de CAV1 (LIU et al., 2018b) promove a proliferação celular e metástase aumentando a expressão de um RNA longo não codificante, HOTAIR (LIU et al., 2018b).

GO-ID	corr p-value	x	X	Description
7076	3.04E-10	6	50	mitotic chromosome condensation
32583	1.07E-09	12	50	regulation of gene-specific transcription
7169	1.41E-09	12	50	transmembrane receptor protein tyrosine kinase signaling pathway
9725	1.88E-09	15	50	response to hormone stimulus
42127	2.51E-09	19	50	regulation of cell proliferation
45787	9.23E-09	8	50	positive regulation of cell cycle
7173	2.56E-08	6	50	epidermal growth factor receptor signaling pathway
43687	2.83E-08	21	50	post-translational protein modification
6468	2.83E-08	16	50	protein amino acid phosphorylation
23034	3.72E-08	20	50	intracellular signaling pathway
51173	3.83E-08	16	50	positive regulation of nitrogen compound metabolic process
10647	6.40E-08	13	50	positive regulation of cell communication
45740	7.51E-08	6	50	positive regulation of DNA replication
7167	9.29E-08	12	50	enzyme linked receptor protein signaling pathway
22403	9.94E-08	13	50	cell cycle phase
279	9.94E-08	12	50	M phase
70	1.06E-07	6	50	mitotic sister chromatid segregation
10551	1.13E-07	9	50	regulation of gene-specific transcription from RNA polymerase II promoter
819	1.19E-07	6	50	sister chromatid segregation
8284	1.41E-07	13	50	positive regulation of cell proliferation
10646	1.41E-07	19	50	regulation of cell communication
16310	1.62E-07	16	50	phosphorylation
35468	1.62E-07	12	50	positive regulation of signaling pathway
42221	1.62E-07	21	50	response to chemical stimulus

Tabela 10 – Ontologias gênicas do Cluster 4 - tabela apresenta GO-ID, valor de p corrigido, x (o número de nós do cluster anotado para esta ontologia), X (número total de genes no cluster) e descrição dos processos biológico

GO-ID	corr p-value	x	X	Description
50793	1.68E-07	16	50	regulation of developmental process
6323	2.06E-07	8	50	DNA packaging
22402	2.11E-07	14	50	cell cycle process
51338	2.41E-07	12	50	regulation of transferase activity
22612	2.73E-07	7	50	gland morphogenesis
6464	2.95E-07	21	50	protein modification process
7059	3.66E-07	7	50	chromosome segregation
43434	3.66E-07	9	50	response to peptide hormone stimulus
32582	4.02E-07	7	50	negative regulation of gene-specific transcription
45840	4.05E-07	5	50	positive regulation of mitosis
51785	4.05E-07	5	50	positive regulation of nuclear division
48732	4.44E-07	9	50	gland development
30879	4.46E-07	7	50	mammary gland development
9888	4.84E-07	15	50	tissue development
71103	5.50E-07	8	50	DNA conformation change
43405	9.70E-07	8	50	regulation of MAP kinase activity
60443	1.04E-06	5	50	mammary gland morphogenesis
43549	1.17E-06	11	50	regulation of kinase activity

Tabela 11 – Ontologias gênicas do Cluster 4 (continuação) - tabela apresenta GO-ID, valor de p, valor de p corrigido, x (é o número de genes que apresentam essa ontologia), X (número total de genes no cluster) e descrição dos processos biológico

Algumas ontologias ganham destaque devido ao fato de não serem tão genéricas e possivelmente estarem envolvidas com o câncer. Na tabelas de processos biológicos do cluster 4, há a via de sinalização do fator de crescimento epidermal, cujos genes do cluster que estão envolvidos são: SHC1, EGF, GRB2, CBL, EGFR, EREG. O fator de crescimento epidermal é o gene EGFR, que com frequência se apresenta desregulado no cânceres de pulmão, especialmente no adenocarcinoma (ROWINSKY, 2004). Esta via de sinalização influencia fortemente na proliferação das células epidermais e na sobrevivência celular durante os processos de organogênese e no reparo de tecidos (ANAGNOSTIS et al., 2013; MARINAŞ et al., 2012; GOLDKORN; FILOSTO, 2010; LEVANTINI et al., 2022). A via de sinalização em questão, em contexto oncológico, pode ativar respostas celulares anti-apoptóticas e pró-sobrevivência: proliferação, motilidade, angiogênese, mimetismo vasculogênico e invasão (PETER et al., 2009; SATO et al., 2007; WEIHUA et al., 2008; PRENZEL et al., 2001; MINDER et al., 2015; LEVANTINI et al., 2022).

Outro mecanismo que chama a atenção é a modificação pós traducional de proteínas. Essas modificações podem alterar a carga, hidrofobicidade e a estabilidade de uma proteína e portanto modificar sua função (HAN et al., 2018), controlando quase todos os processos fisiológicos, como a função imune, a duração e local do processo e a intensidade (LIU;

QIAN; CAO, 2016). A proteína p53 pode passar por diveras modificações (fosforilação, ubiquitinação, sumoilação, acetilação, e metilação). Por exemplo, a sumoilação (ligação a uma proteína SUMO, presente no cluster) pode aumentar a atividade transcricional da p53 (DAI; GU, 2010). Outros processos são essenciais para a proliferação celular, como a condensação mitótica dos chromossomos, a regulação positiva da replicação do DNA, a regulação positiva do ciclo celular.

A via de sinalização MAPK que também aparece nas ontologias gênicas do cluster 4 é bastante estudada em relação ao câncer, sendo ela associada a processos de proliferação, diferenciação, apoptose e resposta ao estresse (GUO et al., 2020).

Os genes que se encontram nos módulos participam de diversos processos biológicos como o reparo do DNA, processo epitélio-mesenquimal, rearranjos de cromatina e ativação da transcrição, biogênese de ribossomos e proliferação celular, sinalizações apoptóticas, metástase, angiogênese, infiltração de células imunes, entre outros.

Nota-se nesta sessão que os genes da rede resultantes da análise de ML já foram intensamente explorados na literatura por sua associação ao câncer. Ou seja, há um embasamento para dizer que eles e também grande parte das ontologias gênicas de processos biológicos são importantes para a patologia e são potenciais biomarcadores. Na próximas sessão, é discutido a importância de cada SNP na patologia dos resultados da análise de variantes pelo Variantspark e qual a sua relação com os genes e ontologias aqui encontrados. Assim, os dados de mutações e espressão serão integrados a partir dos resultados da seleção de SNPs.

5.1.1 Análise de ML para seleção de variantes gênicas

Os primeiros 10 mais relvantes SNPs foram selecionados a partir da análise de mutações proporcionada pelo variantspark. Era esperado que os genes nos quais elas ocorrem tivessem uma intersecção com os resultados gerados a partir dos dados de expressão gênica. Assim, as mutações possivelmente explicariam a regulação gênica responsável pelo distinto padrão de expressão que apresentam as células de tecido tumoral em relação às células de tecido normal e aumentar a confiança da importância do gene e vias de sinalização para o desenvolvimento ou progressão da patologia.

