UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

KARLA CRISTINA TABOSA MACHADO

# INTEGRAÇÃO DE DADOS PROTEÔMICOS DE TECIDOS HUMANOS PARA POTENCIALIZAR A DESCOBERTA DE GENES COMO ALVOS MOLECULARES A NÍVEL PROTEICO

TRABALHO DE CONCLUSÃO DE CURSO DE ESPECIALIZAÇÃO

DOIS VIZINHOS
2022

KARLA CRISTINA TABOSA MACHADO

# INTEGRAÇÃO DE DADOS PROTEÔMICOS DE TECIDOS HUMANOS PARA POTENCIALIZAR A DESCOBERTA DE GENES COMO ALVOS MOLECULARES A NÍVEL PROTEICO

# INTEGRATION OF PROTEOMIC DATA FROM HUMAN TISSUES TO ENHANCE THE DISCOVERY OF GENES AS MOLECULAR TARGETS AT THE PROTEIN LEVEL

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Anderson Chaves Carniel

DOIS VIZINHOS
2022

KARLA CRISTINA TABOSA MACHADO

# INTEGRAÇÃO DE DADOS PROTEÔMICOS DE TECIDOS HUMANOS PARA POTENCIALIZAR A DESCOBERTA DE GENES COMO ALVOS MOLECULARES A NÍVEL PROTEICO

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Data de aprovação: 06/outubro/2022

Anderson Chaves Carniel
Doutorado
Universidade Federal de São Carlos

Rafael Alves Paes de Oliveira
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

Yuri Kaszubowski Lopes
Doutorado
Universidade do Estado de Santa Catarina

DOIS VIZINHOS
2022

## AGRADECIMENTOS

A Deus. Por permitir a conclusão de mais uma etapa importante em minha vida. Por ter guiado os meus passos e ter me mostrado o caminho certo a trilhar durante esta jornada.

Aos meus pais, pelo apoio incondicional, dedicação e comprometimento com a minha educação.

Ao meu namorado, por todo cuidado e paciência nessa reta final desse trabalho.

Ao meu orientador, o professor Dr. Anderson Carniel, por toda dedicação, comprometimento, responsabilidade, compreensão e paciência durante a orientação desse trabalho.

A todos que, de alguma maneira, contribuíram para a realização desse trabalho.

## RESUMO

A proteômica é uma área do conhecimento responsável por analisar todo o conjunto de dados proteicos sintetizados pelo genoma, bem como suas modificações pós-traducionais. Essa área tem passado por grandes avanços tecnológicos na última década em termos de sensibilidade e capacidade de sequenciamento, em consequência, a quantidade de dados proteômicos disponibilizados em repositórios públicos tem aumentado significativamente, possibilitando a integração e análise exploratória de tais dados. Tradicionalmente, análises genômicas e transcriptômicas foram utilizadas para o entendimento do perfil de todo ambiente tumoral. No entanto, o estudo do genoma e do transcriptoma não são suficientes para elucidar os mecanismos moleculares de uma célula, pois a função ocorre em sua maioria a nível proteico. Além disso, a quantidade de mRNA em uma célula não é necessariamente proporcional ao nível de proteína codificado. Com o desenvolvimento das tecnologias proteômicas, tornou-se possível utilizar o proteoma para explorar a caracterização molecular do câncer, bem como para revelar novos biomarcadores de proteínas. Uma dificuldade é que, enquanto estudos transcriptômicos trabalham com centenas de amostras, de células ou tecidos, os estudos proteômicos trabalham com poucas amostras. A integração de dados proteômicos de vários estudos pode solucionar esta problemática, proporcionando uma visão mais global do número das amostras biológicas. O objetivo deste trabalho é integrar dados proteômicos públicos de tecidos humanos de vários estudos e realizar uma análise exploratória desses dados por meio da descoberta de proteínas abundantes em tecidos tumorais, a fim de potencializar a identificação de genes a nível proteico como alvos moleculares para o câncer. Durante o pré-processamento, foram realizados os processos de limpeza, integração e transformação dos dados. Foram processados aproximadamente 10 Tb de dados proteômicos, contendo mais de 500 amostras de tecidos saudáveis, tumores coletados de pacientes e de linhagens celulares imortalizadas usadas como modelo em câncer. Com a integração de dados proteômicos, amostras biológicas de diferentes estudos foram classificadas de acordo com tecido de origem e agrupadas, com o intuito de amplificar o número de amostras por tecido. O agrupamento revelou 140 amostras de tecidos saudáveis, que foram agrupadas em um único tecido e, 385 amostras tumorais, que foram agrupadas de acordo com o tecido tumoral de origem. Em seguida, a variável que indica a abundância das proteínas nos tecidos biológicos foi normalizada e convertida em quatro tipos de categoria ("muito baixa", "baixa", "média", "alta"). Para realizar a identificação de proteínas como alvos moleculares para o câncer, a análise exploratória focou na caracterização de antígenos de câncer/testículo (CTAs) a nível proteico. Para isso, CTAs preditos anteriormente em trabalhos transcriptômicos foram considerados. Como resultado, a integração dos dados proteicos identificou 17200 proteínas únicas e a análise exploratória dos dados revelou 212 CTAs a nível de proteína, dos quais 40 foram categorizadas com expressão "média" em tecidos cancerígenas e "muito baixa" ou "baixa"no grupamento saudável. Em conclusão, a análise exploratória realizada neste trabalho apresenta potencial para permitir futuros avanços na caracterização de proteomas tumorais e consequentemente, na identificação de proteínas como alvos moleculares para o câncer.



**Palavras-chave**: proteína; integração de dados; alvo molecular; antígenos, câncer/testículo.

## ABSTRACT

Proteomics is is an area of knowledge responsible for analyze synthesized protein dataset by the genome, as well as its post-translational modifications. This area has undergone technological breakthroughs in the last decade regarding sensitivity and throughput, and as a result, the size of proteomic data available in public repository has increased significantly, enabling the integration and exploratory data analysis. Traditionally, genomics and transcriptomics analysis were used to understand the profile of the whole tumor environment. But genomic and transcriptomic studies are not sufficient to elucidate all molecular mechanisms in the cell, since function occurs mostly at the protein level. Moreover, the amount of mRNA is not necessarily proportional to the translated protein level. Proteomic approaches allowed to use the proteome to explore cancer molecular characterization, as well as reveal new biomarkers, leading toward personalized medicine. One challenge is that while transcriptomics studies can be done using hundreds of samples, from cells or tissues, proteomics studies work with few samples. The integrated proteomic data from multiple studies can solve this problem, allowing a more comprehensive view of the samples under analysis. This work aims to perform to integrate public protemic data from human tissues and perform an exploratory data analysis through the discovery of abundant proteins in tumoral tissue, in order to enhance the identification genes as molecular targes for cancer at the protein level. During pre-processing, data cleaning, integrating and transforming processes were carried out. Approximately 10 Tb of proteomic data were processed, containing more than 500 samples of healthy tissues, tumors collected from patients and immortalized cell lines used as a model in cancer. To integrate proteomic data, the samples were grouped according to the tissues to which they belonged, for the purpose of amplifying the sample number by tissue. Such clustering revealed 140 samples of healthy tissues were clustered into unique group and, 385 tumoral samples were clustered according to the tumoral tissue of origin. Then, the variable that indicates protein abundance in biological tissues was normalized and converted into four categories (“very low”, “low”, “medium”, “high”). To perform the identification of proteins as molecular targets for the cancer, the exploratory data analysis focused on cancer / testis antigens (CTAs) characterization at protein level. CTAs previously predicted in transcriptomics works were used. As a result, the protemic data integration identify 17200 unique proteins and the exploratory data analysis revealed 222 CTAs at the proteomic level, of which 40 were categorized with ”medium”expression in tumoral tissue and “very low” or “low” in healthy group. In conclusion, the exploratory data analysis performed in this study to show potential to enable future advances in the characterization of tumoral proteome and, consequently, in the identification of proteins as target molecular for the cancer.



**Keywords**: protein; data integration; molecular target; antigens; cancer/testis.

## LISTA DE FIGURAS

Figura 1 – Fluxo do processamento dos dados proteômicos. Durante o processamento, os dados passaram pela etapa de limpeza, integração e transformação dos dados. Como resultado do processamento, foi gerada uma base de dados proteômicos, possibilitando o desenvolvimento de análises biológicas. . . . . 20
Figura 2 – Heatmap com CTAs preditos em tecidos humanos a nível proteico exibidos por uma variável categórica - extremidade esquerda. O eixo x contém os genes detectados nos tecidos biológicos, o eixo y contém o grupamento saudável e os demais tecidos cancerígenas e a variável de resposta, representada pelas cores no heatmap identificam a categoria (“muito”, “baixa”, “baixa”, “média”, “alta”), reflete a intensidade da proteína dentro de cada tecido. Os pixels em branco indicam ausência da proteína no respectivo tecido. . . . . 25
Figura 3 – Heatmap com CTAs preditos em tecidos humanos a nível proteico exibidos por uma variável contínua da extremidade esquerda. O eixo x contém os genes detectados nos tecidos biológicos, o eixo y contém o grupamento saudável e os demais tecidos cancerígenas e a variável de resposta é representada pela mediana do *z-score*. Os pixels em branco indicam ausência da proteína no respectivo tecido. . . . . 26
Figura 4 – Heatmap do gene P75 em amostras biológicas do tecido câncer de ovário e do grupamento saudável. O eixo x contém todas as amostras analisadas para o gene, o eixo y contém os grupamentos câncer de ovário e saudável e a variável de resposta é representada pelo *z-score*. Os pixels em branco indicam ausência da proteína na respectiva amostra. . . . . 27

