UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
INSTITUTO METRÓPOLE DIGITAL
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

LUCAS MARQUES DA CUNHA

DESENVOLVIMENTO DE ABORDAGEM COMPUTACIONAL PARA ANÁLISE E IDENTIFICAÇÃO DE PEPTÍDEOS POLIMÓRFICOS

NATAL - RN
Novembro-2022

LUCAS MARQUES DA CUNHA

DESENVOLVIMENTO DE ABORDAGEM COMPUTACIONAL PARA ANÁLISE E IDENTIFICAÇÃO DE PEPTÍDEOS POLIMÓRFICOS

Tese de Doutorado apresentada ao programa de Pós-Graduação em Bioinformática da Universidade Federal do Rio Grande do Norte, como requisito para obtenção do título de Doutor em Bioinformática.

Área de concentração: Proteogenômica
Orientador: Dr. Gustavo Antônio de Souza.

NATAL - RN
Novembro-2022

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Central Zila Mamede

Cunha, Lucas Marques da.
Desenvolvimento de abordagem computacional para análise e identificação de peptídeos polimórficos / Lucas Marques da Cunha. - 2022.
89 f.: il.

Tese (doutorado) - Universidade Federal do Rio Grande do Norte, Instituto Metrópole Digital, Programa de Pós-graduação em Bioinformática. Natal, RN, 2022.
Orientador: Prof. Dr. Gustavo Antônio de Souza.

1. Peptídeos variantes - abordagem computacional - Tese. 2. Proteogenômica - Tese. 3. Polimorfismo - Tese. 4. Proteômica - Tese. 5. Banco de dados personalizado - Tese. I. Souza, Gustavo Antônio de. II. Título.

RN/UF/BCZM CDU 577.112.6

Elaborado por Kalline Bezerra da Silva - CRB-15-327

LUCAS MARQUES DA CUNHA

DESENVOLVIMENTO DE ABORDAGEM COMPUTACIONAL PARA ANÁLISE E IDENTIFICAÇÃO DE PEPTÍDEOS POLIMÓRFICOS

Natal, 29 de novembro 2022

BANCA EXAMINADORA

_____

Dr. Gustavo Antônio de Souza, UFRN
(Orientador)

_____

Dr. Daniel Carlos Ferreira Lanza, UFRN
(Membro Interno)

_____

Dra. Adriana Ferreira Uchôa, UFRN
(Membro externo ao programa)

_____

Dr. Paulo Costa Carvalho, FIOCRUZ-ICC
(Membro Externo)

_____

Dr. Fabio Passetti, FIOCRUZ-ICC
(Membro Externo)

NATAL - RN
Novembro-2022

DEDICATÓRIA

A Deus por tudo que me proporciona na vida. À minha mãe, Lourdes e meu pai Paulo, os quais amo muito, pelo exemplo de vida e família. Aos meus irmãos por tudo que me ajudaram até hoje. À minha esposa Talita Cunha, pelo carinho, compreensão e companheirismo. Ao meu filho Valentim, por me fazer compreender a vida em sua forma mais pura.

AGRADECIMENTOS

A Deus, por me proporcionar o dom da vida, sabedoria e discernimento.

Aos meus pais, em especial, minha mãe, que sempre me guiou a fazer a melhor escolha.

Aos meus familiares que sempre me apoiaram diante de qualquer situação, contribuindo e incentivando para que eu nunca desistisse de minha jornada.

A minha esposa, Talita Cunha, amiga e companheira de todos os momentos, sempre me motivando e incentivando a ser vitorioso em minhas batalhas. Ao meu filho Valentim, por simbolizar recomeço e trazer um novo olhar sobre a vida.

Aos meus amigos, Suelda, Dêis, Vitória, motivo de força e otimismo. Nunca me deixaram desanimar, nem me desesperar diante dos problemas surgidos durante o doutorado.

Ao professor Gustavo, que me acompanhou durante esse percurso, por meio de ensinamentos e motivação que tornou possível finalizar meu projeto de doutorado acadêmico. Ao professor Sandro, pela oportunidade concedida e todo o conhecimento compartilhado. A professor Sílvia Battistuzzo por todo apoio tecnológico que permitiu o desenvolvimento da pesquisa. À todos os professores do Centro Multiusuário de Bioinformática (BioME) por todo ensinamento.

Aos todos os meus amigos do BioME, em especial, Carol Miranda, Danilo Martins, Iara Dantas, Diêgo Teixeira, Ricardo Victor, Clóvis Reis, Diêgo Coelho, Thaynã Damasceno, Patrick Terrematte, Marília, Thiago, Josivan, Daniela Coelho e Thaís Ratis companheiros de uma jornada, permitiram compartilhar comigo seus conhecimentos e proporcionando grande diversão durante todo esse tempo. Aos colaboradores, Tayná Fiúza, Vandeclécio, Eduardo Kroll pela dedicação e colaboração nesse trabalho.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financiamento desta pesquisa por meio da rede de Biologia Sistêmica do Câncer (BSC).

DA CUNHA LM. Desenvolvimento de abordagens computacionais para análise e identificação de peptídeos polimórficos [Tese]. Natal: Programa de pós-graduação em Bioinformática, Universidade Federal do Rio Grande do Norte; 2022.

RESUMO

A abordagem proteômica permite estudos em larga escala da expressão proteica em diferentes tecidos e fluidos corporais, tendo como objetivo identificar e quantificar o conteúdo proteico total. No processo de análise proteômica, a identificação de proteínas ainda apresenta lacunas, apesar dos grandes avanços na área. Frequentemente, um espectrômetro de massa é utilizado para gerar valores de massa/carga das amostras. Após esse processo, geralmente utiliza-se um banco de dados de proteínas referência (por exemplo, UniProt) para identificação das proteínas. Porém, utilizar uma base de referência limita as análises de identificação das proteínas, uma vez que não contém as variações que ocorrem no DNA, que podem impactar na sequência de aminoácidos, ocasionando identificação incorreta ou impossibilitando o processo. Nesse contexto, existem diversas bases de dados personalizadas que incorporam tais variações genéticas. Embora apresentem bons resultados, também se limitam devido à ausência de algumas mutações, tornando-se outro problema no processo de identificação. Portanto, essa pesquisa tem como objetivo construir m banco de dados de proteogenômica (dbPepVar) combinando informações de variação genética do dbSNP com sequências de proteínas do RefSeq do NCBI. Conjuntos de dados públicos de espectrometria de massa foram usados para realizar uma análise pan-câncer (Ovário, Colorretal, Mama e Próstata), permitindo a identificação de variações genéticas únicas. No total, 3.726 peptídeos variantes foram identificados em amostras de câncer de ovário, 2.543 em próstata, 2.661 em mama e 2.411 em câncer de cólon-retal. Uma análise de frequência mutacional mostrou genes envolvidos nos processos de progressão tumoral, sensibilidade à quimioterapia e risco de suscetibilidade ao câncer. Curiosamente, em muitas amostras, foram identificados peptídeos C-terminais de proteínas encurtadas originárias de eventos de códon de terminação prematura (PTC). Isso indica que tais proteínas escaparam do decaimento mediado por mutações Nonsense (NMD) e, não surpreendentemente, os genes da maquinaria NMD também estão mutados nas mesmas amostras. Isso sugere que o vestígio do transcrito truncado pode estar associado à ineficiência da maquinaria NMD causada por mutações genéticas. Em perspectiva, o portal web desenvolvido bem como as análises realizadas podem direcionar estudos para identificar novos alvos terapêuticos para diferentes tipos de câncer, podendo-se também utilizar nosso banco de dados para caracterização de variantes em amostras de antecedentes genéticos desconhecidos, como amostras arquivadas. O portal está disponível em: https://bioinfo.imd.ufrn.br/dbPepVar/.

Palavras-chave: Proteômica. Proteogenômica. Polimorfismo. Peptídeos variantes. Banco de dados personalizado. Portal *web*.

DA CUNHA LM. Development of computational approaches for analysis and identification of polymorphic peptides [Thesis]. Natal: Postgraduate Program in Bioinformatics, Federal University of Rio Grande do Norte; 2022

ABSTRACT

The proteomic approach allows large-scale studies of protein expression in different tissues and body fluids, aiming to identify and quantify the total protein content. In the proteomic analysis process, protein identification still presents limitations despite major advances in the area. Frequently, a mass spectrometer is used to generate mass/charge values of the samples. After this process, a reference protein database (eg, UniProt) is usually used to identify proteins. However, using a reference database limits the analysis of the identification of the proteins, since it does not contain the variations in the DNA that can impact the sequence of amino acids, causing incorrect identification or making the process impossible. In this context, there are several custom databases that incorporate such genetic variations. Although they present good results, they are also limited due to the absence of some mutations, becoming another problem in the identification process. Therefore, this research aims to build a proteogenomic database (dbPepVar) by combining genetic variation information from dbSNP with protein sequences from the NCBI RefSeq. Public mass spectrometry datasets were used to perform a pan-cancer analysis (Ovarian, Colorectal, Breast, and Prostate), allowing the identification of unique genetic variations. In total 3,726 variant peptides were identified in ovarian cancer samples, 2,543 in prostate, 2,661 in breast and 2,411 in colon-rectal cancer. A mutational frequency analysis showed genes involved in tumor progression processes, sensitivity to chemotherapy, and risk of susceptibility to cancer. Interestingly, in many samples, C-terminal peptides from shortened proteins originating from premature termination codon (PTC) events were identified. This indicates that such proteins had escaped Nonsense-mediated decay (NMD) and, not surprisingly, NMD machinery genes are also mutated in the same samples. This suggests that the vestige of the truncated transcript may be associated with NMD machinery inefficiency caused by gene mutations. In perspective, the web portal developed as well as the analysis performed may direct studies to identify new therapeutic targets for different cancer, and one can also use our database for characterization of variants in samples of unknown genetic background, such as archived samples. The portal is available in: https://bioinfo.imd.ufrn.br/dbPepVar/

Keywords: Proteomics. Proteogenomics. Polymorphism. Variant peptides. Custom database. Web Portal.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

PTM - *Post-translational modification* ([1]Modificação pós-traducional)

SAGE - *Serial analysis of gene expression* (Análise serial da expressão gênica)

RNA - *Ribonucleic Acid* (Ácido ribonucléico)

DNA - *Deoxyribonucleic acid* (ácido desoxirribonucleico)

LC - *Liquid Chromatography* (Cromatografia Líquida)

MS - *Mass Spectrometry* (Espectrometria de massas)

HPLC - *High Performance Liquid Chromatography* (Cromatografia líquida de alta performance)

ESI - *Electrospray ionisation* (ionização por electrospray)

MALDI - *Matrix-assisted laser desorption/ionisation* (dessorção/ionização a laser assistida por matriz)

TOF - *Time of Flight* (Tempo de voo)

PMF - *Peptide Mass Fingerprint* (assinatura digital da massa peptídica)

PFF - *Peptide Fragmentation Fingerprint* (assinatura digital do fragmento peptídico)

ICAT - *isotope-coded affinity tag* (marcação de afinidade codificada com isótopos)

ITRAQ - *Isobaric tags for relative and absolute quantitation* (Isobaric tags for relative and absolute quantitation)

SILAC - *Stable isotope labelling with amino acids in cell culture* (Stable isotope labelling with amino acids in cell culture)

SNP - *Single nucleotide polymorphism* (Polimorfismo de nucleotídeo único)

SAP - *Single Amino Acid Polymorphism* (Polimorfismo de aminoácido único)

UTR - *Untranslated Region* (Região não traduzida)

CNV - *Copy number variation* (Variação no número de cópias)

NCBI - *National Center for Biotechnology Information* (Centro Nacional de informação biotecnológica)

IPI - *International Protein Index* (Sequência proteica internacional)

ORF - *Open Read Frame* (Fase de leitura aberta)

MAF - *Minor frequency allele* (Menor frequência alélica)

[1] Tradução livre.

SUMÁRIO

1 INTRODUÇÃO

Esta seção apresenta os conceitos relacionados à pesquisa proposta com intuito de conduzir a compreensão da área estudada. Serão descritos tópicos referentes à Proteômica e Proteogenômica, elucidando conceitos básicos, abordagens, técnicas, as bases de dados existentes para análises e efeitos das variações genéticas no proteoma.

1.1 PROTEÔMICA: CONCEITOS, ABORDAGENS E TÉCNICAS

As proteínas desempenham papéis fundamentais para estabilidade, manutenção e funcionamento adequado de um organismo, envolvendo-se em grande parte dos processos celulares, podendo atuar como enzimas, anticorpos, hormônios, componentes estruturais e receptores celulares (POLURI; GULATI; SARKAR, 2021; NELSON e COX, 2022). O conjunto dessas macromoléculas que intervêm diretamente e indiretamente nos processos biológicos é conhecido como proteoma que corresponde a quantidade proteica expressa por um organismo, bem como as inúmeras possibilidades de interação proteína-proteína e as modificações pós-traducionais (PTM), tornando-o bastante complexo (LUCK et al, 2017).