Entretanto, nenhum dos genes está presente na análise conduzida sobre os dados de expressão gênica (68 genes que foram selecionados ou nas redes de interação). Nenhuma das variantes foi predita como maligna pelo SIFT ou PolyPhen, além disso, nenhuma tem significado clínico já descrito no ClinVar. É possível que existam limitações na análise de variantes devido aos dados utilizados, que foram de RNA-seq. Ao contrário do DNA-seq, o RNA-seq não tem cobertura homogênea do genoma inteiro por causa do nível de expressão de cada gene, que é variável (JEHL et al., 2021).

chr	start (bp)	SNP	rsID	genes
X	71465008	A-G		TAF1
7	150331699	A-G		LRRC61
1	85256124	T-C		C1orf52
17	75944359	T-C		ACOX1, FBF1
16	29813167	T-C		MAZ, PRRT2, PAGR1, MVP-DT
13	21496582	T-C	rs1034752992	MICU2
2	177219923	A-G		HNRNPA3, NFE2L2
12	120701721	A-G	rs574808706	MLEC
1	155133113	A-G	rs1664274631	EFNA1, SLC50A1
6	150249642	A-G		PPP1R14C

Tabela 12 – Variantes encontradas a partir da análise de ML realizada por meio do VariantSpark. Os dados da tabela mostram, da esquerda para a direita, o cromossomo onde se encontra a variante, o local onde ela está inserida (pares de bases), qual é o SNP que ocorre, o rsID (se tiver), e os genes afetados pela variante

O TAF1 é uma proteína envolvida centralmente na transcrição realizada pela polimerase II (CHENG et al., 2020), seu domínio N-terminal interage com proteínas que se ligam a TATA-box e isso é requerido para iniciação e ativação de genes sob a influência deste tipo de promotor (MAL et al., 2004). Conforme os resultados da análise pelo VEP, a variante no gene TAF1 não afeta o gene na parte codificante de proteína. A variante ocorreu um regiões como a 3' não traduzida, em introns, em transcritos que são alvos do decaimento, em exons não codificantes e à jusante do gene. Como predito pelo VEP (coluna *IMPACT* com a palavra *MODIFIER*), a partir deste tipo de variantes é difícil predizer o impacto que teriam na função ou regulação do gene. Entretanto, é possível que este impacto afete o câncer, já que TAF1 já foi descrita como influente na patalogia (WU et al., 2014).

O gene TAF1 também é um fator de transcrição para o gene TOP2A (ZHANG et al., 2019), que está presente no cluster 4 (20). TOP2A promove o deslocamento de beta-catenina para o núcleo, promovendo resistência ao paclitaxel (ZHANG et al., 2019), droga também usada no tratamento do CPCNP (PEREIRA, 2015). O RNA longo não codificante que promove a expressão de TAF1 (que promove a transcrição de TOP2A) é conhecido por ter um papel crítico em muitos carcinomas, incluindo o adenocarcinoma de pulmão (SHI et al., 2017; MARCHESE et al., 2016; ZHANG et al., 2019). Logo, se a mutação em TAF1 está aumentando sua atividade ou nível de transcrição e portanto aumentando o nível de transcrição de TOP2A, isso terá consequências negativas na patologia do câncer de pulmão.

O gene ACOX1 participa da via de beta oxidação do ácido graxo, a variante afetou ele à jusante e montante, em íntron e na região 3' UTR. Um estudo descobriu que ACOX1

quando defosforilado é responsável por regular a palmitoilação da β-catenina, inibindo sua ubiquitinação, portando ele é um supressor de tumor (ZHANG et al., 2023). Este gene é mais um que está envolvido na via de sinalização da β-catenina assim como ERCC6L, TOP2A e THRA,(HUANG et al., 2022; WANG et al., 2022; GIOLITO et al., 2022), genes que estão no cluster 4.

FBF1 regula diferentes processos envolvendo o centrossomo, como a duplicação dos centríolos, a separação do centrossomo, formação de cílios e polaridade celular (INOKO et al., 2018; SUGIMOTO et al., 2008). Importantes funções do centrossomo são a formação do fuso mitótico e separação dos cromossomos durante a divisão celular (INOKO et al., 2018), portanto as funções do centrossomo são críticas para a proliferação e diferenciação celular (AKIYAMA et al., 2017; INOKO et al., 2018). Aberrações na função do centrossomo pode causar instabilidade dos cromossomos, portanto causando câncer e outras doenças (CONDUIT; WAINMAN; RAFF, 2015; HILDEBRANDT; BENZING; KATSANIS, 2011; NIGG; ČAJÁNEK; ARQUINT, 2014; REITER; LEROUX, 2017; INOKO et al., 2018).

A variante afeta o gene na sua região montante apenas, mais a jusante da região regulatória, por isso é difícil dizer se ela afetaria a regulação do gene ou a sua função. Entretanto, caso afetado, este gene poderia causar distúrbios na proliferação celular, diferenciação celular, estabilidade dos cromossomos e divisão celular, contribuíndo na formação e progressão do câncer. No cluster 4 (10) os genes presentes têm processos biológicos relacionados ao do gene FBF1, como segregação dos cromossomos, regulação positiva da mitose e da divisão nuclear, entre outros, mostrando a relevância deste processo para a patologia.

A próxima mutação (16:29813167) tem uma COSMIC ID associado a ela: COSV99570009 e um dos fenótipos associados a esta mutação é o tumor do trato biliar, contudo não é explicitado à qual gene está atrelado este fenótipo. Entretanto, diferente do VEP, ele só mostra consequências para esta mutação nos genes MVP-DT e PRRT2. Poucas coisas sobre o PRRT2 associado ao câncer existem na literatura, contudo um estudo identificou que o gene tem uma sub-expressão nos cânceres de prostata, pulmão e gástrico e que mutações neste gene eram frequentes para o câncer de ovário, colorretal e endometrial (ALVES et al., 2017). Eles também verificaram que PRRT2 quando mutado (mutação em um microsatélite) se comportava como um oncogene dominante, enquanto o selvagem como um supressor de tumor (ALVES et al., 2017). A mutação nesse gene é missense e está atrelada a um impacto moderado. Pode-se hipotetizar que a proteína pode ter perdido sua eficácia por causa da mutação missense e perdido um pouco de sua capacidade como supressor de tumor e por isso a sua importância na patologia.

A variante de MICU2 tem um rsID (rs1034752992) e não há muita informação sobre ela. Esta proteína está encolvida na importação de cálcio para a mitocôndria e na regulação negativa da concentração do íon cálcio dentro da mitocôndria. A disrupção deste

ciclo está implicado em várias doenças, dentre elas o câncer (GARBINCIUS; ELROD, 2022). O processo biológico de transporte de cálcio está presente no cluster 4.