## LISTA DE TABELAS


Tabela 1 – Seleção da tabela grupo de proteínas, contendo apenas atributos de interesse para a análise. Cada linha contém o grupo de proteína, o nome do gene, a contagem de peptídeos e as intensidades dos grupos de proteínas nas amostras biológicas (A549, Hek293 e MCF7). . . . . 19

Tabela 2 – Seleção do resultado do processamento realizado pelo script *process_protein_groups.r*. Cada linha contém o grupo de proteínas identificadas, o nome dos genes, a contagem de peptídeos e as intensidades dos grupos de proteínas nas amostras biológicas (A549, Hek293 e MCF7). . . . . 21

Tabela 3 – Quantidade de amostras por tumor. . . . . 22

Tabela 4 – Seleção da saída do script *integrating_protein_groups.r*. Cada linha contém o identificador e nome do banco de dados, nome do tecido biológico, identificador e nome da amostra biológica, nome do gene e intensidade do gene. . . . . 23

Tabela 5 – Seleção da saída do script normalization_integrated_data.r Cada linha contém o identificador e nome do banco de dados, nome do tecido biológico, identificador e nome da amostra biológica, nome do gene e *Z-score*. . . . 24

Tabela 6 – Seleção da saída do script conversion_integrated_data.r. Cada linha contém o nome do gene, o tecido biológico, a mediana do z-score e a classe. . . . 24

Tabela 7 – Intervalo de valores e quantidades de amostras biológicas por categoria (baixa, média e alta). . . . . 25

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| CSV | Valores separados por vírgula (do inglês, comma-separated values) |
| CT | Câncer testículo |
| CTAs | Antígenos câncer/testículos (do inglês, cancer-testis antigens) |
| DNA | Ácido Desoxirribonucleico (do inglês, deoxyribonucleic acid) |
| MetaMSD | Dados de meta-análise por espectrometria de massa |
| mRNA | Ácido Ribonucleico Mensageiro (do inglês, messanger ribonucleic acid) |
| MS | Espectrometria de massas (do inglês, Mass spectrometry) |
| NA | Não disponível (do inglês, not avaliable) |
| NGS | Sequenciamento de Próxima Geração (do inglês, next-generation sequencing) |
| ORF | Quadro Aberto de Leitura (do inglês, open reading frame) |
| SQL | Linguagem de consulta padrão (do inglês, Standard query language) |
| PDAC | Adenocarcinoma ductal pancreático |
| TXT | Text (do inglês, texto) |
| XML | Linguagem de marcação expansível (do inglês, Extensible markup language) |

SUMÁRIO


1 INTRODUÇÃO . . . . . 10
1.1 Problema de Pesquisa . . . . . 12
1.2 Objetivos . . . . . 13
1.2.1 Objetivo Geral . . . . . 13
1.2.2 Objetivos Específicos . . . . . 13
1.3 Justificativa . . . . . 13
1.4 Organização da monografia . . . . . 14

2 FUNDAMENTAÇÃO TEÓRICA . . . . . 15
2.1 Coleta de dados . . . . . 15
2.2 Preparação de dados . . . . . 15
2.3 Análise exploratória de dados . . . . . 17

3 TRABALHOS RELACIONADOS . . . . . 18

4 PROCESSAMENTO E ANÁLISE DE DADOS PROTEÔMICOS . 19
4.1 Obtenção dos dados proteômicos . . . . . 19
4.2 Fluxo do processamento de dados . . . . . 20
4.2.1 Limpeza de dados . . . . . 20
4.2.2 Integração de dados . . . . . 21
4.2.3 Transformação de dados . . . . . 22
4.3 Análise exploratória dos dados proteômicos . . . . . 23

5 CONCLUSÕES E TRABALHOS FUTUROS . . . . . 28

REFERÊNCIAS . . . . . 29

# 1 INTRODUÇÃO

O genoma é o repositório de informações biológicas presentes em todos os organismos, sendo formado por ácidos desoxirribonucleicos (do inglês, *deoxyribonucleic acid - DNA*). O DNA serve como molde para a produção de todas as proteínas necessárias para a vida dos organismos (LEWIN, 2004). No entanto, o primeiro produto da expressão do genoma é o transcriptoma, constituído por um conjunto de ácidos ribonucleicos mensageiros (do inglês, *messanger ribonucleic acid, mRNA*) derivados dos genes que codificam proteínas. O segundo produto é o proteoma, composto pelo conjunto de proteínas que descrevem a natureza das reações bioquímicas que a célula pode realizar. O termo proteoma refere-se ao conjunto de proteínas que estão presentes em determinada célula, tecido ou organismo, em dado momento do desenvolvimento ou sob circunstâncias específicas (um estímulo celular ou do ambiente, por exemplo).

Em razão do rápido desenvolvimento das tecnologias de alto rendimento, dados biológicos em diferentes níveis (genômicos, transcriptômicos e proteômicos) aumentam diariamente sua disponibilidade em repositórios públicos (SHAFI et al., 2019). A quantidade desses dados biológicos disponíveis, abordando as mesmas doenças, possibilita que pesquisadores apliquem uma meta-análise a fim de integrar vários conjuntos de dados para obter uma visão abrangente do dado biológico, possibilitando a identificação de melhores marcadores prognósticos (BERGER; PENG; SINGH, 2013; KRISTENSEN et al., 2014; NGUYEN et al., 2016; SHAFI et al., 2019).

Na última década ocorreu um aumento do número de biomarcadores específicos de doenças (por exemplo, câncer) relatados por diferentes grupos de pesquisa, sendo a maioria dos experimentos baseados na análise de transcritos (SHAFI et al., 2019). Sabe-se que a análise de transcritos é uma abordagem padrão para identificar transcritos diferencialmente expressos em tecidos cancerígenos. Contudo, a análise de transcritos pode não ser suficiente para revelar o estado de uma célula, tecido ou organismo, em dado momento sob determinada doença, pois não permite a observação de modificações pós-transcricionais (como por exemplo, splicing alternativo), pós-traducionais (por exemplo, modificações pós-traducionais como fosforilação), além de complexos formados por essas biomoléculas. Ademais, a quantidade de mRNA em uma célula não é necessariamente proporcional ao nível de proteína codificado, pois os níveis de proteínas podem ser influenciados por mecanismos de controle pós-traducional além de mecanismos de regulação dos níveis de mRNA. Assim, não existe uma relação linear entre os genes e o proteoma de uma célula, bem como a existência de um quadro aberto de leitura (do inglês, open reading frame - ORF) em um gene não o caracteriza como funcional. Além disso, como as proteínas são mais dinâmicas que o DNA ou o mRNA, elas carregam mais informações que os ácidos nucléicos. Por esse motivo, considerar as informações derivadas do genoma em paralelo aos dados sobre as diferenças proteicas entre tecidos corporais, normais ou alterados,

tornou-se uma abordagem fundamental para a identificação de biomarcadores confiáveis.

A análise proteômica proporciona uma descrição quantitativa do estado de um sistema biológico por meio da análise quantitativa dos perfis de expressão de proteínas. Tal análise tem potencial para determinar propriedades de sistemas biológicos que não são aparentes apenas pela análise de sequência de DNA ou mRNA (GYGI et al., 1999). A combinação dos dados proteômicos, transcriptômicos e genômicos de amostras clínicas de tumores oferece uma oportunidade sem precedentes para a proteogenômica tumoral (EDWARDS et al., 2015). Inclusive, desenvolvimentos marcantes em tecnologias proteômicas tornou o proteoma um componente fundamental para a compreensão da biologia do câncer, permitindo avanços significativos no diagnóstico precoce e no prognóstico exato de muitos tipos de câncer. Além de revelar novos biomarcadores, conduzindo à medicina personalizada contra o câncer (HUANG et al., 2017).

O ponto chave da proteômica é a identificação de sequências proteicas utilizando dados espectrometria de massa (do inglês, *mass spectrometry - MS*). Essa técnica resume-se sobretudo na ionização de um composto e na avaliação da razão massa/carga (m/z) dos íons (BARBOSA et al., 2012). O produto da MS é um gráfico conhecido como espectro de massa contendo a relação massa/carga (eixo X) e a intensidade (eixo Y) dos íons detectados como uma função de sua razão massa/carga. Os picos encontrados nesses espectros são utilizados pelos softwares de busca para identificar as proteínas em bancos de dados. A análise de proteínas por espectrometria de massa (MS) experimentou grandes desenvolvimentos tecnológicos na última década em termos de sensibilidade e capacidade de sequenciamento. Como consequência, a quantidade de dados produzidos por laboratórios de proteômica aumentou de forma significativa. Adicionalmente, surgiu um consenso para que os pesquisadores compartilhassem seus dados de espectrometria de massa para facilitar a avaliação, reutilização e análise comparativa de tais dados (PEREZ-RIVEROL et al., 2015). A disponibilidade de grandes quantidades de dados proteômicos para o mesmo tipo de estudo, permitiu a integração desses dados por técnicas de meta-análise (GOVEIA et al., 2016). A meta-análise é uma estratégia eficiente para integrar e analisar dados científicos de vários estudos, com o potencial de melhorar a interpretação de um modelo, aumentar o poder estatístico e obter resultados mais confiáveis (GLASS, 1976; WU et al., 2019; HU; ZHOU; TONG, 2021). A meta-análise de dados proteômicos é uma modalidade promissora para explorar a caracterização molecular do câncer, possibilitando a descoberta de biomarcadores a nível de proteína.