O proteoma de uma espécie modifica-se de acordo com o tipo celular, momentos e/ou condições fisiológicas distintas, sendo altamente dinâmico e variável quando comparado ao genoma que é essencialmente estático. Nesse contexto, a abordagem proteômica permite estudos em larga escala da expressão proteica em diferentes tecidos e fluidos corporais, tendo como objetivo identificar e quantificar o conteúdo proteico total, analisar o produto dinâmico do genoma em células em diferentes estados, determinar funções e interações proteicas e obter informações sobre as modificações que podem ocorrer nas proteínas após a sua tradução (HUSTOFT et. al, 2012). Há possivelmente 100,000 proteínas codificadas por aproximadamente 20,235 genes do genoma humano, e determinar explicitamente a função de cada uma é um desafio encontrado nas pesquisas atuais (JIANG et. al, 2020).

De acordo com Barbosa et. al (2012), o estudo da expressão gênica usando técnicas como análise de microarranjo, SAGE e sequenciamento em larga escala

utilizando equipamentos de última geração, fornecem um perfil molecular e oportunidades para identificar mudanças importantes que ocorrem no nível de RNA. No entanto, informações sobre processos que modulam a função e a atividade da proteína, como alterações pós-traducionais, interações proteína-proteína, transporte e degradação, são perdidas na análise de RNA. Além disso, a análise de transcritos também é prejudicada pela não-conformidade entre o transcrito e a concentração de proteínas e pela suscetibilidade a degradação (WANG et al., 2019; FRUNKIM et al., 2018).

A não-conformidade entre o proteoma e o transcriptoma pode ser explicada por diversos fatores tais como a taxa de tradução, modulação da taxa de tradução, modulação da meia vida da proteína, atraso da síntese proteica e transporte das proteínas (LIU; BEYER; AEBERSOLD, 2016). De modo geral, as pesquisas em proteômica provêm não somente a geração e acúmulo de novos dados, mas também permitem investigar e compreender os mecanismos envolvidos em patologias clínicas e desenvolver abordagens terapêuticas para tal fim (BARBOSA et. al, 2012).

O processo para análise proteômica consiste nas etapas de extração, separação, identificação e quantificação das proteínas expressas em amostras biológicas complexas. De modo geral, as metodologias empregadas em proteômica podem ser classificadas nos tipos *bottom up* e *top down*. A abordagem *bottom up* refere-se à caracterização das proteínas por análise dos peptídeos oriundos da digestão proteolítica de proteínas intactas. Em virtude da alta heterogeneidade das misturas proteicas, a escolha da enzima proteolítica é uma etapa importante no processo de identificação. De modo geral, as enzimas se diferem na especificidade da clivagem do resíduo de aminoácido na proteína. A tripsina é a enzima mais comum nesse processo e, trata-se de uma protease que gera fragmentos peptídicos clivando em resíduos de Arginina (R) ou Lisina (K) do lado C-terminal (NELSON e COX, 2022; PERUTKAET e ŠEBELA, 2018). No entanto, espectrômetros de massa de alta precisão reduziram a importância deste critério de filtragem e permitiram a identificação de sequências de proteínas não trípticas e modificações pós-traducionais onde a clivagem de tripsina é inibida (DONNELLY et. al, 2019). A mistura de peptídeos é fracionada e submetida a análise por Cromatografia Líquida acoplada à Espectrometria de Massas, em que cada fragmento peptídico obtido é comparado com os espectros de massa teóricos gerados a partir da digestão *in silico* de uma base de dados de proteínas. A inferência da proteína é realizada atribuindo sequências

peptídicas às proteínas correspondentes. As proteínas identificadas podem ser ainda marcadas e agrupadas com base nos seus peptídeos, visto que algumas proteínas possuem peptídeos exclusivos, e outras possuem peptídeos em comum (MILLER et. al, 2019). Neste trabalho, o foco será a abordagem *bottom-up* e, por esse motivo, as etapas serão descritas em detalhes.

Após a separação das misturas proteicas de interesse por eletroforese ou cromatografia líquida, o segundo processo consiste na análise e caracterização das moléculas. Nesse processo, é utilizado com maior frequência o MS devido à sensibilidade, precisão e poder resolutivo. MS é um método utilizado para detectar moléculas de interesse (analitos), que consiste em gerar, em fase gasosa, íons de compostos orgânicos por método adequado de ionização que são introduzidas em um campo elétrico e/ou magnético, permitindo separá-los de acordo com sua razão de massa/carga (m/z) (NELSON e COX, 2022). O espectrômetro de massas apresenta os seguintes componentes: entrada da amostra, fonte de ionização, analisador de massas e um detector (Figura 1). Um alto vácuo permite que os íons produzidos na fonte atinjam o detector e minimizem colisões entre íons e moléculas de ar. Na fase de ionização, os íons podem ser produzidos a partir de protonação assistida ou por alta voltagem. Os métodos mais comuns nesta fase são ionização por eletroaspersão e dessorção/ionização a laser assistida por matriz. (NELSON e COX, 2022; NIEHAUS et al., 2019).

Após a ionização das moléculas, elas passam para o analisador de massas que tem a função de separá-las de acordo suas massas (*m/z*). Esse processo gera um perfil de fragmentação conhecida como impressão digital da massa peptídica (do inglês, *Peptide Mass Fingerprint*, PMF) que é comparado por meio de um computador com todas as digestões de peptídeos teóricos a partir de uma base de dados de proteínas para identificar as melhores correspondências possíveis (MARQUIONI; NUNES; NOVO-MANSUR, 2021).

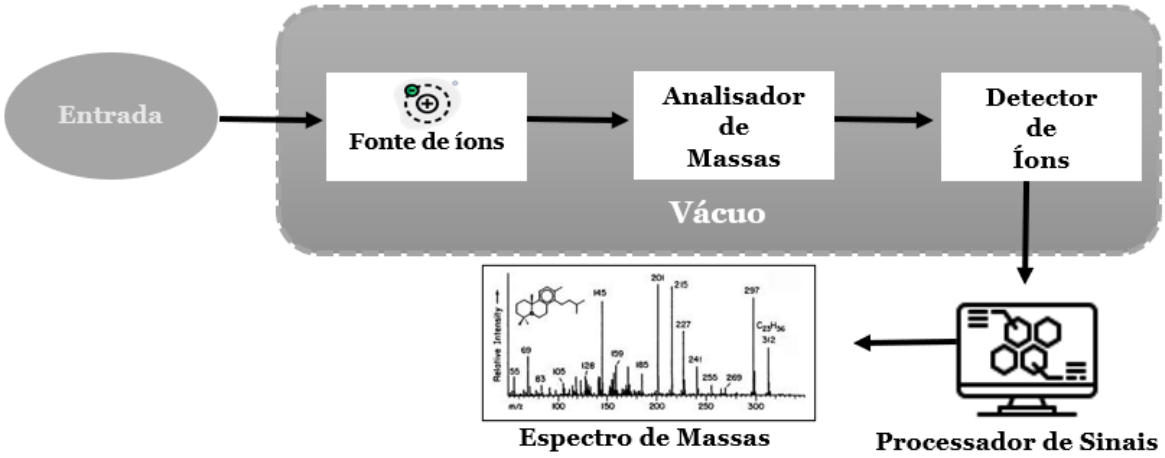


Figura 1 Componentes de um espectrômetro de massas.
As proteínas de interesse ou analitos são submetidas à um sistema à vácuo, que inicialmente ioniza-as removendo ou adicionando um elétron. Em seguida, o analisador de massas mede a razão massa carga das moléculas ionizadas. Por fim, essas moléculas têm o sinal neutralizado pelo detector e a corrente que flui é amplificada e convertida em sinal para ser processada pelo computador. Fonte: Autoria própria.

Em contrapartida, a abordagem *top down* é utilizada para caracterizar proteínas intactas, em que todo processo é realizado sem requerer digestão química ou enzimática. Em alguns casos, essa abordagem permite uma cobertura de 100% da sequência, determinar modificações pós-traducionais e isoformas (ZHANG et. al, 2013; CATHERMAN et. al., 2014). Ambas as abordagens apresentam algum tipo de limitação na caracterização de um proteoma. O método *bottom up* pode deixar de identificar grandes regiões importantes sobre PTMs ou variantes de splicing alternativo que podem ser perdidas após a digestão. Além disso, um ou vários peptídeos podem ser compartilhados entre as proteínas, ocasionando a ineficiência na identificação (Figura 2). O método *top down* limita-se quanto às dificuldades com o fracionamento, ionização da proteína e fragmentação na fase gasosa (MORADIAN et. al, 2014; ZHANG et. al, 2013; CATHERMAN et. al., 2014). Uma terceira abordagem, denominada *middle down*, analisa os fragmentos peptídicos maiores do que o método bottom up, minimizando a redundância peptídica entre as proteínas e permitindo obter mais informações sobre as modificações pós-traducionais, assim como no método *top down* (ZHANG et. al, 2014).

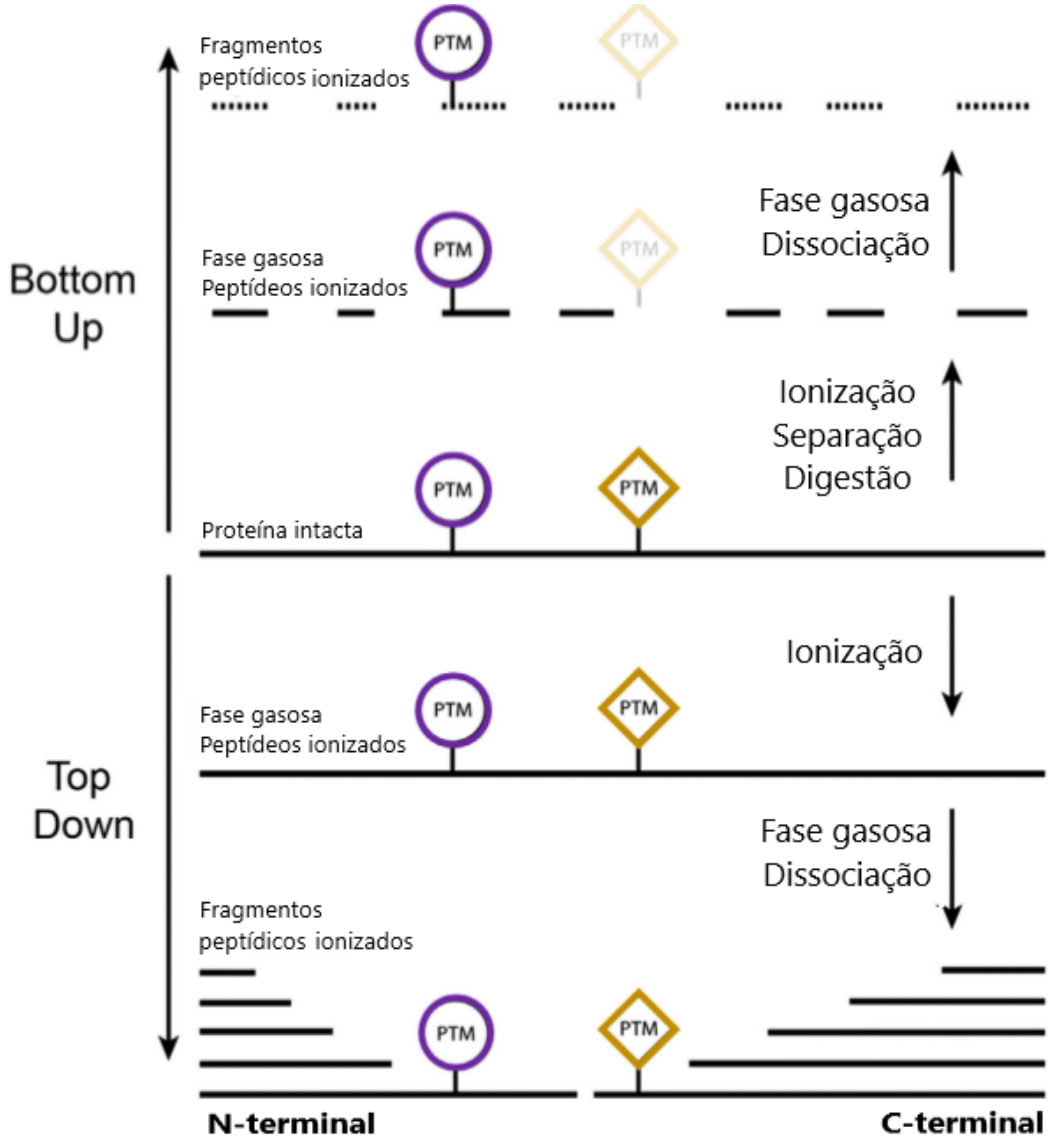


Figura 2 Comparação entre abordagens Top Down e Bottom Up.
Na abordagem Bottom Up, as proteínas intactas inicialmente são digeridas gerando peptídeos que são introduzidos no espectrômetro de massas onde são detectados e fragmentados. Na abordagem Top Down com espectrometria de massas, a proteína é ionizada diretamente, permitindo uma melhor cobertura de sequência e detecção de PTMs. Fonte: (CATHERMAN et. al., 2014)

A Espectrometria de massas do tipo *tandem* ou MS/MS é uma variação da configuração dessa metodologia. Esta configuração combina os pontos positivos de vários tipos de analisadores como forma de obter melhor desempenho. As combinações mais comuns incluem dois ou mais dos seguintes analisadores: Quadrupolos, TOF (do inglês, *time of flight*), ressonância ciclotrônica de íons ou orbitrap (GROSS, 2017).