A atividade do gene HNRNPA3 está associado com o transporte de outro gene, assim como ocorre com TOP2A e a β-catenina. Nesse caso, HNRNPA3 (hnRNP A3) quando subexpresso diminui a quantidade de EGFR no núcleo (WANG et al., 2020), gene esse cuja importância é alta para o CPNPC, considerando sua presença tanto no cluster 4 quanto nos HBs. Diminuindo a quantidade de EGRF no núcleo também há diminuição da capacidade do tumor de crescer (WANG et al., 2020). A variante se encaixa na categoria de impacto MODIFIER, onde é difícil prever os impactos.

O EFNA1, assim como EGFR e ERBB2 (presentes no cluster 4) é um receptor tirosina-quinase e também faz parte de processos relacionados ao desenvolvimento como aqueles que aparecem no cluster 4 (tabelas 10 e 11). Seu papel no câncer é amplamente descrito na literatura, participando de processos como angigênese tumoral, eventos celulares malignos e a capacidade de invasão (HAO; LI, 2020). A variante afetou EFNA1 em íntrons e RNA não-codificante.

SLC50A1 também é um gene que já foi implicado no câncer e tem seu papel no transporte de açúcar, que são vias que se mostraram desreguladas no câncer (WANG et al., 2019) e também estão presentes nos processos biológicos do CLuster 4. Esta variante afetou o gene a jusante.

O gene PPP1R14C também conhecido como KEPI, inibe a proteina fosfatase 1 (PP1), e a sua inibição induz EGR1 (pela via MEK-ERK MAPK), que por sua vez, regula diretamente PTEN (WENZEL et al., 2007). PTEN é parte dos HBs da rede e é um supressor de tumor (WENZEL et al., 2007).

Ao fim desta sessão, é visto que apesar de nenhum dos SNPs ocorrerem em genes da rede de interação, eles ocorrem em genes que compartilham vias de sinalização com aqueles, indicando a relevância desses processos biológicos para o diagnóstico de adenocarcinoma de pulmão, já que o propósito destas análises era distinguir amostras de tecido tumoral de tecido normal. No entanto, é possível que a análise com o VariantSpark não resultou nas variantes mais relevantes quanto poderia se não existisse um desbalanço de quantidade de SNPs entre as classes (câncer e normal). Assim, devido a ausência de dados de mutações de uma das partes, aqueles SNPs que teriam mais importância na distinção entre as classes foram desconsiderados pelo algoritmo. Também podem ter contribuído para a diminuição de dados a filtragem de SNVs que são também SNPs (ocorrem em pelo menos 1% da população), assim, o câncer sendo uma doença heterogênea, aumentaria a dificuldade do algoritmo em encontrar um padrão para diferenciar as amostras a partir de suas mutações.

5.2 Conclusão

A partir de integração das análises de ML com dados de expressão e mutação, selecionando genes e SNPs, foram obtidas novas possibilidades de biomarcadores para o adenocarcinoma de pulmão. Quando analisados, os resultados apontam para possíveis biomarcadores:

1. Para uma mutação em TAF1, cujo gene é necessário para transcrição de outro gene, TOP2A selecionado a partir de dados de sua expressão (Cluster 4). TOP2A já foi estabelecido por ter papel crítico em carcinomas de pulmão. Por isso mais estudos são necessários para estabelecer se há relação entre a mutação e a expressão de TAF1 e TOP2A.

2. Vários genes encontrados nas análises participam da via da β-catenina, reforçando a importância dessa via para o câncer de pulmão.

3. A via da divisão celular, que participam o gene encontrado na sessão de SNPs FBF1 e genes do cluster 4.

4. A mutação no gene PRRT2, que apesar de estar relacionada com alguns tipos de câncer no COSMIC e ser considerada um supressor de tumor, ainda necessita de mais estudos para entender seu papel no adenocarcinoma de pulmão.

5. O processo de biológico de transporte de cálcio, que foi apontado tanto no cluster 4, quanto para o gene de uma das mutações que apresenta rsID (MICU2).

6. A mutação em HNRNPA3, um supressor de tumor com a função de transporte de EGFR, removendo ele do núcleo, sendo que EGFR é um dos HBs da rede de interação de proteínas, além de ser há muito considerado relevante para o adenocarcinoma de pulmão e outros cânceres de pulmão.

7. A mutação de PPP1R14C e sua relação com outros genes supressores de tumores.

Referências

AKHTAR, N.; BANSAL, J. G. Risk factors of lung cancer in nonsmoker. *Current problems in cancer*, Elsevier, v. 41, n. 5, p. 328–339, 2017. Citado na página 13.

AKIYAMA, T. et al. Shg-specificity of cellular rootletin filaments enables naive imaging with universal conservation. *Scientific reports*, Nature Publishing Group UK London, v. 7, n. 1, p. 39967, 2017. Citado na página 55.

ALEKSEYEV, Y. O. et al. A next-generation sequencing primer—how does it work and what can it do? *Academic pathology*, SAGE Publications Sage CA: Los Angeles, CA, v. 5, p. 2374289518766521, 2018. Citado na página 6.

ALPAYDIN, E. *Introduction to machine learning — 3rd ed.* The MIT Press, 2014. ISBN 978-0-262-02818-9. Disponível em: <https://www.google.com.br/books/edition/Introduction_to_Machine_Learning/NP5bBAAAQBAJ?hl=pt-BR&gbpv=0>. Citado na página 27.

ALVES, I. T. et al. A mononucleotide repeat in prrt2 is an important, frequent target of mismatch repair deficiency in cancer. *Oncotarget*, Impact Journals, LLC, v. 8, n. 4, p. 6043, 2017. Citado na página 55.

AMIN, M. B. et al. *AJCC cancer staging manual*. [S.l.]: Springer, 2017. v. 1024. Citado 3 vezes nas páginas 14, 16 e 19.

ANAGNOSTIS, A. et al. Molecular profiling of egfr family in chronic obstructive pulmonary disease: correlation with airway obstruction. *European Journal of Clinical Investigation*, Wiley Online Library, v. 43, n. 12, p. 1299–1306, 2013. Citado na página 52.

ANDREWS, S. et al. *FastQC*. Babraham, UK: [s.n.], 2012. Babraham Institute. Citado na página 34.

ANTUNES, J. L. et al. Sex and socioeconomic inequalities of lung cancer mortality in barcelona, spain and são paulo, brazil. *European Journal of Cancer Prevention*, LWW, v. 17, n. 5, p. 399–405, 2008. Citado na página 12.

ARJMAND, B. et al. Machine learning: a new prospect in multi-omics data analysis of cancer. *Frontiers in Genetics*, Frontiers Media SA, v. 13, 2022. Citado 4 vezes nas páginas 10, 11, 27 e 28.

BABA, A. et al. Pka-dependent regulation of the histone lysine demethylase complex phf2–arid5b. *Nature cell biology*, Nature Publishing Group UK London, v. 13, n. 6, p. 668–675, 2011. Citado na página 43.