Aqui, destacamos os antígenos de câncer / testículo (CTAs) como potenciais candidatos para a imunoterapia contra o câncer (JIN et al., 2018). Os CTAs são um grupo de proteínas que são tipicamente restritas ao testículo entre os tecidos normais, mas são expressas de forma aberrante em vários tipos de câncer. Esses antígenos são excelentes candidatos à imunoterapia contra o câncer, devido a expressão altamente restrita em tecidos normais e à capacidade de provocar uma resposta imune quando expressas em células cancerígenas (PARMIGIANI et al., 2006; CABALLERO; CHEN, 2009; LI et al., 2017). Os CTAs são amplamente reconhecidos

como alvos imunoterapeuticos promissores e existem vários ensaios clínicos em andamento (KRISHNADAS et al., 2015; SILVA et al., 2017). Devido ao seu padrão de expressão restrito, os CTAs representam candidatos a biomarcadores únicos para o diagnóstico / prognóstico do câncer (KULKARNI; SCIENCES; 2017, ).

Muitos estudos transcriptômicos permitiram um exame mais aprofundado de CTAs em potencial para a imunoterapia contra o câncer (WANG et al., 2016; SILVA et al., 2017). Em contraste com a extensa pesquisa transcriptômica, a análise proteômica profunda de CTAs ainda está em desenvolvimento, mas como a camada proteômica reflete com muito mais precisão a função celular, possivelmente a análise a nível proteômico pode revelar novos aspectos que ainda não foram observados a nível transcriptômico. Decidiu-se então realizar a integração de dados proteômicos de tecidos humanos para explorar a expressão de CTAs a nível proteico.

## 1.1 Problema de Pesquisa

Com o desenvolvimento de sequenciadores de próxima geração (do inglês, *next-generation sequencing - NGS*), uma revolução ocorreu na pesquisa genômica, permitindo que informações em diferentes níveis (genômico e transcriptômico) aumentassem diariamente. Enquanto as tecnologias de sequenciamento de ácidos nucleicos desenvolveram-se a uma velocidade impressionante, o mesmo não ocorreu com as tecnologias de sequenciamento de proteínas. Somente nos últimos 5 a 6 anos os estudos de proteoma baseado em espectrometria de massa (MS) desenvolveram-se suficientemente. Em alguns casos, a proteômica já é considerada uma tecnologia de alto rendimento e abrangência (MANN; KELLEHER, 2008), sendo uma estratégia promissora para o estudo da expressão gênica de uma célula ou tecido.

Uma das abordagens padrão para caracterizar alvos moleculares para o câncer é comparar perfis de transcrição da doença versus tecido normal para identificar transcritos diferencialmente expressos. Outra abordagem que recebeu destaque foi a identificação de genes amplificados nos tecidos cancerígenos. Por mais que sejam abordagens relevantes, elas não identificarão a grande maioria dos biomarcadores de proteínas que surgem por mecanismos pós-transcricionais e pós-traducionais. A identificação direta de proteínas que estão diferencialmente expressas no tecido tumoral está dentro da capacidade da MS (AEBERSOLD et al., 2005).

A análise de misturas de peptídeos por cromatografia líquida acoplada à MS em alta resolução surgiu como a principal tecnologia para a caracterização proteômica em larga escala, com publicações relatando a rotineira identificação de mais de quatro mil proteínas em experimentos simples (sem fracionamento) (BONALDI et al., 2008; GRAUMANN et al., 2008; KELSTRUP et al., 2012), ou de 8000 a 10000 proteínas quando um simples fracionamento da amostra é empregado (GEIGER et al., 2012; COSCIA et al., 2016; RIECKMANN et al., 2017a). Como consequência, o tamanho dos dados produzidos por laboratórios de proteômica cresceram significativamente.

No entanto, enquanto estudos transcriptômicos trabalham com centenas de amostras, de células ou tecidos, os estudos proteômicos trabalham com poucas amostras, mas essa

problemática pode ser solucionada por meio da integração de dados proteômicos, possibilitando uma visão mais global do número das amostras. Assim, a análise exploratória de dados proteômicos integrados pode fornecer informações valiosas que uma única análise de conjunto de dados não pode fornecer, melhorando a identificação de proteínas clinicamente relevantes. Dessa forma, esse trabalho objetiva preencher essa lacuna, conforme os objetivos descritos na seção 1.2.

## 1.2 Objetivos

### 1.2.1 Objetivo Geral

Integrar dados proteômicos de tecidos humanos de vários estudos e realizar uma análise exploratória de dados por meio da descoberta de proteínas abundantes em tecidos tumorais, a fim de potencializar a identificação de genes a nível proteico como alvos moleculares para o câncer.

### 1.2.2 Objetivos Específicos

- Processar e integrar dados proteômicos públicos de diferentes estudos;
- Analisar a nível proteômico a expressão de antígenos de câncer/ testículo (CTAs);
- Identificar as proteínas abundantes em tecidos tumorais em relação a tecidos saudáveis.

## 1.3 Justificativa

Os dados públicos de sequências de proteínas têm crescido substancialmente nos últimos anos, permitindo a integração tais dados por meio de técnicas de meta-análise. A meta-análise proteômica tem como objetivo integrar dados de vários estudos proteômicos, a fim de potencializar novas descobertas, tais como a identificação de proteínas de interesse clínico. O estudo do proteoma possibilita identificar as proteínas que estão sendo expressas em um determinado momento, quantificar e observar suas modificações pós-traducionais (EMIDIO et al., 2016).

Notavelmente, o estudo do genoma e do transcriptoma não são suficientes para elucidar os mecanismos moleculares de uma célula, pois o proteoma é central para entender a função celular. Não existe uma relação linear entre os genes e o proteoma de uma célula, bem como a existência de um ORF em um gene não o caracteriza como funcional. Logo, a proteômica é complementar à genômica porque ela explora os produtos gênicos. Além disso, como a quantidade de mRNA em uma célula não é necessariamente proporcional ao nível de expressão das proteínas, indicar de forma direta a expressão das proteínas pode ser uma estratégia fundamental (PANDEY; MANN, 2000).

Como o proteoma traduz com mais precisão o estado dinâmico de uma célula, tecido ou organismo, espera-se que a proteômica seja promissora para a produção de marcadores de

doenças, para o diagnóstico e o monitoramento da terapia. Com o acelerado crescimento das tecnologias proteômicas, surgiram novas oportunidades e desafios no entendimento de doenças. As tecnologias proteômicas associadas com a bioinformática avançada são extremamente utilizadas com a finalidade de identificar assinaturas moleculares de doenças. Inclusive, os avanços significativos das tecnologias proteômicas tornaram o proteoma um componente fundamental para compreender a biologia do câncer, possibilitando aplicar a proteômica na descoberta de alvos moleculares para essa doença (CHO, 2007). Com isso, este trabalho objetiva realizar a integração de dados proteicos de tecidos humanos oriundos de diversas fontes, com a finalidade de identificar proteínas abundantes em tecidos tumorais, potencializando a descoberta de genes como alvos moleculares para o câncer a nível proteico.

## 1.4 Organização da monografia

Este trabalho está organizado da seguinte forma: o Capítulo 2 descreve a fundamentação teórica. O Capítulo 3 apresenta os trabalhos relacionados. A metodologia e os resultados desse trabalho são apresentados no Capítulo 4, intitulado como processamento e análise de dados proteômicos. E as conclusões e trabalhos futuros são apresentadas no Capítulo 5.

# 2 FUNDAMENTAÇÃO TEÓRICA

## 2.1 Coleta de dados

Uma das primeiras atividades ao analisar dados é coletar e preparar os dados em um formato ideal para a análise das amostras. Os dados podem ser coletados de várias fontes, sendo lidos diretamente de um arquivo ou podem ser obtidos por raspagem da web, notícias, redes sociais, etc. Os dados apresentam formatos, tipos e tamanhos diferentes. O método mais apropriado para a análise depende do formato em que os dados estão: texto simples, colunas fixas, CSV, XML, SQL, etc (IGUAL; SEGUí, 2017). Também é importante compreender o tipo de dados que são coletados, sejam eles estruturados, semi estruturados e não estruturados. Os dados estruturados possuem uma estrutura rígida, previamente planejada e cada campo de dados tem um formato bem definido. Os dados semi estruturados possuem estrutura flexível e cada campo de dados tem uma estrutura, mas não existe uma imposição de formato. Já os dados não estruturados são sem estrutura pré-definida, representando dados produzidos a partir de atividades humanas. Assim, nessa fase é fundamental identificar as fontes dos dados e compreender os tipos e formatos de cada dado.

## 2.2 Preparação de dados

Após a fase de coleta de dados, inicia-se a fase de preparação de dados. A sua primeira etapa é a limpeza de dados, que é responsável por remover valores discrepantes, substituir valores ausentes, suavizar dados ruidosos e corrigir dados inconsistentes. Em muitos conjuntos de dados do mundo real, valores errados podem ser registrados por vários motivos, incluindo erros de medição, julgamentos subjetivos e mau funcionamento ou uso indevido de equipamentos de gravação automática. Além disso, os valores de dados podem não ser registrados para todos os atributos, podendo acontecer, meramente porque existem alguns atributos que não são aplicáveis a alguns casos (por exemplo, a expressão de certos genes pode ser significativa apenas para pacientes saudáveis ou para pacientes doentes). Valores ausentes podem acontecer por várias causas, como: informações que não puderam ser obtidas, mau funcionamento do equipamento utilizado para registrar dados; adição de novos campos após alguns dados já terem sido coletados. Uma solução seria descartar esses valores ausentes, evitando a introdução de quaisquer erros de dados, sua desvantagem é que o descarte dos dados pode prejudicar a confiabilidade dos resultados. Então não é recomendado em geral, mas pode ser que seja uma estratégia eficaz quando a proporção de valores ausentes for pequena (BRAMER, 2007).