A espectrometria de massa em *tandem* surgiu como uma ferramenta rápida e eficiente para a identificação de proteínas, podendo ser utilizada para sequenciar

trechos curtos de peptídeos. O método envolve, de maneira geral, dois estágios de análise de massas, juntamente com o processo de dissociação ou reação química que causa alteração na carga ou massa do íon (NELSON e COX, 2022; GROSS, 2017). O processo inicia-se com a hidrólise das proteínas investigadas por meio de uma enzima ou reagente químico formando fragmentos de peptídeos menores. Em seguida, a mistura que contém os fragmentos é injetada no equipamento formado por dois analisadores em tandem (MS1 e MS2). Assim, um primeiro analisador é usado para isolar um íon precursor, que então sofre espontaneamente ou por alguma ativação uma fragmentação para produzir íons de produto e fragmentos neutros. Esse íon então entra na célula de colisão e colide com um gás inerte, como hélio ou argônio, fragmentando o peptídeo que, geralmente, ocorre nas ligações peptídicas. Nesse processo, se a carga é retida no lado N-terminal o íon é do tipo b, se for no lado C-terminal é do tipo y. O segundo analisador de massa mede as razões m/z de todos os fragmentos carregados (Figura 3). Os picos formados consistem em todos os fragmentos que foram carregados pela quebra do mesmo tipo de ligação e inclui apenas fragmentos com cargas retidas, podendo ser no lado N-terminal ou C-terminal (NELSON e COX, 2022; GROSS, 2017).

Ao término desse processo, é gerado um perfil das massas dos peptídeos denominado impressão digital do fragmento peptídico (do inglês, *Peptide Fragmentation Fingerprint*, PFF) (BALLESTÉ, 2018). Esta abordagem não requer a interpretação em termos de sequência dos fragmentos observados. As massas observadas de íons fragmentados são comparadas com as esperadas para os vários peptídeos proteolíticos deduzidos de cada proteína contida na base de dados. Peptídeos como o mesmo conteúdo de aminoácidos e sequências diferentes podem apresentar a mesma massa molecular, mas terão padrões de fragmentação diferentes (BALLESTÉ, 2018).

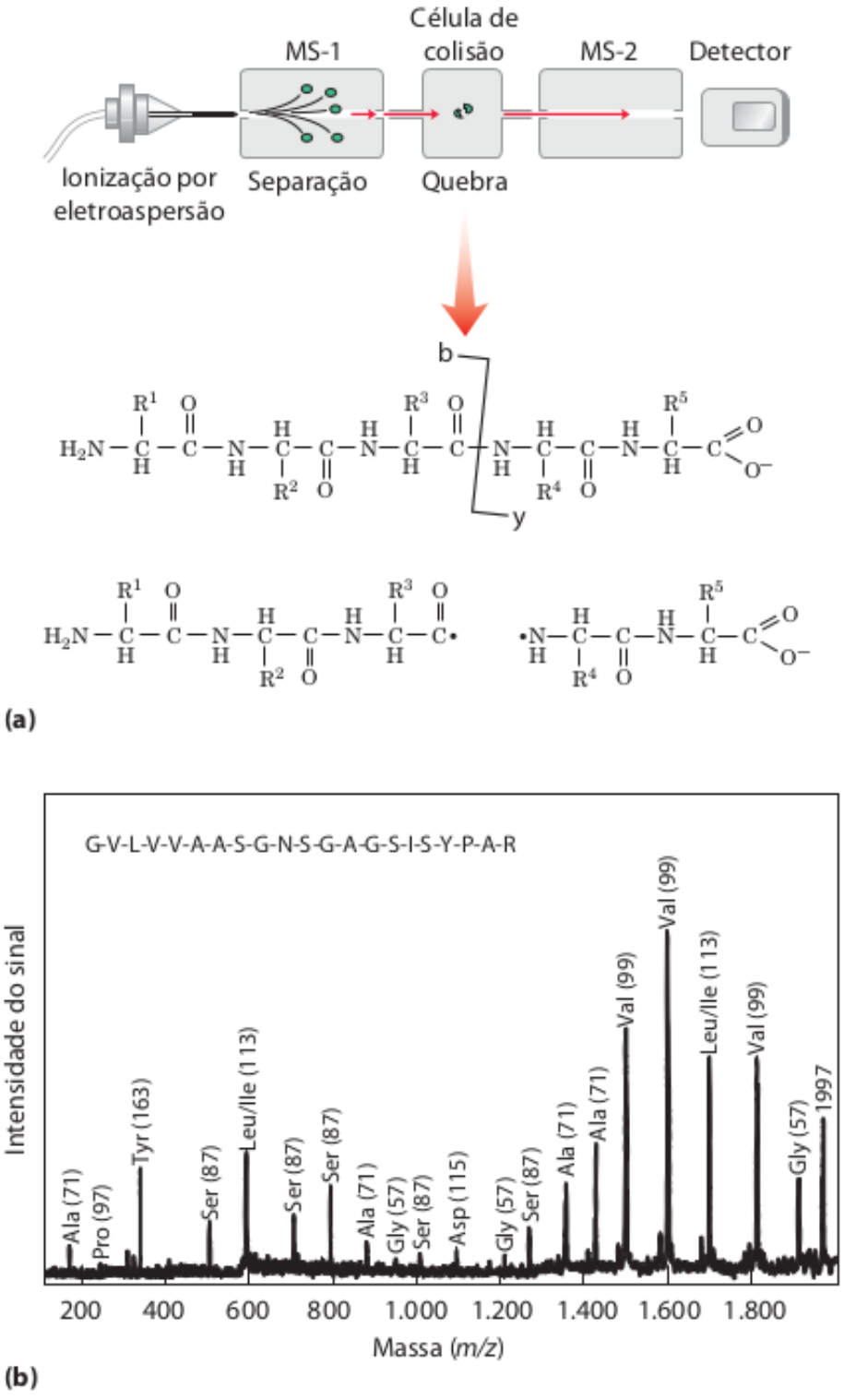


Figura 3. Identificação de proteínas utilizando espectrômetro de massas em tandem.
(a) No MS-1 a mistura proteica é disposta de modo que apenas um peptídeo é selecionado para anaĺise adicional. Em seguida, o peptídeo é fragmentado na célula de colisão. Muitos dos íons fragmentados resultam na quebra da ligação peptídica, gerando íons do tipo b e y. Se carga é retida no lado N-terminal é do tipo b, se for no lado C-terminal é tipo y. No MS-2 é obtido o valor *m/z* do fragmento analisado. **(b)** Picos representando os fragmentos dos peptídeos que foram gerados a partir de uma amostra contendo 21 resíduos de aminoácidos. A massa molecular dos aminoácidos é mostrada em parênteses acima dos picos. No topo está a sequência deduzida. **Fonte:** (NELSON e COX, 2022).

Como descrito, o processo de análise proteômica utiliza uma base de dados de proteínas teóricas para comparar com os valores de massa/carga obtidos pelo espectrômetro de massas, permitindo identificar as proteínas de interesse. Existem diferentes bases de dados para essa finalidade:

- **NCBI-protein**: Apresenta o conjunto das sequências proteicas de várias espécies, incluindo traduções de regiões codificantes anotadas pelo GenBank, RefSeq, TPA, *SwissProt*, PIR, PRF e PDB.

- **SwissProt/UNIPROT:** é um banco de dados ideal para pesquisas de perfil de fragmentação dos peptídeos, nas quais as sequências não possuem redundância (SUZEK et al., 2007). Portanto, pode haver menos correspondências para uma pesquisa do MS em série do que um banco de dados abrangente.

- **dbEST:** é uma base de dados derivada do GenBank que contém sequências de cDNA *single pass* ou *tags* de sequências expressas de um certo número de organismos.

Assim como as bases de dados, existem muitos softwares desenvolvidos para análise em larga escala de dados oriundos de espectrometria de massas. Esses softwares recebem como entrada uma base de dados de interesse do usuário. Nesse cenário, foram desenvolvidos dois tipos de busca em base de dados. O primeiro tipo, chamado *forward search*, compara o novo espectro com os espectros armazenados na base de dados e procura a melhor combinação dos espectros (BALLESTÉ, 2018). O segundo tipo, chamado de pesquisa inversa, verifica a possível presença nos novos espectros a partir de um espectro escolhido na base, gerando uma sequência peptídica revertida (ZHANG et al, 2018). A priori, uma filtragem é realizada para eliminar a maioria dos espectros da base considerados muito diferentes do novo espectro (BALLESTÉ, 2018).

Nesse contexto, há três tipos de algoritmos desenvolvidos para aumentar a precisão do método quando a busca é realizada em base de dados menores - o *Peptide Sequence Tag*, Autocorrelação e o *matching* baseada em probabilidade. O algoritmo *Peptide Sequence Tag* utiliza uma pequena série do fragmento do peptídeo chamada de aminoácido *tag* que pode ser usada para a identificação no banco de proteínas. A menor massa da série contém informações sobre a distância, em

unidades de massa, para um término do peptídeo, e a maior massa contém informações sobre a distância ao outro terminal peptídico. Juntos, o marcador de sequência peptídica consiste em três partes - a massa do N-terminal (m_1), a menor sequência de aminoácidos e a massa do C-terminal (m_3) (STEEN e MANN, 2004).

O algoritmo de autocorrelação é uma técnica de processamento de sinal que determina matematicamente a sobreposição entre o espectro teórico derivado de uma sequência na base de dados e o espectro experimental em questão. Para isso, o algoritmo calcula um *score* para os peptídeos da base de dados teórica e seleciona aqueles que apresentam valor igual ou similar ao valor do *score* do espectro experimental. No algoritmo de *matching* baseado em probabilidade, os fragmentos previstos são combinados aos fragmentos experimentais, começando com os íons -b e -y mais intensos. Para isso, calcula-se a probabilidade de que o número de correspondências de fragmentos seja aleatório e o logaritmo negativo desse número (multiplicado por 10) é a pontuação de identificação. (STEEN e MANN, 2004).

Dentre os softwares desenvolvidos para essa finalidade, pode-se citar o Mascot (*http://www.matrixscience.com/search_form_select.html*), *Protein Prospector* (*http://prospector.ucsf.edu/*) e o MaxQuant (http://www.maxquant.org). O MaxQuant é um pacote de software de proteômica quantitativa que suporta todas as principais técnicas de rotulagem, como SILAC, Di-metil, TMT e iTRAQ, bem como a quantificação sem rótulo. O software é de fácil manuseio e, portanto, permite análises de conjuntos de dados complexos em máquinas desktop por qualquer pesquisador que deseje empregar dados proteômicos. Além disso, ele integra uma infinidade de algoritmos, permitindo a análise completa dos dados LC-MS e oferece módulos adicionais para visualização de espectros e dados 3D. Este software tem como ponto forte a aplicação de algoritmos avançados em seu processo de busca que melhora substancialmente a precisão e a acurácia da massa. Como saída, o MaxQuant gera várias tabelas contendo as características dos peptídeos identificados. A tabela *evidence*, por exemplo, contém as informações combinadas sobre todos os recursos dos peptídeos identificados. A informação de identificação inclui a sequência peptídica e sequência modificada, comprimento e estado de modificação. Os campos como *Score, PEP (Posterior Error Probability), mass deviation, number of MS/MS matches, uncalibrated* e *calibrated m/z, mass error,* e *recalibrated retention time* podem ser usados para inspecionar a qualidade da identificação (TYANOVA et. al., 2016).

A sumarização do processo de anaĺise proteômica pode ser vista na Figura 4.