BADER, G. D.; HOGUE, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, BioMed Central, v. 4, n. 1, p. 1–27, 2003. Citado na página 36.

BENJAMIN, D. et al. Calling somatic snvs and indels with mutect2. *BioRxiv*, Cold Spring Harbor Laboratory, p. 861054, 2019. Citado na página 35.

BIAN, M. et al. trna metabolism and lung cancer: beyond translation. *Frontiers in Molecular Biosciences*, Frontiers Media SA, v. 8, p. 659388, 2021. Citado na página 48.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, Oxford University Press, v. 30, n. 15, p. 2114–2120, 2014. Citado na página 34.

CAI, Z. et al. Machine learning for multi-omics data integration in cancer. *Iscience*, Elsevier, p. 103798, 2022. Citado 3 vezes nas páginas 26, 29 e 32.

CALVAYRAC, O. et al. Molecular biomarkers for lung adenocarcinoma. *European Respiratory Journal*, Eur Respiratory Soc, v. 49, n. 4, 2017. Citado na página 21.

CHATENOUD, L. et al. Trends in cancer mortality in brazil, 1980–2004. *European Journal of Cancer Prevention*, JSTOR, v. 19, n. 2, p. 79–86, 2010. Citado na página 12.

CHEN, D.; LIU, Q.; CAO, G. Ercc6l promotes cell growth and metastasis in gastric cancer through activating nf-κb signaling. *Aging (Albany NY)*, Impact Journals, LLC, v. 13, n. 16, p. 20218, 2021. Citado na página 50.

CHENG, H. et al. Missense variants in taf1 and developmental phenotypes: challenges of determining pathogenicity. *Human mutation*, Wiley Online Library, v. 41, n. 2, p. 449–464, 2020. Citado na página 54.

CIPOLLA-FICARRA, F. V.; QUIROGA, A.; FICARRA, M. C. Quality and web software engineering advances. In: *Handbook of research on software quality innovation in interactive systems*. [S.l.]: IGI Global, 2021. p. 41–82. Citado na página 27.

CONDUIT, P. T.; WAINMAN, A.; RAFF, J. W. Centrosome function and assembly in animal cells. *Nature reviews Molecular cell biology*, Nature Publishing Group UK London, v. 16, n. 10, p. 611–624, 2015. Citado na página 55.

COOPER, G. M. *The cell: A molecular approach*. Sinauer Associates, 2018. ISBN 9781605357713, 1605357715. Disponível em: <https://books.google.com.br/books?id=9ZSoAgAAQBAJ>. Citado na página 10.

DAI, C.; GU, W. p53 post-translational modification: deregulated in tumorigenesis. *Trends in molecular medicine*, Elsevier, v. 16, n. 11, p. 528–536, 2010. Citado na página 53.

DEPRISTO, M. A. et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, Nature Publishing Group, v. 43, n. 5, p. 491–498, 2011. Citado na página 35.

DI, X. et al. The oncogene iars2 promotes non-small cell lung cancer tumorigenesis by activating the akt/mtor pathway. *Frontiers in Oncology*, Frontiers Media SA, v. 9, p. 393, 2019. Citado na página 49.

DOBIN, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, v. 29, n. 1, p. 15–21, 10 2012. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/bts635>. Citado 2 vezes nas páginas 34 e 35.

DONG, Y. et al. Phosphorylation of phf2 by ampk releases the repressive h3k9me2 and inhibits cancer metastasis. *Signal Transduction and Targeted Therapy*, Nature Publishing Group UK London, v. 8, n. 1, p. 95, 2023. Citado na página 45.

DUFFY, M. J. Tumor markers in clinical practice: a review focusing on common solid cancers. *Medical Principles and Practice*, S. Karger AG Basel, Switzerland, v. 22, n. 1, p. 4–11, 2012. Citado na página 20.

DUFFY, M. J.; O'BYRNE, K. Tissue and blood biomarkers in lung cancer: a review. *Advances in clinical chemistry*, Elsevier, v. 86, p. 1–21, 2018. Citado 2 vezes nas páginas 20 e 21.

EID, R. A. et al. Integrative analysis of wdr12 as a potential prognostic and immunological biomarker in multiple human tumors. *Frontiers in Genetics*, Frontiers, v. 13, p. 1008502, 2023. Citado 2 vezes nas páginas 46 e 47.

FABRIS, A. et al. Proteomic-based research strategy identified laminin subunit alpha 2 as a potential urinary-specific biomarker for the medullary sponge kidney disease. *Kidney international*, Elsevier, v. 91, n. 2, p. 459–468, 2017. Citado na página 27.

FERLAY, J. et al. Cancer statistics for the year 2020: An overview. *International journal of cancer*, Wiley Online Library, v. 149, n. 4, p. 778–789, 2021. Citado na página 12.

FU, P. et al. The different functions and clinical significances of caveolin-1 in human adenocarcinoma and squamous cell carcinoma. *OncoTargets and therapy*, Taylor & Francis, p. 819–835, 2017. Citado na página 51.

GARBINCIUS, J. F.; ELROD, J. W. Mitochondrial calcium exchange in physiology and disease. *Physiological reviews*, American Physiological Society Rockville, MD, v. 102, n. 2, p. 893–992, 2022. Citado na página 56.

GIOLITO, M. V. et al. Regulation of the thra gene, encoding the thyroid hormone nuclear receptor trα1, in intestinal lesions. *Molecular Oncology*, Wiley Online Library, v. 16, n. 22, p. 3975–3993, 2022. Citado na página 55.

GOLDKORN, T.; FILOSTO, S. Lung injury and cancer: mechanistic insights into ceramide and egfr signaling under cigarette smoke. *American journal of respiratory cell and molecular biology*, American Thoracic Society, v. 43, n. 3, p. 259–268, 2010. Citado na página 52.

GONZÁLEZ-SANCHO, J. M. et al. Thyroid hormone receptors/thr genes in human cancer. *Cancer letters*, Elsevier, v. 192, n. 2, p. 121–132, 2003. Citado na página 50.

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, Nature Publishing Group, v. 17, n. 6, p. 333–351, 2016. Citado na página 7.

GREBER, B. J.; BAN, N. Structure and function of the mitochondrial ribosome. *Annual review of biochemistry*, Annual Reviews, v. 85, p. 103–132, 2016. Citado na página 49.

GREENER, J. G. et al. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, Nature Publishing Group UK London, v. 23, n. 1, p. 40–55, 2022. Citado 3 vezes nas páginas 28, 29 e 30.

GUO, Y.-J. et al. Erk/mapk signalling pathway and tumorigenesis. *Experimental and therapeutic medicine*, Spandidos Publications, v. 19, n. 3, p. 1997–2007, 2020. Citado na página 53.

HAN, Z.-J. et al. The post-translational modification, sumoylation, and cancer. *International journal of oncology*, Spandidos Publications, v. 52, n. 4, p. 1081–1094, 2018. Citado na página 52.