Uma outra estratégia é substituir um valor ausente por uma estimativa de seu valor verdadeiro, seja preenchendo o valor ausente manualmente, usando um valor padrão, usando tendência central (média, mediana, moda) ou usando um valor mais provável. No entanto, essas estratégias podem, obviamente, introduzir ruído nos dados, mas se a proporção de valores

ausentes para uma variável for pequena, isso provavelmente não terá mais do que um pequeno efeito nos resultados derivados dos dados (BRAMER, 2007). Como os dados podem ter muitas inconsistências, eles devem ser processados e explorados, pois quanto mais limpos estiverem os dados, melhores serão as previsões.

Em seguida, é realizada a etapa de integração dos dados, que possibilita integrar, de forma adequada, dados coletados de várias fontes de dados. Em geral, a integração dos dados pode gerar dados redundantes e inconsistentes. Valores inconsistentes são encontrados quando, valores indicados pelos mesmos atributos possuem discrepâncias em relação a formatação do tipo de dado ou não possuem o mesmo domínio. Para tratar tais valores pode-se remover o valor inconsistente, corrigir manualmente e realizar a análise do esquema das fontes de dados e dos conjuntos de dados para construir procedimentos de correção automática. Já valores redundantes podem ser causados por três motivos: uso de nomenclaturas diferentes para atributos equivalentes; inserção de dados repetidos no conjunto de dados; armazenamento de dados derivados de outros atributos. Para solucionar essas problemáticas é indicado realizar a redução dos dados, seja de forma horizontal, eliminando exemplares, quanto na vertical, eliminando atributos. Além disso, pode-se também selecionar o que é útil para a análise que está sendo realizada, no lugar de procurar o que deve ser eliminado, sempre mantendo a mesma capacidade analítica do conjunto de dados original (SILVA et al., 2017).

Após a integração, os dados serão submetidos à etapa de redução, que busca uma representação compacta do conjunto de dados que seja menor em volume, mas mantenha a integridade do dado original. Como o conjunto de dados pode ser muito grande, pode ser desejável reduzir seu tamanho tanto em termos do número de linhas quanto do número de dimensões. Para reduzir o número de colunas nos dados, por exemplo, pode ser usada uma técnica da fase de transformação dos dados, a seleção de atributo ou recurso, que consiste na seleção de atributos em seus dados que são mais adequados para análise (AGGARWAL, 2015).

A etapa de transformação dos dados busca transformar os dados em um formato apropriado para sua modelagem. Em muitos cenários, os diferentes recursos apresentam diferentes escalas de referência e, como consequência, podem não ser comparáveis entre si. Para resolver esse problema, é comum o uso da técnica de normalização dos dados (AGGARWAL, 2015). Os procedimentos mais comumente utilizados são: a Normalização Min-Max, que dimensiona os dados não normalizados para limites inferiores e superiores predefinidos linearmente, onde esses dados são geralmente redimensionados dentro do intervalo de 0 a 1 ou -1 a 1; e a Normalização *z-score*, onde as medidas de média e desvio padrão são usadas para redimensionar os dados de modo que as características resultantes tenham média zero e uma unidade de variância (SINGH; SINGH, 2020). Outra técnica que permite a transformação dos dados é a conversão, que pode ser de duas formas: conversão de valores numéricos para categóricos; e conversão de valores categóricos para numéricos. A primeira conversão, conhecida como discretização, substitui valores brutos de atributos numéricos por atributos categóricos. Os valores para atributos categóricos nominais ainda podem ser generalizados para conceitos de ordem superior por meio

da técnica de transformação geração do conceito de hierarquia para dados nominais. Enquanto que a segunda conversão, nomeada como codificação, representa atributos categóricos usando números. Por exemplo, para representar tipos de pacientes (doente e saudável), pode-se utilizar um atributo valorado como 0 ou 1, de modo que, normalmente, cada valor seja considerado um atributo descritivo do conjunto de dados (SILVA et al., 2017).

Ainda existem outros procedimentos referentes ao pré-processamento de dados, no entanto, neste capítulo, procurou-se abordar apenas os que serão implementados no trabalho. Concluída a fase de preparação, os dados estão prontos para serem explorados.

## 2.3 Análise exploratória de dados

A análise exploratória de dados proporciona um entendimento dos dados, possibilitando uma visão de como os dados estão distribuídos e qual a maneira que se apresentam. Um dos principais objetivos da análise exploratória é visualizar e resumir a distribuição dos dados, permitindo, assim, fazer suposições provisórias sobre sua distribuição e identificar novos padrões (IGUAL; SEGUí, 2017). Para isso, são empregadas técnicas estatísticas descritivas e gráficas para estudar o conjunto de dados, detectando outliers e anomalias, além de testar suposições do modelo. Durante essa fase, é importante o uso de perguntas, a fim de orientar a investigação dos dados, auxiliando na decisão de quais gráficos, modelos ou transformações devem ser criados. É fundamental compreender que essa fase não influencia somente no resultado da análise, mas durante toda ela, pois, a representação dos dados utilizando estatística descritiva e diferentes tipos de gráficos ajudam a caracterizar dados que estão escondidos em várias tabelas e até entre milhares de linhas de uma tabela, possibilitando identificar novos padrões e características dos dados (ESCOBAR; NETO; TIEZZI, 2020).

# 3 TRABALHOS RELACIONADOS

A capacidade de caracterizar tumores por meio de suas características moleculares é fundamental para o desenvolvimento da terapia personalizada do câncer, pois algumas características a nível molecular são inerentes a tecidos e tipos de tumores. Rosenberg et al. (2010) realizaram uma meta-análise multivariada de dados proteômicos de tumores de próstata e cólon, com o objetivo de encontrar padrões comuns nesses tipos de tumores. A análise realizada identificou 14 proteínas com o mesmo perfil de expressão entre amostras normais e cancerígenas para os dois tipos de tumores, que não foram identificadas ao analisar os conjuntos de dados separadamente.

Assim, a integração de vários conjuntos de dados de espectrometria de massa pode oferecer informações valiosas que uma única análise de conjunto de dados não pode fornecer. Huang et al. (2013), por sua vez, desenvolveram um software de meta-análise, o MetaMSD (do inglês, *Meta Analysis for Mass Spectrometry Data*), projetado para analisar vários conjuntos de dados proteômicos, que podem ser gerados por diferentes técnicas de rotulagem e / ou diferentes tipos de instrumentos de espectrometria de massa. Esse software detecta de forma significativa mais proteínas diferenciais do que a análise baseada no melhor e único experimento.

A extensa quantidade de dados de espectrometria de massa sobre o mesmo tipo de estudo, permitiu que Oliveira et al. (2020) integrassem dados proteômicos do câncer de pâncreas para descobrir proteínas relevantes para o diagnóstico e prognóstico do adenocarcinoma ductal pancreático (PDAC). Essa é uma doença intensamente agressiva, que possui prognóstico desfavorável e não existem biomarcadores para sua detecção precoce. Eles realizaram meta-análise integrada de dados públicos do proteoma e do secretoma PDAC com a finalidade de identificar potenciais biomarcadores da doença e, como resultado, foram identificados um conjunto de 39 proteínas secretadas com potencial para biomarcadores.

Genes com padrão de expressão restrito em tecidos normais e altamente expressos em tecidos cancerígenos são excelentes candidatos a biomarcadores e alvos terapêuticos, entre esses genes, os genes de câncer/testículo (CT) são os mais promissores. Quando os produtos proteicos dos genes de CT provocam uma resposta imune, eles são chamados de antígenos de câncer/testículo (CTAs). Wang et al. (2016), por exemplo, integraram dados transcriptômicos de vários bancos de dados e identificaram 876 CTAs em 19 tipos de câncer. Por sua vez, Silva et al. (2017) identificaram 745 CTAs, sendo 201 novos genes de CT em comparação com o conjunto de genes identificados anteriormente por Wang et al. (2016). Com a finalidade de explorar se os CTAs possuíam um padrão de oncogene ou tumor supressor, um exame mais rigoroso foi aplicado por Silva et al. (2017) e, como resultado, eles identificaram 313 CTAs como possíveis candidatos à imunoterapia contra o câncer. Partindo desses 313 CTAs, este trabalho objetiva realizar uma integração de dados proteicos oriundos de vários estudos e descobrir, se a nível de proteína, tais CTAs também podem ser identificados.

# 4 PROCESSAMENTO E ANÁLISE DE DADOS PROTEÔMICOS

## 4.1 Obtenção dos dados proteômicos

Dados brutos gerados por espectrometria de massa de vários experimentos (GEIGER et al., 2012; GHOLAMI et al., 2013; DEEB et al., 2015; LAWRENCE et al., 2015; COSCIA et al., 2016; TYANOVA et al., 2016; BEKKER-JENSEN et al., 2017; DOLL et al., 2017; RIECKMANN et al., 2017b; HAREL et al., 2019; SINHA et al., 2019; WANG et al., 2019) disponibilizados publicamente foram coletados e submetidos à busca em banco de dados a partir do software de busca MaxQuant versão 1.5.2.8 (COX; MANN, 2008; COX et al., 2011), utilizando o banco de dados de sequências proteicas não-redundantes do Uniprot (CONSORTIUM, 2015) para humanos pelo pesquisador Gustavo Antonio De Souza. Um dos resultados de identificação de proteínas gerados pelo software Maxquant, a tabela grupo de proteínas, foi utilizada como fonte para processamento e análise neste trabalho. Essa tabela contém informações sobre as proteínas identificadas nos arquivos brutos, onde cada linha contém o grupo de proteínas que podem ser reconstruídas a partir de um conjunto de peptídeos. As proteínas que têm peptídeos compartilhados são agrupadas em um grupo de proteínas e quantificadas em conjunto, por exemplo, se todos os peptídeos detectados da proteína A também pertencem à proteína B, A e B formam um grupo de proteínas. Essa tabela pode conter centenas de colunas referentes a identificação das proteínas, como: o grupo de proteína(s) identificada(s), o nome do(s) gene(s) associados à(s) proteínas contidas no grupo, o número total de peptídeos únicos associados ao grupo de proteína(s), o número de proteínas contidas no grupo, as intensidades associadas ao grupo de proteínas, etc. A intensidade representa a soma de todas as intensidades de peptídeos individuais pertencentes a um determinado grupo de proteínas. As tabelas, grupo de proteínas, processadas nesse trabalho, possuíam centenas de atributos sobre a identificação das proteínas, mas só foram selecionados os atributos de interesse para a análise. A Tabela 1 representa uma seleção da tabela grupo de proteínas, contendo apenas atributos de interesse.