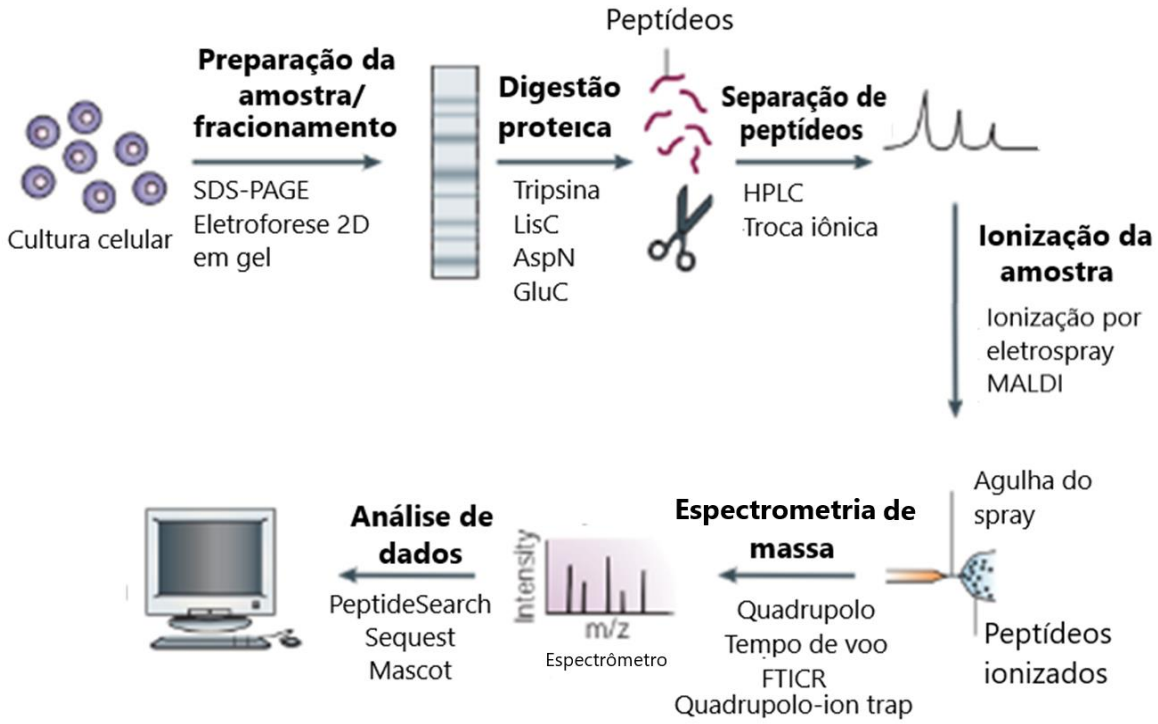


Figura 4. Etapas do processo de análise proteômica.
O processo inicia-se com a separação das proteínas de interesse utilizando alguma das técnicas, cromatografia ou eletroforese em gel. Em seguida, as proteínas de interesse são submetidas à uma enzima que gera fragmentos de peptídeos. Os peptídeos podem ser submetidos diretamente ao espectrômetro ou serem separados e fragmentados inicialmente. O espectrômetro gera como resultado o valor de massa/carga dos fragmentos que são utilizados para efetuar a busca no banco de dados teórico para a identificação das proteínas. Fonte: adaptado (STEEN e MANN, 2004)

1.2 EFEITOS DAS VARIAÇÕES GENÉTICAS NO PROTEOMA

Esta seção descreve como as alterações no DNA podem impactar na sequência de aminoácidos de uma proteína e interferir no processo de identificação por MS. O impacto das mutações descritas nesta seção também ajuda a compreender o processo de construção da base de dados proposta neste trabalho. Ademais, será descrito a relação das mutações com algumas doenças, como o câncer.

De modo geral, quando o aminoácido é substituído por outro quimicamente similar é chamada de substituição conservativa e, neste caso, a alteração pode afetar a estrutura e função da proteína em uma gravidade menor. A substituição não-conservativa produz um impacto maior na proteína, podendo torná-la inativa (LUO et. al., 2017; LENINGER et al., 2019). Para a análise proteômica, esses dois tipos de substituições podem afetar diretamente no processo de identificação da sequência, dependendo da base teórica utilizada. Por exemplo, a base de dados RefSeq/NCBI

possui somente sequências de referência, ou seja, não apresentam proteínas que possuem variações genéticas. Logo, utilizar essa base de dados permite que alguns peptídeos importantes para determinadas doenças não sejam identificados. Isso ocorre em virtude dos aminoácidos possuírem diferentes valores de massa molecular, com exceção da Leucina e Isoleucina como pode ser visto na Tabela 1. Desse modo, o perfil de fragmentação do peptídeo será alterado tornando o processo de identificação por MS ineficiente.

A principal alternativa para melhorar esse processo é o desenvolvimento de bancos de dados personalizados de proteínas em que são incorporados à sequência as possibilidades de variações de acordo com informações genômicas ou transcriptômicas (VÉGVÁRI, 2016). Na perspectiva proteômica, essas variações genéticas criam proteoformas (SMITH e KELLEHER, 2013) que tornam a complexidade proteica ainda maior, assim a identificação dessas alterações também fornece mecanismos para o desenvolvimento de drogas eficientes (VÉGVÁRI, 2016).

Tabela 1. Símbolos e massas dos resíduos dos aminoácidos que constituem as proteínas.

Nome	Símbolo	Massa do resíduo
Alanina	A, Ala	71,079
Arginina	R, Arg	156,188
Asparagina	N, Asn	114,104
Ácido aspártico	D, Asp	115,089
Cisteína	C, Cys	103,089
Glutamina	Q, Gln	128,131
Ácido glutâmico	E, Glu	129,116
Glicina	G, Gly	57,052
Histidina	H, His	137,141
Isoleucina	I, Ile	113,160
Leucina	L, Leu	113,160
Lisina	K, Lys	128,17
Metionina	M, Met	131,199

Fenilalanina	F, Phe	147,177
Prolina	P, Pro	97,117
Serina	S, Ser	87,078
Treonina	T, Thr	101,105
Tirosina	Y, Tyr	163,176
Valina	V, Val	99,133

As diferenças nas sequências de DNA humano, também conhecidas como polimorfismo, contribuem para as variações fenotípicas e influenciam a suscetibilidade de um indivíduo à doenças, à respostas ao meio ambiente e a tratamentos com drogas. O termo polimorfismo, relacionado à genômica, refere-se à presença de duas ou mais formas variantes de uma sequência específica de DNA que pode ocorrer entre diferentes indivíduos ou populações. Outros polimorfismos podem ser muito maiores, envolvendo trechos mais longos de DNA. O tipo mais comum de polimorfismo envolve variação em um único nucleotídeo único (do inglês, *Single Nucleotide Polymorphism*, SNP) e ocorrem aproximadamente a cada 1,200 bases na população humana em geral (STEF et. al., 2013). Na maioria das vezes, os SNPs ocorrem na região não-codificante do genoma. SNPs em regiões codificantes do DNA podem resultar em mudanças de sequências de aminoácidos através de substituições de aminoácidos ou por meio da introdução de códon de parada prematura. Além dos SNPs, podem ocorrer inserções ou deleções (INDELS) que podem gerar mudança no quadro de leitura e alterar a sequência de uma proteína significativamente. Essas alterações também podem ocorrer em regiões não-codificantes e codificantes (STEF et. al., 2013). Os INDELs parecem ocorrer em frequências menores que as SNPs, porém afetam coletivamente mais pares de bases devido ao seu tamanho maior de alterações. Pequenas deleções são aproximadamente três vezes mais comuns do que pequenas inserções e, para ambos os tipos de variação, a frequência de mutação diminui com o aumento do tamanho do fragmento (NOLL et al., 2016).

Nesse contexto, as variações que afetam diretamente a proteína podem ser classificadas em diversos tipos, porém serão descritas apenas àqueles que são necessárias para a compreensão deste trabalho. Assim, de acordo com o tipo de alteração causada pela mutação genética que afeta a proteína, ela pode ser categorizada como polimorfismo de aminoácido único (do inglês, *Single Amino Acid*

Polymorphism, SAP) e INDELS (KREBS et, al, 2014). O SAP ocorre em consequência de SNPs que alteram a sequência de DNA, impactando diretamente na proteína. O polimorfismo genético está na categoria de mutações pontuais que são classificadas como sinônimas e não sinônimas. Essa classificação é baseada em sua capacidade de alterar os aminoácidos codificados e, em alguns casos, afetar a função da proteína (Zheng et. al., 2014). As mutações não sinônimas podem ser subdivididas em missense, nonsense, variação de região de tradução e perda de códon de parada.

A mutação sinônima ou silenciosa ocorre quando a troca de base nucleotídica modifica o códon, mas não altera o aminoácido (Figura 5). Isso acontece devido o código genético ser degenerado, permitindo que um aminoácido seja codificado por mais de um códon (BUHR et. al, 2016). Embora esse tipo de mutação não altere a sequência de aminoácidos da proteína, alguns estudos têm revelado que podem estar relacionados com patogenicidade e alteração nos processos celulares. Dentre esses processos essa mutação interfere no processo de conformação da proteína e no *codon usage* que é relacionado à taxa de tradução proteica, podem alterar a estrutura secundária do RNA mensageiro e podem impactar na cis-regulação de sequência de sítios de splice, miRNA e sítios de fatores de transcrição (BUHR et. al, 2016; HUNT et. al., 2014).

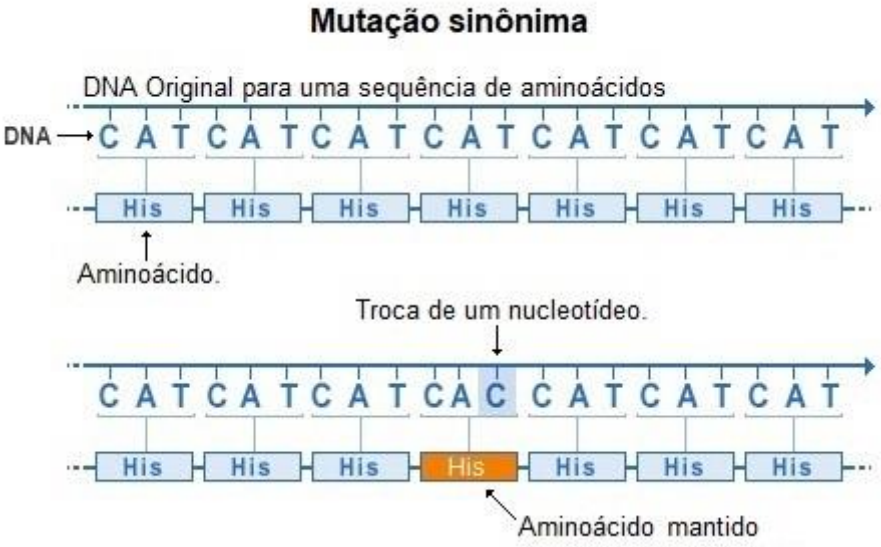


Figura 5. Mutação do tipo sinônima.
Neste caso, ocorre uma troca de base T para C no códon CAT que codifica uma Histidina. Porém, o novo códon gerado CAC também codifica Histidina, mantendo a sequência proteica inalterada. Fonte: Adaptado - U.S. National Library of Medicine

A mutação *missense* ocorre quando a troca de base altera o códon e, consequentemente, o aminoácido, como pode ser visto na Figura 6. Essa mutação

pode afetar a estrutura 3D, alterar a estabilidade ou afinidade de ligação de um complexo proteico e causar perturbações significativas ou remoção completa da função desta proteína particular (STEFL, et al. 2013; CHEN et al., 2020). Em alguns casos, pode provocar a alteração da flexibilidade de toda a molécula ou apenas uma pequena região, a alteração do equilíbrio entre diferentes conformações ou ainda, afetar toda a dinâmica conformacional da molécula (STEFL, et al. 2013; CHEN et al., 2020). Em virtude dessas modificações, essa mutação está associada a alguns tipos de câncer como câncer de ovário e colorretal, e doenças como parkinson e diabetes (LI et. al. 2013; UVERSKY et. al., 2009; KUMAR 2013).

Além das variações *missense*, uma grande fração de variações genéticas humanas patológicas é causada por mutações *nonsense* onde uma variação de um único nucleotídeo introduz um códon de terminação prematuro, gerando uma proteína mais curta, como pode ser visto na Figura 7 (FOLKMAN et. al., 2015). Muitas dessas proteínas geradas são funcionais, porém elas não são sintetizadas em quantidade suficiente devido a degradação do RNA mensageiro pelo mecanismo *nonsense-mediated decay.* Este mecanismo é responsável pela rápida degradação de muitos RNAs mensageiros aberrantes que contém códon de parada prematura (VICENTE-CRESPO e PALACIOS, 2010).