HANAHAN, D.; WEINBERG, R. A. Hallmarks of cancer: the next generation. *cell*, Elsevier, v. 144, n. 5, p. 646–674, 2011. Citado na página 9.

HAO, Y.; LI, G. Role of efna1 in tumorigenesis and prospects for cancer therapy. *Biomedicine & Pharmacotherapy*, Elsevier, v. 130, p. 110567, 2020. Citado na página 56.

HERBST, R. S.; MORGENSZTERN, D.; BOSHOFF, C. The biology and management of non-small cell lung cancer. *Nature*, Nature Publishing Group UK London, v. 553, n. 7689, p. 446–454, 2018. Citado 2 vezes nas páginas 21 e 22.

HILDEBRANDT, F.; BENZING, T.; KATSANIS, N. Ciliopathies. *New England Journal of Medicine*, Mass Medical Soc, v. 364, n. 16, p. 1533–1543, 2011. Citado na página 55.

HOLZEL, M. et al. Mammalian wdr12 is a novel member of the pes1–bop1 complex and is required for ribosome biogenesis and cell proliferation. *The Journal of cell biology*, Rockefeller University Press, v. 170, n. 3, p. 367–378, 2005. Citado na página 46.

HORTON, J. R. et al. Structural basis for human phf2 jumonji domain interaction with metal ions. *Journal of molecular biology*, Elsevier, v. 406, n. 1, p. 1–8, 2011. Citado na página 43.

HOSEOK, I.; CHO, J.-Y. Lung cancer biomarkers. *Advances in clinical chemistry*, Elsevier, v. 72, p. 107–170, 2015. Citado 2 vezes nas páginas 20 e 22.

HOU, G. et al. Ercc6l is a biomarker and therapeutic target for non–small cell lung adenocarcinoma. *Medical Oncology*, Springer, v. 39, n. 5, p. 51, 2022. Citado na página 50.

HUANG, X. et al. Overexpression of ercc6l correlates with poor prognosis and confers malignant phenotypes of lung adenocarcinoma. *Oncology Reports*, Spandidos Publications, v. 48, n. 1, p. 1–18, 2022. Citado 2 vezes nas páginas 50 e 55.

IJ, H. Statistics versus machine learning. *Nat Methods*, v. 15, n. 4, p. 233, 2018. Citado na página 27.

INOKO, A. et al. Albatross/fbf1 contributes to both centriole duplication and centrosome separation. *Genes to Cells*, Wiley Online Library, v. 23, n. 12, p. 1023–1042, 2018. Citado na página 55.

JANG, J. Y. et al. Multiple micrornas as biomarkers for early breast cancer diagnosis. *Molecular and clinical oncology*, Spandidos Publications, v. 14, n. 2, p. 1–1, 2021. Citado na página 23.

JEHL, F. et al. Rna-seq data for reliable snp detection and genotype calling: interest for coding variant characterization and cis-regulation analysis by allele-specific expression in livestock species. *Frontiers in Genetics*, Frontiers Media SA, v. 12, p. 655707, 2021. Citado 2 vezes nas páginas 8 e 53.

JEMAL, A. et al. Global patterns of cancer incidence and mortality rates and trends. *Cancer epidemiology, biomarkers & prevention*, AACR, v. 19, n. 8, p. 1893–1907, 2010. Citado na página 12.

JIANG, T.; GRADUS, J. L.; ROSELLINI, A. J. Supervised machine learning: a brief primer. *Behavior Therapy*, Elsevier, v. 51, n. 5, p. 675–687, 2020. Citado na página 29.

KALEMKERIAN, G. P. et al. Molecular testing guideline for the selection of patients with lung cancer for treatment with targeted tyrosine kinase inhibitors: American society of clinical oncology endorsement of the college of american pathologists/international association for the study of lung cancer/association for molecular pathology clinical practice guideline update. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, American Society of Clinical Oncology, v. 36, n. 9, p. 911, 2018. Citado na página 20.

KARIM, L.; KOSMIDER, B.; BAHMED, K. Mitochondrial ribosomal stress in lung diseases. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, American Physiological Society Rockville, MD, v. 322, n. 4, p. L507–L517, 2022. Citado na página 49.

KASE, S. et al. Expression of e-cadherin and beta-catenin in human non-small cell lung cancer: Clinical significance and prognosis. *Lung Cancer*, v. 1, n. 29, p. 196, 2000. Citado na página 11.

KATSYV, I. et al. Eprs is a critical regulator of cell proliferation and estrogen signaling in er+ breast cancer. *Oncotarget*, Impact Journals, LLC, v. 7, n. 43, p. 69592, 2016. Citado na página 49.

KIM, H.-J.; MAITI, P.; BARRIENTOS, A. Mitochondrial ribosomes in cancer. In: ELSEVIER. *Seminars in cancer biology*. [S.l.], 2017. v. 47, p. 67–81. Citado na página 49.

KIM, J. H. et al. Evolution of the multi-trna synthetase complex and its role in cancer. *Journal of Biological Chemistry*, ASBMB, v. 294, n. 14, p. 5340–5351, 2019. Citado 2 vezes nas páginas 48 e 49.

KNIGHT, S. B. et al. Progress and prospects of early detection in lung cancer. *Open biology*, The Royal Society, v. 7, n. 9, p. 170070, 2017. Citado na página 19.

KOC, E. C. et al. A new face on apoptosis: death-associated protein 3 and pdcd9 are mitochondrial ribosomal proteins. *FEBS letters*, Wiley Online Library, v. 492, n. 1-2, p. 166–170, 2001. Citado na página 49.

KOONIN, M. Y. G. E. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Kluwer Academic, 2003. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK20263/>. Citado na página 7.

KUKURBA, K. R.; MONTGOMERY, S. B. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, Cold Spring Harbor Laboratory Press, v. 2015, n. 11, p. pdb–top084970, 2015. Citado na página 8.

KULASINGAM, V.; DIAMANDIS, E. P. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature clinical practice Oncology*, Nature Publishing Group UK London, v. 5, n. 10, p. 588–599, 2008. Citado na página 20.

LEE, J. H. et al. Histone demethylase gene phf2 is mutated in gastric and colorectal cancers. *Pathology & Oncology Research*, Springer, v. 23, p. 471–476, 2017. Citado na página 45.

LEE, K. et al. Phf2 histone demethylase acts as a tumor suppressor in association with p53 in cancer. *Oncogene*, Nature Publishing Group, v. 34, n. 22, p. 2897–2909, 2015. Citado 2 vezes nas páginas 44 e 45.

Lesk, A. M. *Introduction to Genomics (2nd Edition)*. Oxford University Press Inc., 2012. ISBN 9780199564354. Disponível em: <https://www.google.com.br/books/edition/Introduction_to_Genomics/WbScAQAAQBAJ?hl=pt-BR&gbpv=0>. Citado na página 24.