Tabela 1 – Seleção da tabela grupo de proteínas, contendo apenas atributos de interesse para a análise. Cada linha contém o grupo de proteína, o nome do gene, a contagem de peptídeos e as intensidades dos grupos de proteínas nas amostras biológicas (A549, Hek293 e MCF7).

| *Protein ID's* | *Gene names* | *Peptides count (unique)* | *Intensity Sample A549* | *Intensity Sample Hek293* | *Insentisy Sample MCF7* |
|---|---|---|---|---|---|
| O15360 | FANCA | 12 | 6056300 | 17061000 | 14313000 |
| O14640;P54792 | DVL1;DVL1P1 | 6;3 | 1907200 | 22815000 | 5790600 |

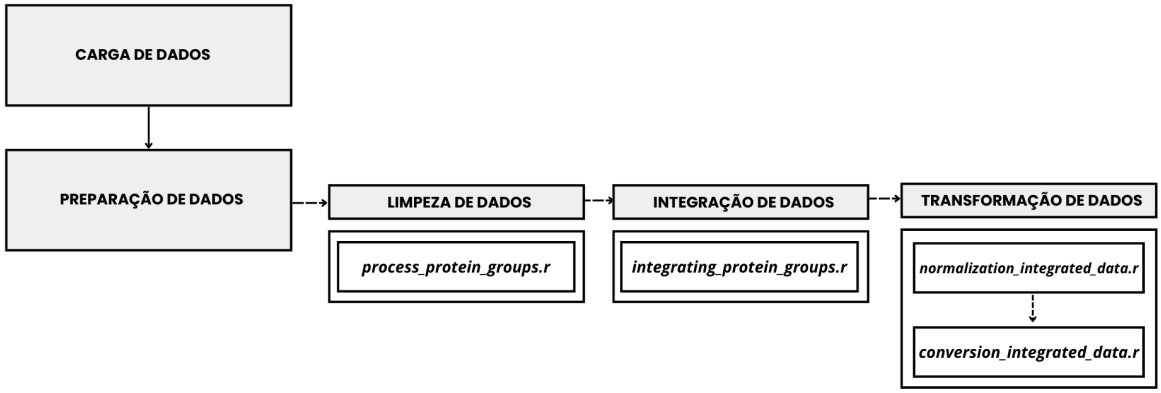
Fonte: Autoria própria.

Nesse trabalho, os estudos focaram-se no processamento e na integração de 15 tabelas grupos de proteínas, somando aproximadamente 10 Tb de dados proteômicos, contendo mais de 500 amostras de tecidos saudáveis, tumores coletados de pacientes e de linhagens celulares imortalizadas usadas como modelo em câncer. Os dados foram lidos de arquivos no formato txt, estruturados, em que cada coluna possuía um formato bem definido. Como os dados coletados não estavam prontos para análise imediata, iniciou-se a próxima fase, que é o pré-processamento.

## 4.2 Fluxo do processamento de dados

A fim de preparar os dados proteômicos para a análise exploratória, discutida na seção 4.3, desenvolveram-se scripts utilizando a linguagem de programação Rscript versão 4.0.5, responsáveis pelos processos de limpeza, integração e transformação de dados, conforme mostrado pela Figura 1.

Figura 1 – Fluxo do processamento dos dados proteômicos. Durante o processamento, os dados passaram pela etapa de limpeza, integração e transformação dos dados. Como resultado do processamento, foi gerada uma base de dados proteômicos, possibilitando o desenvolvimento de análises biológicas.




Fonte: Autoria própria.

### 4.2.1 Limpeza de dados

Com a finalidade de realizar a limpeza de dados foi desenvolvido o script *process_protein_groups.r*. Esse script é responsável por remover as proteínas consideradas como contaminantes e sinalizadas como falsas, além de selecionar os dados de interesse, como: identificador das proteínas, nome dos genes, número de peptídeos associados às proteínas identificadas e intensidades das proteínas nas amostras biológicas. As proteínas que possuem peptídeos compartilhados, têm seus identificadores agrupados e separados por ";". Dessa forma, para cada grupo, com mais de um identificador de proteína, o script seleciona a isoforma proteica e o gene codificador daquela proteína que detectou o maior número de peptídeos. Ainda

foram removidos genes com nome ausente e genes com nome duplicado do conjunto de dados. Uma seleção do resultado do processamento realizado pelo script *process_protein_groups.r* pode ser observada por meio Tabela 2.

Tabela 2 – Seleção do resultado do processamento realizado pelo script *process_protein_groups.r*. Cada linha contém o grupo de proteínas identificadas, o nome dos genes, a contagem de peptídeos e as intensidades dos grupos de proteínas nas amostras biológicas (A549, Hek293 e MCF7).

| *Protein ID's* | *Gene names* | *Peptides count (unique)* | *Intensity Sample A549* | *Intensity Sample Hek293* | *Insentisy Sample MCF7* |
|---|---|---|---|---|---|
| O15360 | FANCA | 12 | 6056300 | 17061000 | 14313000 |
| O14640 | DVL1 | 6 | 1907200 | 22815000 | 5790600 |

Fonte: Autoria própria.

### 4.2.2 Integração de dados

Em seguida, iniciou-se a etapa de integração de dados. Inicialmente foi realizada uma alteração no formato das tabelas grupos de proteínas, elas foram alongadas, aumentando o número de linhas e diminuindo o número de colunas, pois, percebeu-se que o conjunto de proteínas identificadas não era o mesmo em todas as tabelas processadas. Depois, as amostras biológicas procedentes de diferentes estudos, que encontravam-se em diferentes tabelas, foram classificadas de acordo com o tecido de origem e agrupadas, possibilitando amplificar o número de amostras por tecido. Do total de amostras, 140 pertenciam a tecidos normais, enquanto 385 pertenciam a diferentes tipos de câncer. Os tumores analisados foram: câncer de cérebro; câncer de mama, câncer cervical; câncer de cólon; câncer de rim; câncer de pulmão; câncer de fígado; linfoma; melanoma; osteosarcoma; mesotelioma; câncer de ovário e câncer de próstata. O número de amostras para cada tipo de tumor pode ser visto na Tabela 3. Mesmo com o agrupamento das amostras biológicas por tecido de origem, para alguns tecidos cancerígenos a quantidade de amostras ainda é pequena, então, para a análise exploratória, considerou-se apenas os tecidos com o número de amostras igual ou superior a 10. Dessa forma, câncer de cérebro, câncer cervical, câncer de fígado, osteosarcoma e mesotelioma foram desconsiderados. No total, foram analisados 8 tipos de tumores. Em relação às amostras que pertenciam a tecidos normais, todas elas agrupadas em um único grupo.

A alteração no formato do dado e sua integração foi realizada por meio do script *integrating_protein_groups.r*. Também foram adicionadas informações descritivas referentes às amostras, gerando um dataset final (Tabela 4) com as seguintes colunas: identificador e nome do banco de dados, nome do tecido biológico, identificador e nome da amostra biológica, nome do gene e intensidade do gene. Assim, esse dataset final identifica as intensidades de um

Tabela 3 – Quantidade de amostras por tumor.

| *Tumor* | *n* |
|---|---|
| *Brain Tumor* | 7 |
| *Breast Cancer* | 72 |
| *Cervical Cancer* | 4 |
| *Colon Cancer* | 34 |
| *Kidney Cancer* | 10 |
| *Liver Cancer* | 2 |
| *Lung Cancer* | 10 |
| *Lymphoma* | 29 |
| *Melanoma* | 130 |
| *Gliobastoma* | 1 |
| *Osteosarcoma* | 1 |
| *Ovary Cancer* | 43 |
| *Prostate Cancer* | 41 |

Fonte: Autoria própria.

gene em diferentes amostras biológicas referentes a determinados tecidos biológicos que foram extraídos de diferentes experimentos.

### 4.2.3 Transformação de dados

Como os dados processados foram disponibilizados por diferentes estudos, sendo coletados de forma independente, eles não têm a mesma performance, sendo necessário realizar a normalização de tais intensidades. Para não interferir na normalização, todas as intensidades quantificadas como zero foram substituídas por NA (do inglês, *not avaliable*). O valor zero para a intensidade não significa necessariamente que a proteína não foi identificada na amostra biológica. Na verdade, o zero tanto pode indicar que a proteína está ausente na amostra biológica, quanto que ela está com uma expressão tão baixa, que não foi identificada pela máquina de espectrometria. Desse modo, foi desenvolvido o script *normalization_integrated_data.r*, que realizou a normalização *Z-score* sobre as intensidades das proteínas identificadas para cada um dos estudos. Uma seleção do resultado dessa transformação pode ser visualizada por meio Tabela 5.