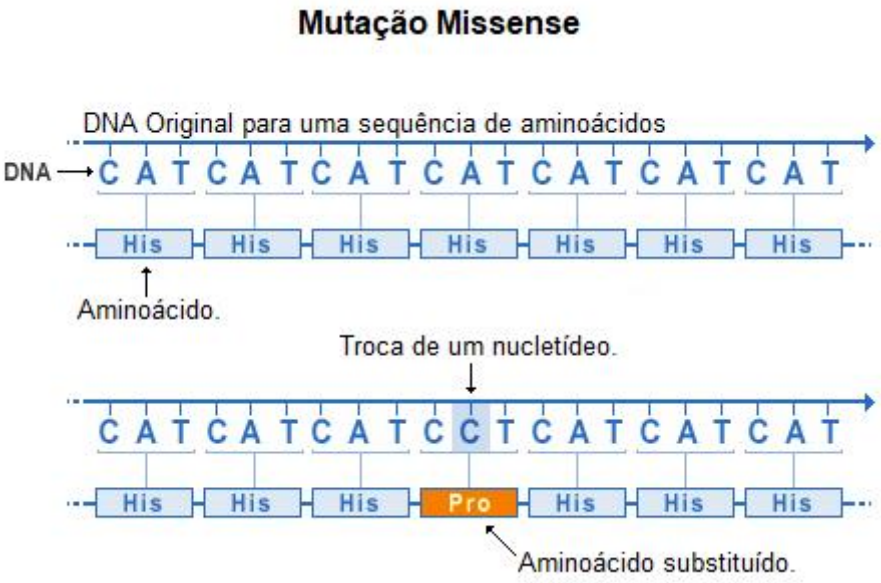


Figura 6. Mutação do tipo *missense*.
Neste caso, ocorre uma troca de base A para C no códon CCT que codifica uma Histidina. O novo códon gerado CTC agora codifica uma Prolina, modificando a sequência da proteína. Fonte: Adaptado - U.S. National Library of Medicine

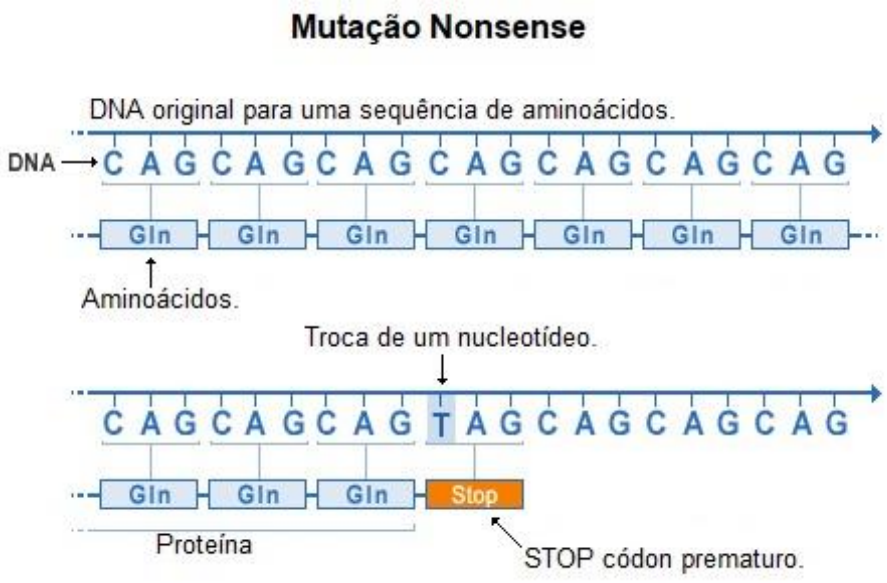


Figura 7. Mutação do tipo *nonsense*.
Neste caso, ocorre uma troca de base C para T no códon CAG que codifica uma Glutamina. Essa troca gera o códon TAG, inserindo na proteína um códon de parada prematura. Fonte: Adaptado - U.S. National Library of Medicine

Como resultado da mutação *nonsense*, estudos descrevem que ela tem um impacto funcional significativo e é potencialmente causadora de doenças, como Alzheimer, Demência lentamente progressiva, Insuficiência autonômica central com comprometimento cognitivo, alucinações visuais e acústicas (GUERREIRO et. al., 2014). Além disso, pacientes que apresentam os genes ATM (E1978X) e BLM (Q548X) com mutação nonsense estão susceptíveis ao câncer de mama (PROKOFYEVA et al, 2013; BOGDANOVA et al, 2009). No entanto, o trabalho descrito por Antczak et al. (2013) mostrou que o gene BLM (Q548X) não está associado ao aumento do risco de câncer de próstata, e não afeta a sobrevivência de homens com esse tipo de câncer.

A mutação *missense* também pode alterar a região não traduzida do RNA mensageiro (do inglês, *Untranslated Region*, UTR) por meio da substituição de base no códon inicial. Essa alteração faz com que o ribossomo busque outro início de tradução, como pode ser visto na Figura 8. Desse modo, a maquinaria de tradução composta pelo complexo de pré-iniciação 43S, varre a região 5' não traduzida do RNAm até a tripla de códon inicial de nucleotídeos ATG utilizando complementaridade com o anticódon de Met-RNAt (HINNEBUSCH et. al., 2016; CHOI et. al., 2015; KO e CHOW, 2003).

Em eucariotos, o ATG é quase exclusivamente o códon usado para iniciação. Nas células de mamíferos, o ACG e o CTG foram encontrados como códons iniciadores em alguns mRNAs. Embora seja mais comum em procariotos, estudos envolvendo a mutagênese do códon ATG demonstrou que o CTG, TTG, GTG, AAG, ACG, AGG, ATA, ATC e ATT podem iniciar a síntese proteica em vários graus *in vivo* e *in vitro*. Estes códons de iniciação mutantes diferem do ATG por um único nucleotídeo. Como eles usam o iniciador do tipo selvagem tRNA para iniciação, a síntese proteica em todos esses casos é iniciada com metionina (REUTER et. al., 2016; DRABKIN et. al., 1998).

As mutações que afetam o códon inicial da proteína estão associadas à algumas patologias como Nistagmo ocular infantil e problemas neurológicos relacionados ao atraso no desenvolvimento, epilepsia e microcefalia (CHOI et. al., 2015; PAGNAMENTA et. al., 2018). A mutação do códon de início em CLN3 - distúrbio que afeta predominantemente a retina e cérebro - está associada a um curso prolongado da doença (KUPER et al., 2020). No trabalho descrito por Cyr et al. (2012) relata uma paciente com câncer de ovário que carrega uma variante germinativa MSH2 que altera o códon de iniciação da tradução e, embora não seja um alelo de doença forte e de alto risco, pode ter um impacto moderado no fenótipo da doença.

Além da alteração no códon inicial, a mutação *missense* também pode ocasionar a perda de códon de parada, que são responsáveis pela sinalização da terminação da tradução. Neste caso, um processo no qual um códon de parada não é reconhecido, a síntese proteica será finalizada no próximo códon de parada, gerando uma proteína com uma extensão, como pode ser visto na Figura 9. (ZHANG et. al., 2012). Em um estudo feito na população chinesa mostrou que mutações em códon de parada estão associadas a doenças neurológicas como doença de Parkinson e tremor essencial, apesar de serem raras (HE e HUANG, 2016). Zhang et. al. (2012), sugere em seu estudo feito com hamsters chineses que a substituição do códon de parada TAA (Stop) para GAA (Glu) na Imunoglobulina IgG1 pode aumentar o risco de imunogenicidade porque a sequência adicional não faz parte da região constante da cadeia. No estudo desenvolvido por Marlin et al. (2020) mostraram que a mutação germinativa (p.Ter285Lysfs) rs77179853 no gene HOXB13 está associada com câncer de próstata. Dhamija et al. (2020) mostram em sua pesquisa que as mutações no códon de parada podem ser funcionalmente

importantes no câncer e caracterizar seu impacto de perda de função no supressor de tumor SMAD4.

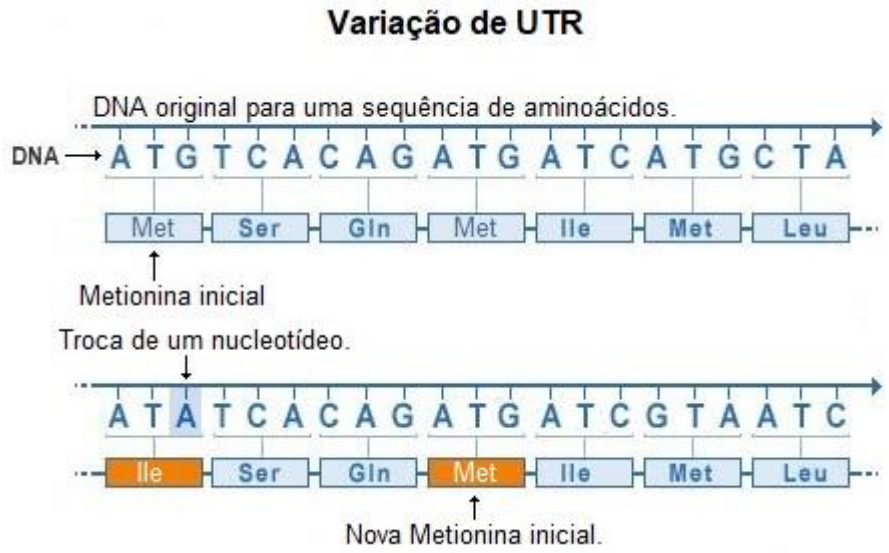


Figura 8. Mutação do tipo Variação de UTR.
Esse tipo de mutação altera o início de tradução da proteína, por meio da modificação do códon inicial. Neste exemplo, ocorre a troca de G para A no códon ATG que codifica a Metionina inicial, gerando o novo códon ATA que codifica uma Isoleucina. Essa alteração faz com que o ribossomo busque outro início de tradução, ampliando a região não traduzida e reduzindo a proteína no início. Fonte: Adaptado - U.S. National Library of Medicine

Figura 9. Mutação do tipo *Stop Loss*.
A substituição no códon de parada da proteína permite que o ribossomo continue o processo de tradução até localizar o próximo códon de parada, ocasionando uma extensão na cadeia proteica. Fonte: Adaptado - U.S. National Library of Medicine

A mutação do tipo INDEL é a segunda categoria de mutações polimórficas e podem alterar ou manter a janela de leitura. Quando as inserções e deleções são múltiplas de 3 e ocorrem entre os códons, o frame de leitura (*inframe*) permanece

inalterado, ocasionando apenas a inserção de novos aminoácidos, conforme mostrado na Figura 10. Essas mutações desempenham papéis distintos em doenças como câncer, podendo atuar como supressores ou oncogênicos. No estudo realizado por Baeissa et. al. (2017) envolvendo diversos tipos de mutações em domínios proteicos, foram encontrados 5 domínios (zf-C2H2, IL6Ra-bind, bZIP_2, PI3K_p85B e Myb_DNA-bind_6) com mutações do tipo indel associados à oncogênese, e apenas 2 domínios supressores de tumor (PIK3R1 e TP53) cada um de uma única proteína. De modo similar, Guan et. al. (2012) identificaram mutações indel em cânceres ginecológicos que possuem impacto nos mecanismos relacionados à função supressora de tumor do domínio proteico ARID1A.

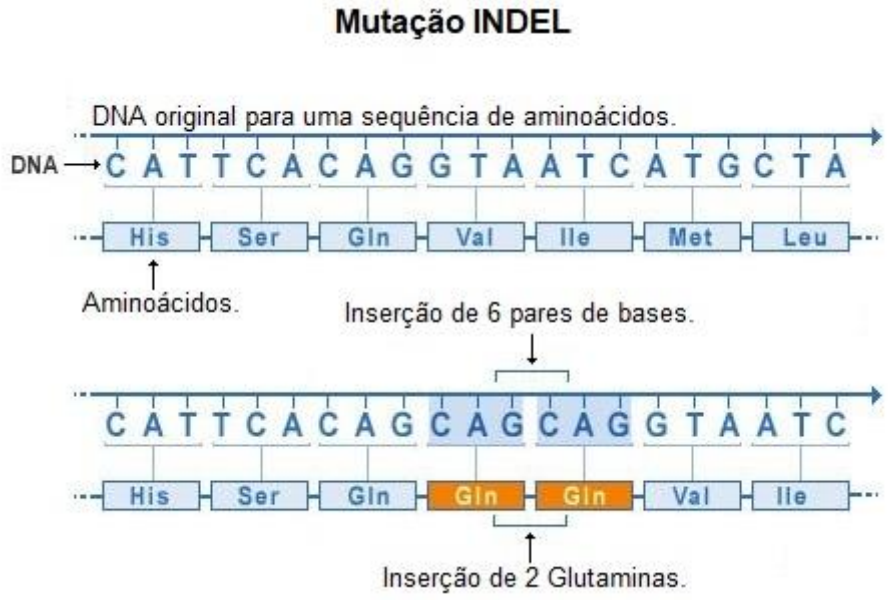


Figura 10. Mutação do tipo INDEL.
Este tipo de mutação é decorrente da inserção ou deleção de pares de bases que, quando ocorridas na região codificante, altera a sequência proteica. Neste exemplo, houve a inserção de 6 pares de bases (CAGCAG) dentro do *frame* de leitura, ocasionando na inserção de dois novos resíduos de Glutamina. O processo de deleção ocorre de modo inverso, em que as bases são deletadas do genoma e, consequentemente, os aminoácidos codificados por elas.
Fonte: Adaptado - U.S. National Library of Medicine

As mutações INDEL que alteram a janela de leitura (*frameshift*) têm um efeito maior na sequência proteica. Esse tipo de alteração é decorrente de inserção ou deleção que não é múltipla de 3 ou quando ocorre entre os códons, assim, o frame de leitura é alterado, resultando na modificação da cadeia de aminoácidos, a partir do ponto da mutação, como pode ser visto na Figura 11. Em mutações desse tipo, é comum ocorrer códon de parada prematura ou extensão da cadeia proteica.