LEVANTINI, E. et al. Egfr signaling pathway as therapeutic target in human cancers. In: ELSEVIER. *Seminars in Cancer Biology*. [S.l.], 2022. v. 85, p. 253–275. Citado na página 52.

LI, B.; DEWEY, C. N. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, Springer, v. 12, p. 1–16, 2011. Citado na página 34.

LI, J.-L. et al. Integrative genomic analyses identify wdr12 as a novel oncogene involved in glioblastoma. *Journal of Cellular Physiology*, Wiley Online Library, v. 235, n. 10, p. 7344–7355, 2020. Citado na página 46.

LINDEMAN, N. I. et al. Updated molecular testing guideline for the selection of lung cancer patients for treatment with targeted tyrosine kinase inhibitors: guideline from the college of american pathologists, the international association for the study of lung cancer, and the association for molecular pathology. *Archives of pathology & laboratory medicine*, the College of American Pathologists, v. 142, n. 3, p. 321–346, 2018. Citado na página 20.

LIU, J.; QIAN, C.; CAO, X. Post-translational modification control of innate immunity. *Immunity*, Elsevier, v. 45, n. 1, p. 15–30, 2016. Citado na página 53.

LIU, J. et al. shrna knockdown of dna helicase ercc6l expression inhibits human breast cancer growth. *Molecular Medicine Reports*, Spandidos Publications, v. 18, n. 3, p. 3490–3496, 2018. Citado na página 50.

LIU, W. et al. Cav-1 promote lung cancer cell proliferation and invasion through lncrna hotair. *Gene*, Elsevier, v. 641, p. 335–340, 2018. Citado na página 51.

MAERE, S.; HEYMANS, K.; KUIPER, M. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, Oxford University Press, v. 21, n. 16, p. 3448–3449, 2005. Citado na página 36.

MAL, T. K. et al. Structural and functional characterization on the interaction of yeast tfiid subunit taf1 with tata-binding protein. *Journal of molecular biology*, Elsevier, v. 339, n. 4, p. 681–693, 2004. Citado na página 54.

MALARKEY, D. E.; HOENERHOFF, M.; MARONPOT, R. R. Carcinogenesis: mechanisms and manifestations. *Haschek and Rousseaux's Handbook of Toxicologic Pathology*, Elsevier, p. 107–146, 2013. Citado 2 vezes nas páginas 10 e 11.

MARCHESE, F. P. et al. A long noncoding rna regulates sister chromatid cohesion. *Molecular cell*, Elsevier, v. 63, n. 3, p. 397–407, 2016. Citado na página 54.

MARIAN, A. J. Clinical interpretation and management of genetic variants. *Basic to Translational Science*, American College of Cardiology Foundation Washington DC, v. 5, n. 10, p. 1029–1042, 2020. Citado 2 vezes nas páginas 24 e 25.

MARINAŞ, M. et al. Egfr, her2/neu and ki67 immunoexpression in serous ovarian tumors. *Rom J Morphol Embryol*, v. 53, n. 3, p. 563–567, 2012. Citado na página 52.

MENDELSOHN, J. et al. *The Molecular Basis of Cancer E-Book: The Molecular Basis of Cancer E-Book*. Elsevier Health Sciences, 2014. ISBN 9780323261968. Disponível em: <https://books.google.com.br/books?id=9ZSoAgAAQBAJ>. Citado 2 vezes nas páginas 10 e 11.

MI, L. et al. Suppression of ribosome biogenesis by targeting wd repeat domain 12 (wdr12) inhibits glioma stem-like cell growth. *Frontiers in oncology*, Frontiers Media SA, v. 11, p. 751792, 2021. Citado na página 46.

MINDER, P. et al. Egfr regulates the development and microarchitecture of intratumoral angiogenic vasculature capable of sustaining cancer cell intravasation. *Neoplasia*, Elsevier, v. 17, n. 8, p. 634–649, 2015. Citado na página 52.

MISRA, B. B. et al. Integrated omics: tools, advances and future approaches. *Journal of molecular endocrinology*, Bioscientifica Ltd, v. 62, n. 1, p. R21–R45, 2019. Citado na página 26.

MOHAMED, A. E. Comparative study of four supervised machine learning techniques for classification. *International Journal of Applied*, v. 7, n. 2, p. 1–15, 2017. Citado 2 vezes nas páginas 27 e 28.

MORDENTE, A. et al. Cancer biomarkers discovery and validation: state of the art, problems and future perspectives. *Advances in Cancer Biomarkers*, Springer, p. 9–26, 2015. Citado na página 11.

MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, Academic Press, v. 92, n. 5, p. 255–264, 2008. Citado na página 7.

NANAVATY, P.; ALVAREZ, M. S.; ALBERTS, W. M. Lung cancer screening: advantages, controversies, and applications. *Cancer control*, SAGE Publications Sage CA: Los Angeles, CA, v. 21, n. 1, p. 9–14, 2014. Citado na página 19.

NGIAM, K. Y.; KHOR, W. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, Elsevier, v. 20, n. 5, p. e262–e273, 2019. Citado na página 27.

NIGG, E. A.; ČAJÁNEK, L.; ARQUINT, C. The centrosome duplication cycle in health and disease. *FEBS letters*, Elsevier, v. 588, n. 15, p. 2366–2372, 2014. Citado na página 55.

NOORELDEEN, R.; BACH, H. Current and future development in lung cancer diagnosis. *International journal of molecular sciences*, MDPI, v. 22, n. 16, p. 8661, 2021. Citado 3 vezes nas páginas 15, 17 e 19.

O'BRIEN, A. R. et al. Variantspark: population scale clustering of genotype information. *BMC genomics*, Springer, v. 16, p. 1–9, 2015. Citado na página 35.

PAPPA, S. et al. Phf2 histone demethylase prevents dna damage and genome instability by controlling cell cycle progression of neural progenitors. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 116, n. 39, p. 19464–19473, 2019. Citado na página 45.

PECORARO, A. et al. Ribosome biogenesis and cancer: overview on ribosomal proteins. *International journal of molecular sciences*, MDPI, v. 22, n. 11, p. 5496, 2021. Citado na página 46.

PEI, Y.-f.; YIN, X.-m.; LIU, X.-q. Top2a induces malignant character of pancreatic cancer through activating β-catenin signaling pathway. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, Elsevier, v. 1864, n. 1, p. 197–207, 2018. Citado na página 50.

PELLETIER, J.; THOMAS, G.; VOLAREVIĆ, S. Ribosome biogenesis in cancer: new players and therapeutic avenues. *Nature Reviews Cancer*, Nature Publishing Group UK London, v. 18, n. 1, p. 51–63, 2018. Citado na página 46.

PEREIRA, R. técnico D. M. B. *Paclitax (paclitaxel): Solução Injetável*. São Paulo, Brasil, 2015. Citado na página 54.

PERVEZ, M. T. et al. A comprehensive review of performance of next-generation sequencing platforms. *BioMed Research International*, Hindawi, v. 2022, 2022. Citado 2 vezes nas páginas 6 e 7.