Com a finalidade de explorar a expressão dos genes em cada tecido biológico do conjunto de dados, calculou-se, para cada gene, a mediana dos valores de intensidades normalizados de todas as amostras do mesmo tecido, assim, o valor obtido representa o valor central das intensidade das amostras de cada gene dentro de cada tecido. Em seguida, buscou-se realizar uma categorização com os valores das medianas. Assim, a intensidade, que é uma variável contínua, foi convertida, inicialmente, em três tipos de categorias (“baixa”, “média”, “alta”), dividindo a distribuição dos dados em três partes, podendo ter partes de tamanhos diferentes, desde que cada parte tivesse seu intervalo de comprimento igual. Ainda foi criada uma quarta categorização

Tabela 4 – Seleção da saída do script *integrating_protein_groups.r*. Cada linha contém o identificador e nome do banco de dados, nome do tecido biológico, identificador e nome da amostra biológica, nome do gene e intensidade do gene.

| *Dataset* | *Dataset name* | *Tissue* | *Sample* | *Sample name* | *Gene names* | *Intensity* |
|---|---|---|---|---|---|---|
| 11CL | 11 linhagens celulares | Lung Cancer | A549 | Adenocarcinoma de pulmão | FANCA | 6056300 |
| 11CL | 11 linhagens celulares | Kidney Cancer | Hek293 | Células embrionárias de rim | FANCA | 17061000 |
| 11CL | 11 linhagens celulares | Breast Cancer | MCF7 | Cancer de mama | FANCA | 14313000 |
| 11CL | 11 linhagens celulares | Lung Cancer | A549 | Adenocarcinoma de pulmão | DVL1 | 1907200 |
| 11CL | 11 linhagens celulares | Kidney Cancer | Hek293 | Células embrionárias de rim | DVL1 | 22815000 |
| 11CL | 11 linhagens celulares | Breast Cancer | MCF7 | Cancer de mama | DVL1 | 5790600 |

Fonte: Autoria própria.

("muito baixa") para os genes com intensidade NA. Tanto o cálculo da mediana, quanto essas categorizações foram implementadas por meio do script *conversion_integrated_data.r*. Por fim, esse script também implementou uma comparação entre a expressão dos genes nas amostras biológicas provenientes de tecidos normais e de tecidos cancerígenas. Esse script filtrou genes cujas intensidades foram categorizadas como "média" ou "alta" em tecidos cancerígenas e "muito baixa" ou "baixa" no tecido normal. A Tabela 6 ilustra uma seleção da saída desse script.

Os estudos sequentes focaram-se na caracterização da expressão de CTAs a nível proteico. Para esses estudos, CTAs anteriormente preditos por Silva et al. (2017) foram considerados. Na realidade, foi realizada uma intersecção entre as proteínas identificadas pela integração dos dados proteômicos com os CTAs preditos pela transcriptômica. Considerando o resultado dessa intersecção, o passo seguinte foi identificar quais dessas proteínas eram mais abundantes nos tecidos tumorais e pouco abundantes ou ausentes no grupamento saudável.

## 4.3 Análise exploratória dos dados proteômicos

A integração dos dados proteômicos de tecidos humanos identificou 17200 proteínas únicas. A fim de explorar se tais proteínas podem ser identificadas como candidatas a alvos

Tabela 5 – Seleção da saída do script normalization_integrated_data.r Cada linha contém o identificador e nome do banco de dados, nome do tecido biológico, identificador e nome da amostra biológica, nome do gene e *Z-score*.

| *Dataset* | *Dataset name* | *Tissue* | *Sample* | *Sample name* | *Gene names* | *Z-score* |
|---|---|---|---|---|---|---|
| 11CL | 11 linhagens celulares | Lung Cancer | A549 | Adenocarcinoma de pulmão | FANCA | -0.687 |
| 11CL | 11 linhagens celulares | Kidney Cancer | Hek293 | Células embrionárias de rim | FANCA | -0.584 |
| 11CL | 11 linhagens celulares | Breast Cancer | MCF7 | Cancer de mama | FANCA | -0.655 |
| 11CL | 11 linhagens celulares | Lung Cancer | A549 | Adenocarcinoma de pulmão | DVL1 | -1.15 |
| 11CL | 11 linhagens celulares | Kidney Cancer | Hek293 | Células embrionárias de rim | DVL1 | -0.470 |
| 11CL | 11 linhagens celulares | Breast Cancer | MCF7 | Cancer de mama | DVL1 | -1.02 |

Fonte: Autoria própria.

Tabela 6 – Seleção da saída do script conversion_integrated_data.r. Cada linha contém o nome do gene, o tecido biológico, a mediana do z-score e a classe.

| *Gene names* | *Tissue* | *Median Z-score* | *Class* |
|---|---|---|---|
| DVL1 | Breast Cancer | -1.05 | medium |
| DVL1 | Colon Cancer | -1.15 | low |
| DVL1 | Kidney Cancer | -0.714 | medium |
| DVL1 | Lung Cancer | -1.09 | low |
| DVL1 | Lymphoma | -1.09 | low |
| DVL1 | Healthy Tissue | -1.08 | low |

Fonte: Autoria própria.

moleculares para o câncer, investigou-se sua abundância nos tecidos biológicos, seja ele saudável ou cancerígeno. Para isso, essas proteínas foram comparadas com os 313 CTAs anteriormente preditos por Silva et al. (2017), com o intuito de descobrir se tais CTAs também eram identificados a nível proteico. Como resultado, 222 CTAs foram identificados.

Com a finalidade de investigar a abundância das proteínas no grupamento saudável e nos tecidos tumorais, um heatmap foi criado. Como citado anteriormente, as amostras

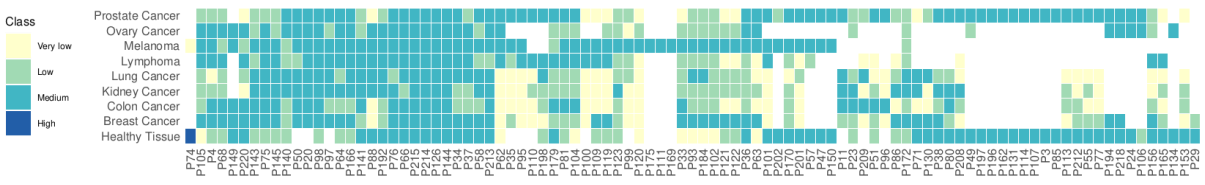
cancerígenas foram agrupadas de acordo com os diferentes tipos de câncer, enquanto que todas as amostras saudáveis foram agrupadas em um único grupo. Essencialmente, um heatmap necessita de 3 variáveis: uma variável resposta e duas outras variáveis para compor os eixos x e y. Não havendo restrição quanto ao tipo de variável, podendo ser quantitativa ou qualitativa. O eixo x, do heatmap criado, indica as 222 proteínas preditas como antígenos CT (SILVA et al., 2017), seus identificadores reais foram suprimidos, a fim de preservar a descoberta biológica, sendo nomeadas como P1, P2, P3 e assim por diante até P222. Já o eixo y demonstra os grupamentos contendo todos os tecidos tumorais e o grupamento saudável. A variável de resposta representa a categoria ("muito baixa", "baixa", "média" ou "alta") da intensidade da proteína dentro de cada tecido. O intervalo de valores e quantidade de amostras de cada categoria pode ser visualizado por meio da Tabela 7. A Figura 2 contém a extremidade esquerda do heatmap criado.

Tabela 7 – Intervalo de valores e quantidades de amostras biológicas por categoria (baixa, média e alta).

| Categorias | Intevalo de valores | Tamanho |
|---|---|---|
| "Baixa" | [-3.5,-1.06) | 24332 |
| "Média" | [-1.06,1.38) | 81507 |
| "Alta" | [1.38,3.82) | 5014 |

Fonte: Autoria própria.

Figura 2 – Heatmap com CTAs preditos em tecidos humanos a nível proteico exibidos por uma variável categórica - extremidade esquerda. O eixo x contém os genes detectados nos tecidos biológicos, o eixo y contém o grupamento saudável e os demais tecidos cancerígenas e a variável de resposta, representada pelas cores no heatmap identificam a categoria ("muito", "baixa", "baixa", "média", "alta"), reflete a intensidade da proteína dentro de cada tecido. Os pixels em branco indicam ausência da proteína no respectivo tecido.



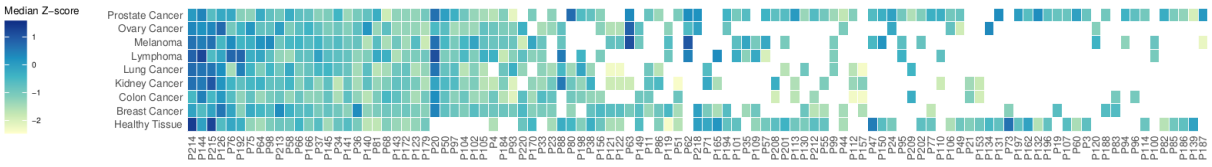

Fonte: Autoria própria.