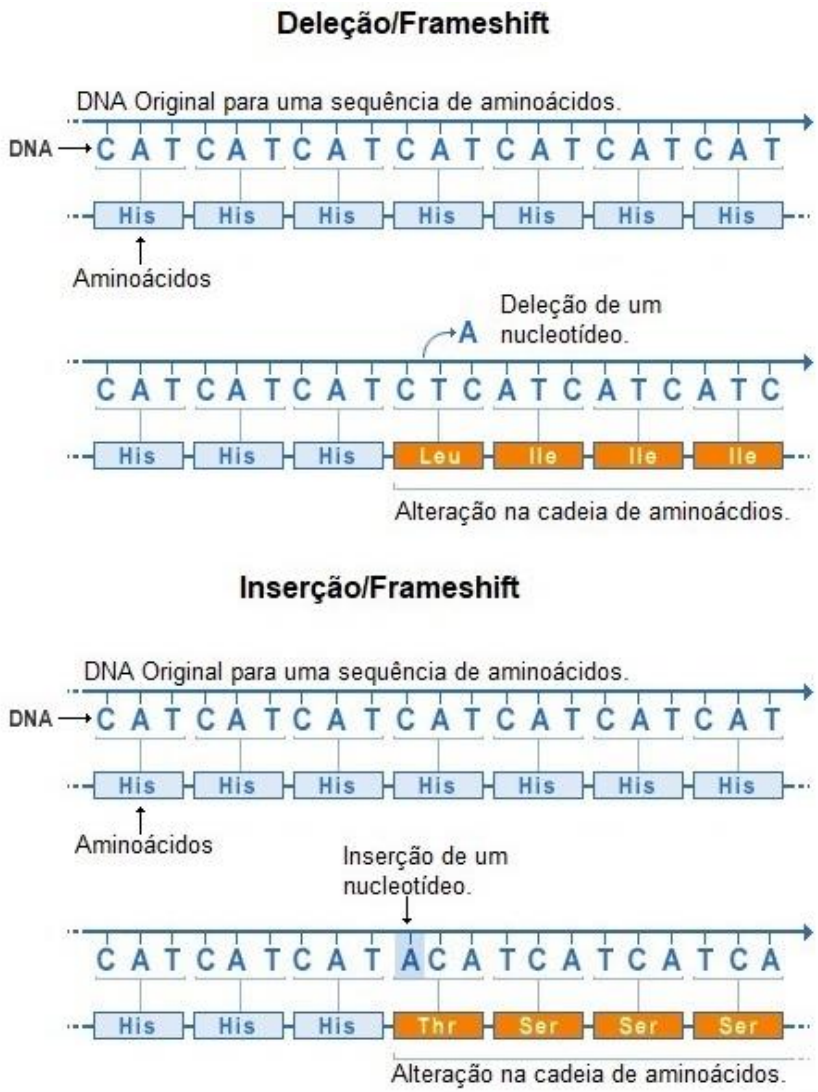


Figura 11. Mutação do tipo *frameshift*.
No primeiro caso, ocorre uma deleção da base A no códon CAT (His), gerando o códon CTC (Leu), movendo do frame para esquerda. De modo similar, no segundo caso, ocorre a inserção da base A antes do códon CAT (Hist) gerando o códon ACA (Thr), movendo o frame para direita. Em ambos os casos, a sequência de aminoácidos é modificada a partir do ponto da mutação, encerrando-se até a localização do primeiro códon de parada. Fonte: Adaptado - U.S. National Library of Medicine

As mutações *frameshift* estão associadas à oncogênese, causando diminuição na expressão gene TP53 relacionado à supressão tumoral (MERTINS et. al., 2016). No fator de transcrição GATA3, esse tipo de mutação ocorre em aproximadamente 15% dos cânceres de mama com receptor de estrogênio positivo, ou tipo luminal. Este gene é essencial para o desenvolvimento da mama e, quando mutado, estimula o crescimento tumoral *in vivo*, por meio da regulação positiva e negativa em células mutantes (GUSTIN et. al., 2017).

1.3 PROTEOGENÔMICA: CONCEITOS, APLICAÇÕES E ABORDAGENS COMPUTACIONAIS

O termo proteogenômica foi introduzido por Jaffe *et* al. (2004) e inicialmente usado para descrever estudos em que dados proteômicos são usados para melhorar a anotação genômica e a caracterização das proteínas codificantes. Esta abordagem permite que novos peptídeos sejam identificados por meio de busca dos espectros MS/MS contra banco de dados de proteínas personalizados, contendo novas sequências proteicas e sequências variantes que são gerados utilizando informações genômicas e transcriptômicas. O principal ponto desta abordagem é que permite não somente melhorar a anotação genômica a nível proteico e refinar os modelos de genes, mas também permite aprimorar os bancos de dados de sequências proteicas (NESVIZHSKII, 2014).

A análise de proteogenômica fornece informação única para a anotação do gene, (a) confirmando a tradução e separação dos pseudogenes dos genes codificantes; (b) estabelecendo que uma proteína não é alvo de degradação (c) determinando automaticamente o *frame*, mesmo múltiplos frames sobrepostos; (d) restringir a localização dos locais de início e fim de tradução, bem como locais de processamento pós-tradução; (e) identificação de limites exatos de *splicing* e formas alternativas de *splice*, se o peptídeo for dividido entre exons; e (f) predizer um gene completamente novo, mapeando para uma localização genômica não caracterizada (CASTELLANA e BAFNA, 2010; KROLL et al., 2014; MACHADO et al., 2019; MADISON et al., 2022).

No contexto de anotação gênica, os peptídeos acetilados no N-terminal podem ser utilizados para a validação ou correção da atribuição de metionina inicial em genes anotados como codificantes. Esse processo normalmente ocorre durante eventos co- e pós-traducionais que envolve a clivagem da metionina inicial por aminopeptidases seguida por acetilação do aminoácido seguinte por N-acetiltransferases (RENUSE et. al., 2011). Essa abordagem torna-se bastante útil visto que em sítios de início de tradução de proteínas, os códons de início são difíceis de prever, já que o contexto da sequência geralmente não é bem definido. Isso pode ocorrer devido o reconhecimento ineficiente de um códon de iniciação, fazendo com que uma parte do complexo de

pré-iniciação 43S de tradução continue lendo e iniciando a tradução em um sítio *downstream*, um processo conhecido como “varredura com vazamento” (LEE et. al., 2012; JACKSON et. al., 2010; SMIRNOVA et al., 2022). A proteogenômica também pode ser aplicada à análise de variações genéticas, anotação de genomas não sequenciados, estudos de mecanismos patológicos e metaproteômica (NESVIZHSKII, 2014; RENUSE et. al. 2011; MACHADO et al., 2019).

Em relação às variações genéticas, sabe-se que elas podem alterar as funções em produtos gênicos e serem responsáveis por patogenicidade específica a um indivíduo. A identificação dessas variações no nível da proteína fornece uma oportunidade para reduzir o conjunto de candidatos a investigação do seu papel funcional ou relevância clínica (NESVIZHSKII, 2014). Além dessas, existem outras múltiplas fontes de variação do genoma de alta significância biológica que resulta potencialmente em transcritos codificadores de proteínas novos ou variantes. Estes incluem a edição de RNA, que ocorre durante o processamento pós-tradução e cujo papel e significado biológico ainda não foram totalmente compreendidos (SORTINO et al., 2022). Nesse caso, a proteogenômica pode fornecer evidências valiosas de nível de proteína para alguns desses supostos eventos de edição de RNA. Pode também fornecer evidências de expressão proteica para novas fusões gênicas e transcritos quiméricos e transcritos anotados como pseudogenes (NESVIZHSKII, 2014).

As alterações genômicas específicas no genoma de uma pessoa podem resultar em uma doença que geralmente é causada pela alteração da função proteica. Esta alteração pode ser produto de um gene recombinante que contém porções de dois ou mais genes diferentes de modo que suas sequências codificadoras estejam na mesma matriz de leitura, gerando uma proteína de fusão (SNUSTAND e SIMMONS, 2013). Embora abordagens principalmente genômicas sejam usadas para a identificação de genes de fusão, abordagens proteômicas como espectrometria de massas também podem ser usadas para a identificação e caracterização de proteínas de fusão, o que pode ser útil no estudo dos estados alterados ou doentes. A identificação de peptídeos de fusão utilizando proteogenômica pode ser útil para desvendar mecanismos de doença, como algumas PTMs, como a fosforilação que desempenham um papel importante em muitas doenças (RENUSE et. al., 2011).

Por fim, alguns dos elementos essenciais para proteogenômica são listados abaixo, de acordo com Renuse et. al. (2011):

- **Disponibilidade dos dados da sequência do genoma**: É importante ter os dados da sequência do genoma (idealmente montados) para o organismo em estudo. Na falta de informação disponível sobre a sequência do genoma, a disponibilidade de genomas semelhantes pode ser utilizado como alternativa.
- **Dados de espectrometria de massa de alta resolução e alta precisão:** Como o espaço de pesquisa é aumentado ao pesquisar em bancos de dados de genoma, é importante ter alta resolução e alta dados de precisão para reduzir a ocorrência de falsos positivos.
- **Ferramentas de pesquisa e anotação de banco de dados genoma:** A maioria dos algoritmos de busca em banco de dados não permite pesquisas diretas contra bancos de dados de genoma traduzidos dos seis *frames* de leitura. Mesmo quando isso é possível, o mapeamento direto de peptídeos em estruturas de genes não é trivial.

1.4 PROBLEMATIZAÇÃO, HIPÓTESE E JUSTIFICATIVA.

Como descrito, o processo de análise proteômica consiste em etapas como separação do complexo proteico, digestão proteolítica por uma enzima sítio-específica, ionização, detecção e definição dos valores de razão massa/carga dos fragmentos peptídicos utilizando espectrometria de massas. Esse processo é concluído utilizando softwares analisadores de espectros que buscam em base de dados de proteínas teóricas, fragmentos peptídicos com massa correspondente ao fragmento experimental. Nesse processo, existem algoritmos que realizam análise de correlação ou de marcação de série dos peptídeos e geram como saída todos os possíveis peptídeos teóricos com uma correspondência aceitável quando comparado ao experimento.

Desse modo, o desafio da área é mapear corretamente sequências peptídicas identificadas para sequências de proteínas, particularmente para proteínas redundantes e isoformas. Além disso, à medida que os métodos proteômicos continuam a se tornar mais sensíveis e abrangentes e os bancos de dados de proteínas aumentam de tamanho, a capacidade de atribuir adequadamente os peptídeos às proteínas tornou-se ainda mais desafiadora. Quando as sequências peptídicas são compartilhadas entre as proteínas, elas são mais prevalentes e

abundantes do que suas contrapartes peptídicas únicas e, portanto, também mais facilmente identificadas (ZHANG et. al., 2013). Por outro lado, peptídeos únicos são mais difíceis de identificar por serem menos abundantes e, infelizmente, carregam a maior evidência experimental para a identificação inequívoca e confiável de proteínas homólogas e redundantes (ZHANG et. al., 2013). O aumento do espaço de busca em banco de dados teóricos gera mais peptídeos candidatos para ser pontuado contra um espectro experimental MS/MS, aumentando a probabilidade de que a melhor pontuação corresponda ao espectro incorreto, e também se torna mais difícil distinguir entre as identificações verdadeira e falsa (NEVIZHSKII, 2014; RENUSE et. al., 2014). Esse processo é ilustrado na Figura 12A.