PETER, Y. et al. Epidermal growth factor receptor and claudin-2 participate in a549 permeability and remodeling: implications for non-small cell lung cancer tumor colonization. *Molecular Carcinogenesis: Published in cooperation with the University of Texas MD Anderson Cancer Center*, Wiley Online Library, v. 48, n. 6, p. 488–497, 2009. Citado na página 52.

Pevsner, J. *Bioinformatics and Functional Genomics (3nd Edition)*. Wiley, 2015. ISBN 9781118581780, 1118581784. Disponível em: <https://www.google.com.br/books/edition/Bioinformatics_and_Functional_Genomics/dgmeCAAAQBAJ?hl=pt-BR&gbpv=0>. Citado na página 24.

PICARD toolkit. [S.l.]: Broad Institute, 2019. <https://broadinstitute.github.io/picard/>. Citado na página 35.

PRENZEL, N. et al. The epidermal growth factor receptor family as a central element for cellular signal transduction and diversification. *Endocrine-related cancer*, Bioscientifica Ltd, v. 8, n. 1, p. 11–31, 2001. Citado na página 52.

RAJULA, H. S. R. et al. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, MDPI, v. 56, n. 9, p. 455, 2020. Citado na página 27.

RAMPINELLI, C. et al. Exposure to low dose computed tomography for lung cancer screening and risk of cancer: secondary analysis of trial data and risk-benefit analysis. *bmj*, British Medical Journal Publishing Group, v. 356, 2017. Citado na página 19.

REEL, P. S. et al. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, Elsevier, v. 49, p. 107739, 2021. Citado na página 27.

REITER, J. F.; LEROUX, M. R. Genes and molecular pathways underpinning ciliopathies. *Nature reviews Molecular cell biology*, Nature Publishing Group UK London, v. 18, n. 9, p. 533–547, 2017. Citado na página 55.

RIBATTI, D.; TAMMA, R.; ANNESE, T. Epithelial-mesenchymal transition in cancer: a historical overview. *Translational oncology*, Elsevier, v. 13, n. 6, p. 100773, 2020. Citado na página 11.

RITCHIE, M. D. et al. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, Nature Publishing Group UK London, v. 16, n. 2, p. 85–97, 2015. Citado 2 vezes nas páginas 26 e 32.

RIVERA, M. P. Lung cancer in women: the differences in epidemiology, biology and treatment outcomes. *Expert review of respiratory medicine*, Taylor & Francis, v. 3, n. 6, p. 627–634, 2009. Citado na página 13.

RODRIGUEZ-CANALES, J.; PARRA-CUENTAS, E.; WISTUBA, I. I. Diagnosis and molecular classification of lung cancer. In: _____. *Lung Cancer: Treatment and Research*. Cham: Springer International Publishing, 2016. p. 25–46. ISBN 978-3-319-40389-2. Disponível em: <https://doi.org/10.1007/978-3-319-40389-2_2>. Citado na página 17.

ROWINSKY, E. K. The erbb family: targets for therapeutic development against cancer and therapeutic strategies using monoclonal antibodies and tyrosine kinase inhibitors. *Annu. Rev. Med.*, Annual Reviews, v. 55, p. 433–457, 2004. Citado na página 52.

RUDDON, R. W. *Cancer biology*. [S.l.]: Oxford University Press, 2007. Citado na página 11.

SÁNCHEZ-ORTEGA, M.; CARRERA, A. C.; GARRIDO, A. Role of nrf2 in lung cancer. *Cells*, MDPI, v. 10, n. 8, p. 1879, 2021. Citado na página 18.

SATO, M. et al. A translational view of the molecular pathogenesis of lung cancer. *Journal of thoracic oncology*, Elsevier, v. 2, n. 4, p. 327–343, 2007. Citado na página 52.

SCARDONI, G. et al. Biological network analysis with centiscape: centralities and experimental dataset integration. *F1000Research*, Faculty of 1000 Ltd, v. 3, 2014. Citado na página 36.

SCATENA, R. *Advances in Cancer Biomarkers*. [S.l.]: Springer Dordrecht, 2015. ISBN 978-94-017-7215-0. Citado na página 21.

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. Citado 4 vezes nas páginas 12, 13, 15 e 17.

SCOTT, R. A. et al. Common genetic variants highlight the role of insulin resistance and body fat distribution in type 2 diabetes, independent of obesity. *Diabetes*, Am Diabetes Assoc, v. 63, n. 12, p. 4378–4387, 2014. Citado na página 24.

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. Citado na página 20.

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. Citado na página 36.

SHARMA, M. R. et al. Structure of the mammalian mitochondrial ribosome reveals an expanded functional role for its component proteins. *Cell*, Elsevier, v. 115, n. 1, p. 97–108, 2003. Citado na página 49.

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. Citado 2 vezes nas páginas 7 e 8.

SHERIF, Z. A.; SULTAN, A. S. Divergent control of cav-1 expression in non-cancerous li-fraumeni syndrome and human cancer cell lines. *Cancer Biology & Therapy*, Taylor & Francis, v. 14, n. 1, p. 29–38, 2013. Citado na página 51.

SHI, M. et al. Ddx11-as1 as potential therapy targets for human hepatocellular carcinoma. *Oncotarget*, Impact Journals, LLC, v. 8, n. 27, p. 44195, 2017. Citado na página 54.

SHI, Z. et al. Feature selection methods for protein biomarker discovery from proteomics or multiomics data. *Molecular & Cellular Proteomics*, ASBMB, v. 20, 2021. Citado na página 31.

SINHA, S. et al. Alterations in candidate genes phf2, fancc, ptch1 and xpa at chromosomal 9q22. 3 region: pathological significance in early-and late-onset breast carcinoma. *Molecular cancer*, Springer, v. 7, p. 1–13, 2008. Citado na página 45.

STENDER, J. D. et al. Control of proinflammatory gene programs by regulated trimethylation and demethylation of histone h4k20. *Molecular cell*, Elsevier, v. 48, n. 1, p. 28–38, 2012. Citado na página 45.

SUGIMOTO, M. et al. The keratin-binding protein albatross regulates polarization of epithelial cells. *The Journal of cell biology*, Rockefeller University Press, v. 183, n. 1, p. 19–28, 2008. Citado na página 55.

SZKLARCZYK, D. et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, Oxford University Press, v. 51, n. D1, p. D638–D646, 2023. Citado na página 36.

TALSETH-PALMER, B. A.; SCOTT, R. J. Genetic variation and its role in malignancy. *International journal of biomedical science: IJBS*, Master Publishing Group, v. 7, n. 3, p. 158, 2011. Citado na página 24.

TANNOCK, I. F.; HICKMAN, J. A. Limits to personalized cancer medicine. *N Engl J Med*, v. 375, n. 13, p. 1289–1294, 2016. Citado 3 vezes nas páginas 21, 26 e 32.