Ao analisar a Figura 2, observou-se quais proteínas eram categorizadas como "média" ou "alta" em tecidos cancerígenas e "muito baixa" ou "baixa" no grupamento saudável de acordo com a categorização realizada. No total, 40 proteínas satisfizeram essa condição. Também notou-se que muitas proteínas foram categorizadas com intensidade "muito baixa" em vários tecidos. Essa categoria tanto pode indicar que a proteína está ausente em alguns dos tecidos

biológicos, quanto que ela está com uma expressão tão baixa, que o espectrômetro de massa não foi capaz de detectar. Mas, quando de fato, uma proteína estava ausente em alguns dos tecidos biológicos, era atribuído o pixel em branco a ela. Esses valores ausentes não foram descartados, pois, por mais que um gene não tenha sido identificado em um tecido tumoral, o mesmo pode ter sido detectado com intensidade significativa em outros, sugerindo que ele possa ser um possível candidato a alvo molecular para o câncer, desde que tenha apresentado baixa intensidade no grupamento saudável.

A fim de investigar a variação das intensidades das proteínas dentro dos tecidos biológicos, criamos um heatmap com a variável de resposta sendo a mediana do *z-score*, calculada para cada proteína em todas as amostras do respectivo tecido, no entanto, para isso, foi necessário remover as proteínas quantificadas com valor de intensidade igual a 0, que durante o processamento de dados tiveram suas intensidades substituídas por NA, totalizando 211 das 222 proteínas preditas como CTAs. A Figura 3 contém a extremidade esquerda do heatmap criado. Mais uma vez, os nomes das proteínas foram suprimidas, para preservar a descoberta biológica.

Figura 3 – Heatmap com CTAs preditos em tecidos humanos a nível proteico exibidos por uma variável contínua da extremidade esquerda. O eixo x contém os genes detectados nos tecidos biológicos, o eixo y contém o grupamento saudável e os demais tecidos cancerígenas e a variável de resposta é representada pela mediana do *z-score*. Os pixels em branco indicam ausência da proteína no respectivo tecido.

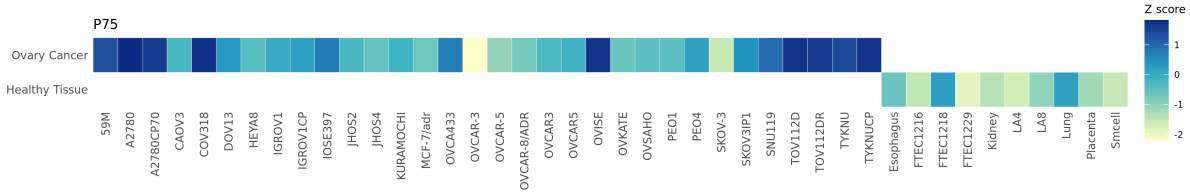



Fonte: Autoria própria.

Pode ser observado por meio da Figura 2, que o gene identificado como P75 teve sua intensidade categorizada como "média" no tecido câncer de ovário e "baixa" no grupamento saudável. Esse resultado pode ser confirmado na Figura 3, onde observa-se uma diferença entre a mediana do *z-score* do tecido câncer de ovário em relação ao grupamento saudável. Com o intuito de aprofundar essa análise, investigou-se o valor das intensidades de cada uma das amostras do tecido de câncer de ovário e das amostras saudáveis para o gene P75, gerando outro heatmap (Figura 4).

É possível notar que o *z-score* é bem mais expressivo para as amostras do tecido câncer de ovário do que do grupamento saudável, sugerindo que a proteína P75 possa ser um possível candidato a alvo molecular para o câncer.

Figura 4 – Heatmap do gene P75 em amostras biológicas do tecido câncer de ovário e do grupamento saudável. O eixo x contém todas as amostras analisadas para o gene, o eixo y contém os grupamentos câncer de ovário e saudável e a variável de resposta é representada pelo *z-score*. Os pixels em branco indicam ausência da proteína na respectiva amostra.




Fonte: Autoria própria.

# 5 CONCLUSÕES E TRABALHOS FUTUROS

A integração de dados proteômicos identificou 17200 proteínas únicas. Para isso, dados proteômicos públicos de diferentes estudos foram integrados, com o objetivo de identificar essas proteínas em tecidos tumorais, potencializando a descoberta de genes como alvos moleculares para o câncer a nível proteico. Amostras biológicas foram agrupadas de acordo com o tipo de tecido tumoral, a fim de amplificar o número de amostras por tumor. O agrupamento revelou 385 amostras tumorais obtidas de 8 diferentes tipos de câncer, foram considerados: câncer de mama (n = 72), câncer de cólon (n = 34); câncer de pulmão (n = 10); câncer de rim (n = 10); linfoma (n = 29); melanoma (n = 130); câncer de ovário (n = 43) e câncer de próstata (n = 41), onde o "n" representa a quantidade de amostras por tecido. As amostras do tecido saudável também foram agrupadas em um único grupo com 140 amostras.

Para potencializar a descoberta de proteínas como alvos moleculares para o câncer, as análises subsequentes focaram-se em explorar a expressão de CTAs a nível de proteína. Dessa forma, antígenos CT anteriormente preditos por Silva et al. (2017) foram considerados, com a finalidade de identificá-los a nível proteico. No total, foram identificados 222 CTAs, dos quais 40 registraram intensidades categorizadas como "média" em tecidos cancerígenas e "muito baixa" ou "baixa" no grupamento saudável. Em conclusão, a integração de dados proteômicos realizada neste trabalho apresenta potencial para permitir futuros avanços na caracterização de proteomas tumorais e consequentemente, na identificação de proteínas como alvos moleculares para o câncer.

Como trabalhos futuros, pretende-se investigar técnicas para lidar com valores faltantes existentes no conjunto de dados proteicos. É comum a presença desses valores ausentes para a intensidade de uma proteína em uma amostra biológica, muitas vezes, decorrente de medições de baixa abundância. Com isso, deseja-se avaliar métodos de imputação, a fim de estimar valores que possam substituir os valores faltantes, sem causar ruídos. Além disso, deseja-se realizar uma validação para fortalecer as descobertas dos 40 CTAs a nível proteico, investigando a sua significância estatística. Além de realizar outras investigações, como por exemplo analisar quais fatores de transcrição (proteínas que regulam a transcrição de genes) estão enriquecidos nos tecidos tumorais.

## Referências


AEBERSOLD, R. et al. Perspective: A program to improve protein biomarker discovery for cancer †. **ACS Publications**, v. 4, p. 1104–1109, 7 2005. Disponível em: <http://pubs.acs.org>. Citado na página 12.

AGGARWAL, C. C. Data mining. Springer International Publishing, 2015. Disponível em: <http://link.springer.com/10.1007/978-3-319-14142-8>. Citado na página 16.

BARBOSA, E. B. et al. Proteômica: metodologias e aplicações no estudo de doenças humanas. **Revista da Associação Médica Brasileira**, Associação Médica Brasileira, v. 58, p. 366–375, 2012. ISSN 0104-4230. Disponível em: <http://www.scielo.br/j/ramb/a/rbzzwDkg4gWNqcPSLNMK6wD/abstract/?lang=pt>. Citado na página 11.

BEKKER-JENSEN, D. B. et al. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. **Cell systems**, Elsevier, v. 4, n. 6, p. 587–599, 2017. Citado na página 19.

BERGER, B.; PENG, J.; SINGH, M. Computational solutions for omics data. **Nature reviews genetics**, Nature Publishing Group, v. 14, n. 5, p. 333–346, 2013. Citado na página 10.

BONALDI, T. et al. Combined use of rnai and quantitative proteomics to study gene function in drosophila. **Molecular cell**, Elsevier, v. 31, n. 5, p. 762–772, 2008. Citado na página 12.

BRAMER, M. A. M. A. Principles of data mining. Springer, p. 343, 2007. Citado 2 vezes nas páginas 15 e 16.

CABALLERO, O. L.; CHEN, Y. T. Cancer/testis (ct) antigens: Potential targets for immunotherapy. **Cancer Science**, v. 100, p. 2014–2021, 11 2009. ISSN 13479032. Citado na página 11.

CHO, W. C. Proteomics technologies and challenges. **Genomics, Proteomics and Bioinformatics**, v. 5, p. 77–85, 2007. ISSN 16720229. Citado na página 14.

CONSORTIUM, T. U. Uniprot: a hub for protein information. **Nucleic Acids Research**, v. 43, 2015. Disponível em: <http://www.uniprot.org/proteomes/UP000000803.> Citado na página 19.

COSCIA, F. et al. Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status. **Nature communications**, Nature Publishing Group, v. 7, n. 1, p. 1–14, 2016. Citado 2 vezes nas páginas 12 e 19.

COX, J.; MANN, M. Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. **Nature Biotechnology**, Nature Publishing Group, v. 26, p. 1367–1372, 12 2008. ISSN 10870156. Disponível em: <https://www.nature.com/articles/nbt.1511>. Citado na página 19.

COX, J. et al. Andromeda: A peptide search engine integrated into the maxquant environment. **Journal of Proteome Research**, American Chemical Society, v. 10, p. 1794–1805, 4 2011. ISSN 15353893. Disponível em: <https://pubs.acs.org/doi/full/10.1021/pr101065j>. Citado na página 19.

DEEB, S. J. et al. Machine learning-based classification of diffuse large b-cell lymphoma patients by their protein expression profiles. **Molecular & Cellular Proteomics**, ASBMB, v. 14, n. 11, p. 2947–2960, 2015. Citado na página 19.

DOLL, S. et al. Region and cell-type resolved quantitative proteomic map of the human heart. **Nature communications**, Nature Publishing Group, v. 8, n. 1, p. 1–13, 2017. Citado na página 19.

EDWARDS, N. J. et al. The cptac data portal: A resource for cancer proteomics research. **Journal of Proteome Research**, American Chemical Society, v. 14, p. 2707–2713, 6 2015. ISSN 15353907. Disponível em: <https://pubs.acs.org/doi/abs/10.1021/pr501254j>. Citado na página 11.