Como mostrado na subseção 1.2, as variações genéticas têm grande impacto na sequência proteica, podendo haver simples trocas de aminoácidos até alterações de grande parte da cadeia polipeptídica que dificultam o processo de identificação correta dos peptídeos. Isso ocorre em virtude da limitação dos bancos de dados quanto à incorporação dessas variações (por exemplo, UniProt/SwissProt). Como mostrado, os bancos de dados apresentam sequência proteicas consideradas referências para um organismo específico, desconsiderando mutações e polimorfismos genéticos. Sabe-se que organismos da mesma espécie podem ter quase 99,9% de similaridade genômica e que os SNPs, inserções e deleções e variações no número de cópias (CNVs) contribuem para variações nos genomas, podendo causar alterações na função de um produto gênico. Essas variações também podem ser responsáveis por respostas específicas de indivíduos a patógenos, como bactérias e vírus (VÉGVÁRI, 2016; RENUSE et. al., 2011). Desse modo, utilizar esses bancos para identificação proteica nas distintas populações genéticas, se torna uma tarefa desafiadora para a análise proteômica. Essa descrição é ilustrada na Figura 12B.

Para os problemas mencionados, existem diversos estudos que utilizam como saída, a customização de banco de dados específicos às mutações e/ou polimorfismos genéticos em determinadas populações. Uma das alternativas amplamente utilizadas é a criação de uma base de dados gerada a partir da tradução dos 6 quadros de leitura. Porém, essa abordagem limita-se por aumentar o espaço de busca da base de dados e falha na captura de peptídeos presentes na junção entre exons (NEVIZHSKII, 2014; RENUSE et. al., 2011).

Outras abordagens desenvolvem base de dados incorporando as variações nas sequências de referência (UniProt, por exemplo), muitas vezes utilizando um caractere especial para distinguir a sequência variante. As variações podem ser baixadas do banco de dados dBSNP do NCBI e complementadas com mutações de doenças conhecidas do banco de dados públicos como *Online Mendelian Inheritance* in Man e *Protein Mutant Database*. Para criar bancos de dados personalizados a partir de dados de RNA-seq, o customProDB pode combinar pequenas variações com algumas já conhecidas extraídos do banco de dados dbSNP (NEVIZHSKII, 2014; VÉGVÁRI, 2016). Essas abordagens apresentam eficiência na identificação de peptídeos variantes, porém necessitam de métodos para análise de redundância para melhorar a confidência da identificação.

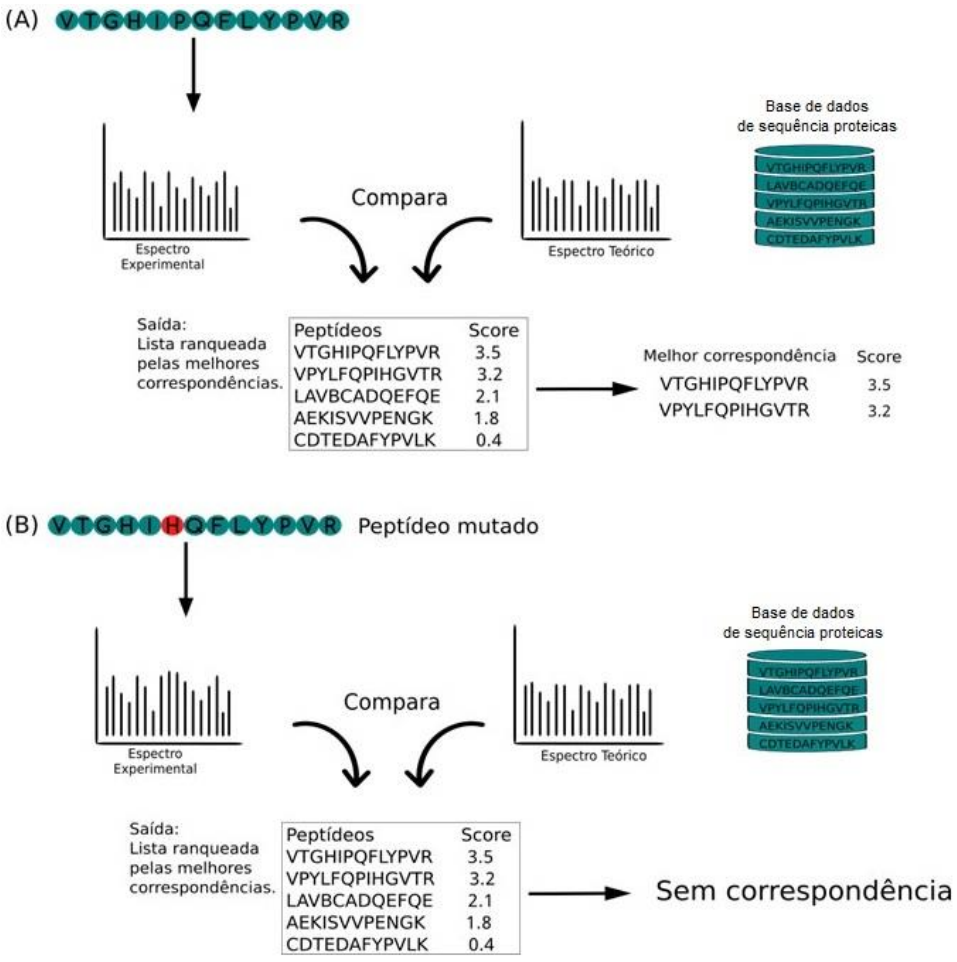


Figura 12. Processo de identificação dos peptídeos por softwares analisadores de espectros de massa. **(A)** O espectrômetro resulta no perfil de fragmentação do peptídeo que é utilizado para realizar a busca em base de dados teórica por peptídeos com espectro correspondente. Neste caso, dois peptídeos obtiveram perfil semelhante ao peptídeo experimental, sendo que o primeiro apresenta melhor score. O número de correspondência pode aumentar de acordo com o tamanho do banco de dados. **(B)** O peptídeo experimental apresenta uma substituição do resíduo de aminoácido P por H, alterando o perfil de fragmentação do peptídeo. A busca é ineficiente visto que a base de referência não apresenta esta variação. Fonte: Autoria própria

Nesse sentido, a criação de um banco de dados personalizado utilizando informações de base de dados públicas contendo polimorfismos e a análise posteriori dos peptídeos identificados por essa base, pode ser uma alternativa para reduzir problemas como espaço de busca, redundância de peptídeos e identificação de variantes proteicas. Para tanto, as bases de dados referência e variantes seriam submetidas separadamente ao software identificador e em seguida, analisados de acordo com critérios estabelecidos para aumentar confidência da identificação.

1.5 TRABALHOS RELACIONADOS

No contexto de banco de dados personalizados, existem alguns trabalhos que incorporaram variações relacionadas aos polimorfismos e que tenham algum impacto clínico. Assim, Schandorff et. al. (2007) modificaram o banco de dados *International Protein Index* (IPI) com o objetivo de mantê-lo relativamente compacto e maximizar a chance de identificar variantes de sequência. Para isso, as entradas originais do IPI foram mantidas intactas, alongando-as apenas com sequências peptídicas adicionais. Para permitir que os mecanismos de pesquisa de banco de dados pudessem compreender os dados das novas variações presentes foram atribuídas letras que não representam aminoácidos, por exemplo J, como espaçadores entre o final da sequência e os peptídeos subsequentes referentes às variações. Como resultado, foram identificados peptídeos N-terminal e de SNPs em amostras proteômicas, sem aumentar substancialmente o tamanho do banco de dados. De maneira similar, Bunger et. al. (2007) desenvolveram um banco de dados de peptídeos trípticos criado a partir de informações encontradas no dbSNP do NCBI. Neste trabalho, eles adicionaram na base de dados criada somente os peptídeos trípticos mutados e os peptídeos referência correspondentes, distinguindo-se da abordagem desenvolvida por Schandorff et. al. (2007) que mantiveram a sequência referência completa. Em relação às variações, foram incorporadas somente mutações não sinônimas e indels que não geram *frameshift*. Como resultado, foram identificados 629 SNPs não sinônimos, dos quais 36 não encontrados nos bancos de dados de referência NCBI ou IPI.

Em uma perspectiva clínica, o trabalho desenvolvido por Li et. al. (2010) buscou a identificação e anotação de proteínas relacionadas à oncogênese e/ou progressão do tumor. Para isso, eles criaram um banco de dados denominado *Cancer Proteome*

Variation Database (*CanProVar*) que integra informações de variações proteicas contidas em dados públicos e que pode ser consultada por meio de uma plataforma também desenvolvida na pesquisa. Essa plataforma fornece acesso à variações conhecidas em proteínas relacionadas aos tipos de câncer e o impacto dessas variações em suas características funcionais. Nessa pesquisa, conseguiram identificar variações enriquecidas em certos domínios de proteína e que as proteínas variantes eram mais propensas a interagir umas com as outras na rede de interação de proteínas. Essa plataforma está disponível em: http://canprovar2.zhang-lab.org/.

No trabalho desenvolvido por Sheynkman et. al. (2013), criaram uma base de dados utilizando dados de RNA-seq. Para a criação da base, eles utilizaram um script disponível no Ensembl para a conversão dos SNPs para as coordenadas dos aminoácidos, gerando como saída a posição na proteína do aminoácido alterado de acordo com o RefSeq de proteínas. O cabeçalho FASTA inclui a alteração e posição de aminoácidos referente ao identificador NP do RefSeq, ligado a cada sequência contendo trocas de aminoácidos e todas estas sequências foram anexadas à proteína RefSeq e ao *common Repository of Adventitious Proteins* (cRAP) FASTA. Essa abordagem permitiu a detecção de 421 peptídeos polimórficos mapeados para 395 Polimorfismos genéticos não sinônimos. Além disso, compararam peptídeos identificados com o banco de dados do dbSNP para análise de falsos positivos e descobriram que mais de 70% dos peptídeos não constavam no dbSNP.

No trabalho descrito por Song et. al. (2014), foi desenvolvido um banco de dados associada à variação para a quantificação de variações de aminoácidos individuais. Para isso, utilizaram as sequências variantes únicas do banco de dados *humsavar* e do banco de dados MS-CanProVar para integrar ao banco de dados de proteínas canônicas UniProtKB/Swiss-Prot. O banco de dados foi construído associando as variações mais abrangentes, denominado Swiss-CanSAAVs. Para cada variação de um único aminoácido, um peptídeo tríptico independente com dois locais de clivagem perdidos ao redor do local de variação central foi extraído das sequências de proteínas, e um identificador pré-fixado com SAAV foi adotado para diferenciar da sequência proteica canônica. Como resultado, até 282 locais de variações únicas foram quantificadas nos tecidos do fígado humano, e as associações entre variações e o desenvolvimento de carcinoma hepatocelular. A redução do espaço de busca dessa abordagem torna-se um ponto positivo, embora apenas as

variações de sequência genômica conhecidas possam ser identificadas por essa pesquisa de banco de dados.

Tan et. al. (2017) utilizou Swiss-CanSAAVs e o UniProtKB/Swiss-Prot para identificar perfis de variantes de aminoácidos de subpopulações na linha celular de câncer de mama MCF-7. Ao pesquisar no banco de dados Swiss-CanSAAVs, 374 aminoácidos variantes exclusivos foram identificados no total, onde 27 são relacionados ao câncer. Além disso, 135 aminoácidos variantes únicos foram encontrados na população de células-tronco cancerosas em comparação com as células maduras luminais. Desse modo, o desenvolvimento de banco de dados customizados permite não somente um acréscimo de dados relacionados à análise proteômica, mas também permite a utilização por pesquisadores em vários contextos, especialmente na clínica. O banco de dados dbSAP apresenta um conjunto de variantes derivadas de oito bancos de dados SNP diferentes e foi usado para caracterizar mutações em vários tipos de câncer (CAO et al., 2017). Um trabalho semelhante foi feito por Alfaro et al. (2017) usando uma combinação de variantes populacionais publicamente disponíveis (dbSNP e UniProt) e variações somáticas no câncer (COSMIC), juntamente com dados genômicos e transcriptômicos específicos da amostra para examinar a variação no proteoma dentro e entre 59 linhagens de células cancerígenas.

2 OBJETIVOS

2.1 GERAL

- Desenvolver uma abordagem computacional para identificação de peptídeos polimórficos utilizando informações do NCBI/RefSeq e dbSNP que permita aumentar a confidência na identificação por MS e associar a patogenicidade.

2.2 ESPECÍFICOS

- Selecionar as variações do dbSNP mapeadas em região codificantes.
- Desenvolver programa computacional para criação do banco de dados e para classificação das variações dos peptídeos identificados.
- Desenvolver plataforma *web* para visualização e análise dos peptídeos identificados.
- Construir o pacote da aplicação dentro de um container, permitindo torná-lo portátil para qualquer usuário, caso o servidor de hospedagem da aplicação fique indisponível.
- Buscar espectros MS experimentais para aplicação da abordagem proteogenômica desenvolvida.
- Utilizar a base de dados desenvolvida em amostras de câncer para validar a abordagem desenvolvida.