TEAM, N. L. S. T. R. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, Mass Medical Soc, v. 365, n. 5, p. 395–409, 2011. Citado na página 18.

THAI, A. et al. Seminar lung cancer. *Lancet*, v. 398, p. 535–554, 2021. Citado na página 12.

THANDRA, K. C. et al. Epidemiology of lung cancer. *Contemporary Oncology/Współczesna Onkologia*, Termedia, v. 25, n. 1, p. 45–52, 2021. Citado na página 12.

TIAN, T. et al. The zatt-top2a-pich axis drives extensive replication fork reversal to promote genome stability. *Molecular Cell*, Elsevier, v. 81, n. 1, p. 198–211, 2021. Citado na página 50.

TOUMAZIS, I. et al. Risk-based lung cancer screening: a systematic review. *Lung Cancer*, Elsevier, v. 147, p. 154–186, 2020. Citado na página 13.

VARGAS, A. J.; HARRIS, C. C. Biomarker development in the precision medicine era: lung cancer as a case study. *Nature Reviews Cancer*, Nature Publishing Group UK London, v. 16, n. 8, p. 525–537, 2016. Citado 2 vezes nas páginas 21 e 22.

VEGA, I. Alonso-de et al. Phf2 regulates homology-directed dna repair by controlling the resection of dna double strand breaks. *Nucleic Acids Research*, Oxford University Press, v. 48, n. 9, p. 4915–4927, 2020. Citado na página 44.

WANG, S. et al. Caveolin-1: an oxidative stress-related target for cancer prevention. *Oxidative medicine and cellular longevity*, Hindawi, v. 2017, 2017. Citado na página 51.

WANG, T.-H. et al. Profiling of subcellular egfr interactome reveals hnrnp a3 modulates nuclear egfr localization. *Oncogenesis*, Nature Publishing Group UK London, v. 9, n. 4, p. 40, 2020. Citado na página 56.

WANG, X. et al. Oncogenic role and potential regulatory mechanism of topoisomerase iiα in a pan-cancer analysis. *Scientific Reports*, Nature Publishing Group UK London, v. 12, n. 1, p. 11161, 2022. Citado 2 vezes nas páginas 50 e 55.

WANG, Y. et al. The novel sugar transporter slc50a1 as a potential serum-based diagnostic and prognostic biomarker for breast cancer. *Cancer Management and Research*, Taylor & Francis, p. 865–876, 2019. Citado na página 56.

WEIHUA, Z. et al. Survival of cancer cells is maintained by egfr independent of its kinase activity. *Cancer cell*, Elsevier, v. 13, n. 5, p. 385–393, 2008. Citado na página 52.

WEINBERG, R.; HANAHAN, D. et al. The hallmarks of cancer. *Cell*, v. 100, n. 1, p. 57–70, 2000. Citado na página 9.

WEINBERG, R. A.; WEINBERG, R. A. *The biology of cancer*. [S.l.]: WW Norton & Company, 2006. Citado na página 10.

WEN, H. et al. Recognition of histone h3k4 trimethylation by the plant homeodomain of phf2 modulates histone demethylation. *Journal of Biological Chemistry*, ASBMB, v. 285, n. 13, p. 9322–9326, 2010. Citado na página 43.

WENZEL, K. et al. Expression of the protein phosphatase 1 inhibitor kepi is downregulated in breast cancer cell lines and tissues and involved in the regulation of the tumor suppressor egr1 via the mek-erk pathway. Walter de Gruyter, 2007. Citado na página 56.

WETTERSTRAND, K. A. *The Cost of Sequencing a Human Genome*. 2021. Disponível em: <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>. Citado na página 7.

WU, F. et al. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nature communications*, Nature Publishing Group UK London, v. 12, n. 1, p. 2540, 2021. Citado na página 22.

WU, Y. et al. Phosphorylation of p53 by taf1 inactivates p53-dependent transcription in the dna damage response. *Molecular cell*, Elsevier, v. 53, n. 1, p. 63–74, 2014. Citado na página 54.

XIE, Y. et al. Ercc6l promotes cell growth and invasion in human colorectal cancer. *Oncology letters*, Spandidos Publications, v. 18, n. 1, p. 237–246, 2019. Citado na página 50.

XU, D. et al. Multi-scale supervised clustering-based feature selection for tumor classification and identification of biomarkers and targets on genomic data. *BMC genomics*, BioMed Central, v. 21, n. 1, p. 1–17, 2020. Citado na página 30.

XU, J.; LAMOUILLE, S.; DERYNCK, R. Tgf-β-induced epithelial to mesenchymal transition. *Cell research*, Nature Publishing Group, v. 19, n. 2, p. 156–172, 2009. Citado na página 11.

YANG, J. et al. Epigenetic regulation of megakaryocytic and erythroid differentiation by phf2 histone demethylase. *Journal of Cellular Physiology*, Wiley Online Library, v. 233, n. 9, p. 6841–6852, 2018. Citado na página 45.

YIN, Y. et al. Identification of wdr12 as a novel oncogene involved in hepatocellular carcinoma propagation. *Cancer Management and Research*, Taylor & Francis, p. 3985–3993, 2018. Citado na página 46.

YING, X. An overview of overfitting and its solutions. In: IOP PUBLISHING. *Journal of physics: Conference series*. [S.l.], 2019. v. 1168, p. 022022. Citado na página 30.

YOO, Y. A. et al. Mitochondrial ribosomal protein l41 suppresses cell growth in association with p53 and p27kip1. *Molecular and cellular biology*, Taylor & Francis, v. 25, n. 15, p. 6603–6616, 2005. Citado na página 49.

ZHANG, Q. et al. Reprogramming of palmitic acid induced by dephosphorylation of acox1 promotes β-catenin palmitoylation to drive colorectal cancer progression. *Cell Discovery*, Springer Nature Singapore Singapore, v. 9, n. 1, p. 26, 2023. Citado na página 55.

ZHANG, S. et al. The resistance of esophageal cancer cells to paclitaxel can be reduced by the knockdown of long noncoding rna ddx11-as1 through taf1/top2a inhibition. *American journal of cancer research*, e-Century Publishing Corporation, v. 9, n. 10, p. 2233, 2019. Citado na página 54.

ZHANG, X. et al. T3 promotes glioma cell senescence and apoptosis via thra and thrb. *Journal of Environmental Pathology, Toxicology and Oncology*, Begel House Inc., v. 40, n. 4, 2021. Citado na página 50.

ZHOU, W.; CHRISTIANI, D. C. East meets west: ethnic differences in epidemiology and clinical behaviors of lung cancer between east asians and caucasians. *Chinese Journal of Cancer*, BioMed Central, v. 30, n. 5, p. 287, 2011. Citado na página 22.

ZHOU, Z. et al. Roles of aminoacyl-trna synthetase-interacting multi-functional proteins in physiology and cancer. *Cell Death & Disease*, Nature Publishing Group UK London, v. 11, n. 7, p. 579, 2020. Citado na página 48.