EMIDIO, N. B. et al. Proteômica: uma introdução aos métodos e aplicações. **HU Revista**, v. 41, p. 101–111, 4 2016. ISSN 0103-3123. Disponível em: <https://periodicos.ufjf.br/index.php/hurevista/article/view/2482>. Citado na página 13.

ESCOBAR, D. B.; NETO, G. H.; TIEZZI, D. G. Pré-processamento de dados clínicos do consórcio tcga. In: . [s.n.], 2020. Disponível em: <http://www.fatecrp.edu.br/WorkTec/edicoes/2020-2/trabalhos/II-Worktec-Danielle_Escobar.pdf>. Citado na página 17.

GEIGER, T. et al. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. **Molecular and Cellular Proteomics**, American Society for Biochemistry and Molecular Biology Inc., v. 11, 2012. ISSN 15359484. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/22278370/https://pubmed.ncbi.nlm.nih.gov/22278370/?from_single_result=Comparative+proteomic+analysis+of+eleven+common+cell+lines+reveals+ubiquitous+but+varying+expression+of+most+proteins>. Citado 2 vezes nas páginas 12 e 19.

GHOLAMI, A. M. et al. Global proteome analysis of the nci-60 cell line panel. **Cell reports**, Elsevier, v. 4, n. 3, p. 609–620, 2013. Citado na página 19.

GLASS, G. V. Primary, secondary, and meta-analysis of research. **Educational Researcher**, American Educational Research Association (AERA), v. 5, p. 3–8, 11 1976. ISSN 0013-189X. Citado na página 11.

GOVEIA, J. et al. Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities. **EMBO Molecular Medicine**, EMBO, v. 8, p. 1134–1142, 10 2016. ISSN 1757-4676. Citado na página 11.

GRAUMANN, J. et al. Stable isotope labeling by amino acids in cell culture (silac) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. In: . [S.l.]: Mol Cell Proteomics, 2008. v. 7, p. 672–683. Citado na página 12.

GYGI, S. P. et al. Correlation between protein and mrna abundance in yeast. **MOLECULAR AND CELLULAR BIOLOGY**, v. 19, p. 1720–1730, 1999. Disponível em: <http://mcb.asm.org/>. Citado na página 11.

HAREL, M. et al. Proteomics of melanoma response to immunotherapy reveals mitochondrial dependence. **Cell**, Elsevier, v. 179, n. 1, p. 236–250, 2019. Citado na página 19.

HU, Z.; ZHOU, Y.; TONG, T. Meta-analyzing multiple omics data with robust variable selection. **Frontiers in Genetics**, Frontiers Media S.A., v. 12, p. 1029, 7 2021. ISSN 16648021. Citado na página 11.

HUANG, Y. H. et al. Identification and enhancement of hla-a2.1-restricted ctl epitopes in a new human cancer antigen-pote. **PLOS ONE**, Public Library of Science, v. 8, p. e64365, 6 2013. ISSN 1932-6203. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064365>. Citado na página 18.

HUANG, Z. et al. Proteomic profiling of human plasma for cancer biomarker discovery. **Proteomics**, Wiley-VCH Verlag, v. 17, 3 2017. ISSN 16159861. Disponível em: <https://pubmed-ncbi-nlm-nih.ez18.periodicos.capes.gov.br/27550791/>. Citado na página 11.

IGUAL, L.; SEGUí, S. Introduction to data science. Springer, Cham, p. 1–4, 2017. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-50017-1_1>. Citado 2 vezes nas páginas 15 e 17.

JIN, S. et al. Cancer/testis antigens (ctas) expression in resected lung cancer. **OncoTargets and therapy**, Dove Press, v. 11, p. 4491, 2018. ISSN 11786930. Disponível em: </pmc/articles/PMC6078192//pmc/articles/PMC6078192/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6078192/>. Citado na página 11.

KELSTRUP, C. D. et al. Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole orbitrap mass spectrometer. **Journal of Proteome Research**, J Proteome Res, v. 11, p. 3487–3497, 6 2012. ISSN 15353893. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/22537090/https://pubmed.ncbi.nlm.nih.gov/22537090/?from_single_result=Optimized+fast+and+sensitive+acquisition+methods+for+shotgun+proteomics+on+a+quadrupole+orbitrap+mass+spectrometer>. Citado na página 12.

KRISHNADAS, D. K. et al. A phase i trial combining decitabine/dendritic cell vaccine targeting mage-a1, mage-a3 and ny-eso-1 for children with relapsed or therapy-refractory neuroblastoma and sarcoma. **Cancer Immunology, Immunotherapy**, Springer, v. 64, n. 10, p. 1251–1260, 2015. Citado na página 12.

KRISTENSEN, V. N. et al. Principles and methods of integrative genomic analyses in cancer. **Nature Reviews Cancer**, Nature Publishing Group, v. 14, n. 5, p. 299–313, 2014. Citado na página 10.

KULKARNI, P.; SCIENCES, V. U. I. journal of molecular; 2017 undefined. Cancer/testis antigens:"smart" biomarkers for diagnosis and prognosis of prostate and other cancers. **mdpi.com**. Disponível em: <https://www.mdpi.com/189160>. Citado na página 12.

LAWRENCE, R. T. et al. The proteomic landscape of triple-negative breast cancer. **Cell reports**, Elsevier, v. 11, n. 4, p. 630–644, 2015. Citado na página 19.

LEWIN, B. **Genes Viii**. [S.l.: s.n.], 2004. Citado na página 10.

LI, Y. et al. Roles of cancer/testis antigens (ctas) in breast cancer. **Cancer Letters**, Elsevier Ireland Ltd, v. 399, p. 64–73, 7 2017. ISSN 18727980. Citado na página 11.

MANN, M.; KELLEHER, N. L. Precision proteomics: the case for high resolution and high mass accuracy. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 105, n. 47, p. 18132–18138, 2008. Citado na página 12.

NGUYEN, T. et al. Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. **Scientific Reports**, Nature Publishing Group, v. 6, n. 1, p. 1–14, 2016. Citado na página 10.

OLIVEIRA, G. de et al. An integrated meta-analysis of secretome and proteome identify potential biomarkers of pancreatic ductal adenocarcinoma. **Cancers**, MDPI AG, v. 12, 3 2020. ISSN 20726694. Disponível em: <https://pubmed-ncbi-nlm-nih.ez18.periodicos.capes.gov.br/32197468/>. Citado na página 18.

PANDEY, A.; MANN, M. Proteomics to study genes and genomes. **Nature**, Nature, v. 405, p. 837–846, 2000. Citado na página 13.

PARMIGIANI, R. B. et al. Characterization of a cancer/testis (ct) antigen gene family capable of eliciting humoral response in cancer patients. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 103, n. 48, p. 18066–18071, 2006. Citado na página 11.

PEREZ-RIVEROL, Y. et al. Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. **Proteomics**, Wiley-VCH Verlag, v. 15, p. 930–950, 3 2015. ISSN 16159861. Citado na página 11.

RIECKMANN, J. C. et al. Social network architecture of human immune cells unveiled by quantitative proteomics. **Nature Immunology**, Nature Publishing Group, v. 18, p. 583–593, 4 2017. ISSN 15292916. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/28263321/https://pubmed.ncbi.nlm.nih.gov/28263321/?from_single_result=Social+network+architecture+of+human+immune+cells+unveiled+by+quantitative+proteomics.> Citado na página 12.

RIECKMANN, J. C. et al. Social network architecture of human immune cells unveiled by quantitative proteomics. **Nature immunology**, Nature Publishing Group, v. 18, n. 5, p. 583–593, 2017. Citado na página 19.

ROSENBERG, L. H. et al. Multivariate meta-analysis of proteomics data from human prostate and colon tumours. **BMC Bioinformatics**, v. 11, 9 2010. ISSN 14712105. Citado na página 18.

SHAFI, A. et al. A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures. **Frontiers in Genetics**, Frontiers Media S.A., v. 10, p. 159, 2019. ISSN 16648021. Citado na página 10.

SILVA, V. L. da et al. Genome-wide identification of cancer/testis genes and their association with prognosis in a pan-cancer analysis. **Oncotarget**, Impact Journals, LLC, v. 8, n. 54, p. 92966, 2017. Citado 8 vezes nas páginas 12, 16, 17, 18, 23, 24, 25 e 28.

SINGH, D.; SINGH, B. Investigating the impact of data normalization on classification performance. **Applied Soft Computing**, Elsevier, v. 97, p. 105524, 12 2020. ISSN 1568-4946. Citado na página 16.

SINHA, A. et al. The proteogenomic landscape of curable prostate cancer. **Cancer Cell**, Elsevier, v. 35, n. 3, p. 414–427, 2019. Citado na página 19.

TYANOVA, S. et al. Proteomic maps of breast cancer subtypes. **Nature communications**, Nature Publishing Group, v. 7, n. 1, p. 1–11, 2016. Citado na página 19.

WANG, C. et al. Systematic identification of genes with a cancer-testis expression pattern in 19 cancer types. **Nature Communications**, Nature Publishing Group, v. 7, 1 2016. ISSN 20411723. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/26813108/>. Citado 2 vezes nas páginas 12 e 18.

WANG, D. et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. **Molecular systems biology**, v. 15, n. 2, p. e8503, 2019. Citado na página 19.

WU, C. et al. A selective review of multi-level omics data integration using variable selection. **High-Throughput 2019, Vol. 8, Page 4**, Multidisciplinary Digital Publishing Institute, v. 8, p. 4, 1 2019. ISSN 2571-5135. Disponível em: <https://www.mdpi.com/2571-5135/8/1/4/htmhttps://www.mdpi.com/2571-5135/8/1/4>. Citado na página 11.