3 TRABALHO PRINCIPAL

Essa seção apresenta o artigo produzido durante a pesquisa proposta neste trabalho. O artigo foi publicado no periódico IEEE Access (Qualis A1), Volume 10, páginas: 90982 - 90994, ISSN: 2169-3536, DOI: 10.1109/ACCESS.2022.3201897. Esse trabalho teve a colaboração dos pesquisadores: Dr. Patrick Terrematte, Ms. Tayná Fiúza, Dr. Vandeclécio Lira, Dr. José Eduardo Kroll, Dr. Sandro José de Souza e Dr. Gustavo Antônio de Souza (Orientador).

3.1 DBPEPVAR: UM NOVO BANCO DE DADOS DE PROTEOGENÔMICA DO CÂNCER

A Figura 13 apresenta o resumo do processo utilizado para construção da base de dados, análise e interpretação dos resultados e desenvolvimento do portal web para visualização e obtenção dos dados gerados. O resumo gráfico é uma exigência do periódico cujo trabalho foi publicado.

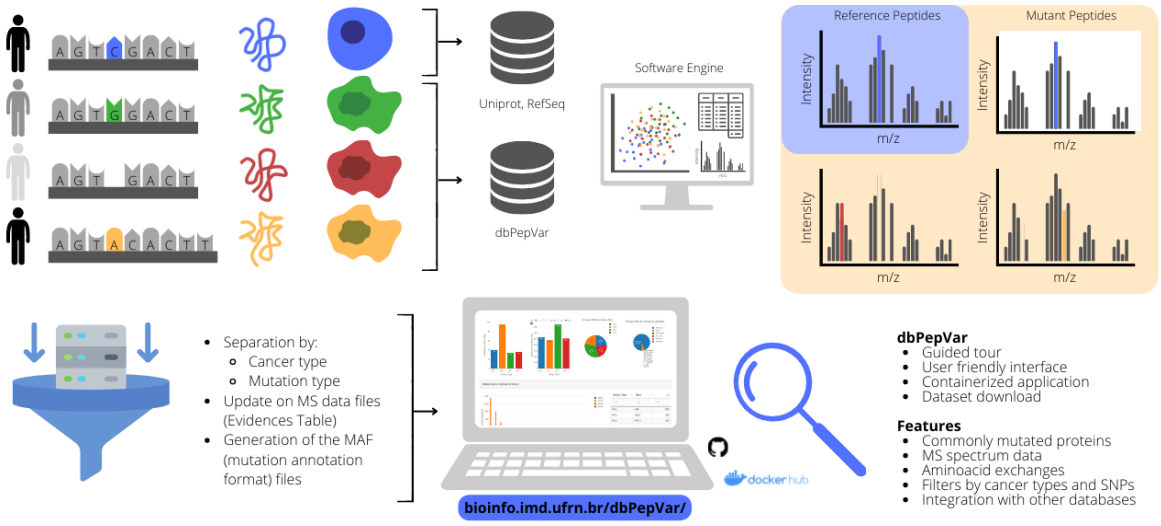


Figura 13. dbPepVar - Graphical Abstract: Genetic variation among people may generate mutant proteins, which might result in diseases such as cancer.
Mutant proteins are not present in reference databases. In this work we gathered reference databases and added information regarding mutant proteins derived from peptide mass spectrometry data of four cancer types. The processed data is available at a web portal: bioinfo.imd.ufrn.br/dbPepVar/.
Autoria: Tayná da Silva Fiúza.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 August 2022, accepted 22 August 2022, date of publication 26 August 2022, date of current version 2 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3201897

METHODS

dbPepVar: A Novel Cancer Proteogenomics Database

LUCAS MARQUES DA CUNHA[1,2], PATRICK TERREMATTE[1,3], TAYNÁ DA SILVA FIÚZA[1], VANDECLÉCIO LIRA DA SILVA[1,4], JOSÉ EDUARDO KROLL[1,5], SANDRO JOSÉ DE SOUZA[1,6], AND GUSTAVO ANTÔNIO DE SOUZA[1,7]

[1] Bioinformatics Multidisciplinary Environment (BioME), Federal University of Rio Grande do Norte (UFRN), Natal 59078-970, Brazil
[2] Academic Department of Computer Science, Federal University of Rondonia (UNIR), Porto Velho 76801-058, Brazil
[3] Metropolis Digital Institute, UFRN, Natal 59078-970, Brazil
[4] Beneficencia Portuguesa Hospital, São Paulo 01323-001, Brazil
[5] Diagnosticos da America S.A., Boa Viagem 51210-001, Brazil
[6] Brain Institute, UFRN, Natal 59078-970, Brazil
[7] Department of Biochemistry, UFRN, Natal 59078-970, Brazil

Corresponding author: Lucas Marques da Cunha (lucas.marques@unir.br)

The work of Gustavo Antônio de Souza was supported in part by the Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior under Grant 23038.004629/2014-19, and in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under Grant 406630/2016-0 and Grant 304422/2017-7. The work of Lucas Marques da Cunha, Tayná da Silva Fiúza (88887.488330/2020-00), Vandeclécio Lira da Silva, and José Eduardo Kroll was supported by the CAPES—Coordination for the Improvement of Higher Education Personnel Formation Programs. This work was supported by the Fund of the Federal University of Rio Grande do Norte.

ABSTRACT Cancers arise from the acquisition of DNA mutations, such as substitutions, deletions, amplifications, and rearrangements. Understanding the distribution and correlation of such mutations in cancer may aid the characterization of the disease and subsequent identification of biomarkers for diagnosis and treatment. The proteogenomics database (dbPepVar) created here combines genetic variation information from dbSNP with protein sequences from NCBI's RefSeq. Public mass spectrometry datasets (Ovarian, Colorectal, Breast, and Prostate) were used to perform a pan-cancer analysis, allowing the identification of unique genetic variations. As a result, 3,726 variant peptides were identified in samples from patients with ovarian cancer, 2,543 in prostate, 2,661 in breast and 2,411 in colon-rectal cancer patients. Data resulting from the proteogenomics approach employed and connected to other biological databases is now available in an intuitive and dynamic web portal where novice users can explore general aspects of the dataset in graph or table format, or dive in to filter the data with click and select options or using more advanced queries with regex. All data can be downloaded in csv or pdf format. In perspective, the web portal developed may direct studies to identify new therapeutic targets for different cancers, and one can also use our database for characterization of variants in samples of unknown genetic background, such as archived samples.

INDEX TERMS Cancer proteomics, genetic variation, proteogenomic database.

I. INTRODUCTION

Mass spectrometry (MS) based proteomics has become the primary method for comprehensive protein detection and characterization. Peptide identification is often based on challenging experimental peptide MS spectra against theoretical peptide data created from a protein sequences database such as RefSeq, Uniprot, or Gencode [1]. Those databases for protein identification do not take into consideration genetic variations in populations. This genetic variability gives individuals unique phenotypic characteristics, vulnerability to diseases, and profiles of responses to drugs. Variations in genomes might affect protein-coding sequences, producing not only a single amino acid change but also changing the reading frame, originating abnormal sequences, or removing whole portions of the protein through a premature termination codon insertion [2], [3]. Thus, peptides whose exact sequences are not found in the databases remain unidentified. Some missing sequences might have a central biological role in non-annotated protein-coding regions, specific variations of individuals, or for a specific disease mutation. So the characterization of those new proteoforms is essential for

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Salehzadeh-Yazdi.

understanding human biology [1]. To grasp full information on protein variations, it is usual to apply proteogenomics methods, a field on the intersection between genomics and proteomics.

Proteogenomics assists in the identification of new coding variants by searching the MS/MS spectra against a database of customized proteins. The protein sequences in these databases are constructed using genomic and transcriptomic information that is absent in conventional protein databases. The approach improves the annotation at the protein level, refines gene models, characterizes protein isoforms, restricts the location of the translation start and end sites, identifies splice sites, and alternative forms of splicing [4], [5], [6], [7]. In precision medicine, proteogenomics is widely used as an alternative to detect mutations in different cancers to find new biomarkers for tumors, analyze differences in levels of gene expression, compare variants to patient survival and support the development of new drugs [8], [9].

The customization of such proteogenomics databases is achieved according to the research goal. For instance, one of the widely used alternatives is the creation of a database generated from the translation of the six reading frames. However, this approach is limited because of the increase in the database search space and the fail to capture peptides present at the junction between exons [3], [6]. Other approaches develop a database incorporating mutations into populations of selected reference sequences and report the variant sequence with a special character [10]. In another case, a database of tryptic peptides was created from information found in NCBI's dbSNP [11] by adding to the database only the mutated tryptic peptides and the corresponding reference peptides, an approach distinct from [10], which maintained the complete reference sequence. Applications of such databases include the investigation of neglected tropical diseases [12], Alzheimer's disease [13], and other complex conditions, like cancer.

Cancer, as a complex disease, arises from combinations of mutations on the same cells over time [14], [15], triggered by external and endogenous factors [16], [17]. Distinct cancer types present different combinations of mutations [14], [15]. Mutations may alter the efficiency of molecules by hindering their stability and activity. Current techniques seek to identify these changes and determine the impact generated on gene products such as RNAs and proteins, which play a major role in the cellular functions of an organism's processes. In cancer, aberrant proteins stimulate initiation, progression, and response to treatment. The abundances of protein and mRNA molecules are partially correlated and determining how the flow of information culminates in proteomic changes in tumors is a major under-explored issue in cancer biology [18]. An exclusive set of alterations can define a subtype profile in a cancer type [14], [19]. The Single nucleotide polymorphism (SNPs) play a fundamental role in distinct responses to the treatment of cancer patients, and also might characterize the risk of low survival outcomes [14], [15], [20]. The presence of mutations in coding regions might affect cellular signaling pathways, as well as the levels of oncogenic and tumor suppressor proteins [20].

Researchers have been developing solutions for better integrating variant discovery into proteomics studies. The Cancer Proteome Variation Database (CanProVar) integrates public data information on protein variations, provides access to known variations in proteins related to cancer types, and evaluates the impact of these variations on their functional characteristics [21], [22]. The CanProVar provides a base rich in genetic variations related to different cancer, incorporating missense, nonsense mutations, and single-base insertions and deletions derived from specific cancer bases, such as TCGA, HPI, COSMIC, and OMIM. Although there is a great diversity of variations incorporated, the base does not include untranslated region (UTR) mutations. The database Swiss-CanSAAVs was developed using the unique variant sequences of the *Humsavar* database, which contains only missense mutations, and integrating the MS-CanProVar database with the canonical protein database UniProtKB/Swiss-Prot [23]. For each single amino acid variation, an independent tryptic peptide with two missing cleavage sites around the central variation site was extracted from the protein sequences, and an identifier prefixed with SAAV was adopted to differentiate from the canonical protein sequence. For instance, the database Swiss-CanSAAVs was used to identify profiles of amino acid variants of subpopulations in the breast cancer cell line MCF-7 [24], identifying protein sequences [23]. The dbSAP database presents a set of variants derived from eight different SNP databases and was used to characterize mutations in various types of cancer [25]. Similar work has been done by [26] using a combination of publicly available population variants (dbSNP and UniProt) and somatic variations in cancer (COSMIC), along with sample-specific genomic and transcriptomic data to examine the variation in the proteome within and across 59 cancer cell lines.

Those databases kept variant and reference sequences within a single file. Most peptide search engines use algorithms based on statistical analysis. Such an algorithm might not assign a mutated peptide correctly because of minimal differences in each score value between the variant and reference sequences [27]. Furthermore, the approaches described above are limited to some mutations, allowing only databases with missense mutations and small insertions/deletions mutations. The mutations with the most significant impact on the sequence and protein function, such as frameshifts, variation of start translation, and loss of stop codon also are not included [28], [29], [30].

In this article, we present a variant database called dbPepVar, which contains mutated peptides built from a proteogenomics perspective. The main objective of this work is to assist in the identification of genetic variants associated with cancer at the protein level by providing a ready-to-use web portal containing processed datasets. Compared to other approaches as those described above, our database reports a greater diversity of variants, including mutations that alter

the translational starting site. Using public MS data from four types of cancer, a majority of SNPs were identified, but cancer-shared mutations were also present in a lower amount.

II. MATERIALS AND METHODS

A. DATA SOURCE

To generate the dbPepVar variant database, we used the protein RefSeq data and the dbSNP, both available on the NCBI portal at (https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/) and (https://ftp.ncbi.nlm.nih.gov/snp/). To carry out identifying variants in different types of cancer, we used experimental mass spectrometry data derived from samples from four studies: ovarian cancer [31], prostate cancer [32], colorectal cancer [9], and breast cancer [33]. All MS raw data is available at the ProteomeXchange repository (https://www.proteomexchange.org).

B. DATA PREPROCESSING

Initially, the dbSNP data were preprocessed to remove redundancies, inconsistencies, and incompleteness. The mutations of Leu to Ile or Ile to Leu were discarded since both amino acids have the same molecular mass and are not distinguished by mass spectrometry. We also removed mutations leading to alternative splicing and synonymous mutations. At the end of this process, three new files were generated containing information about the SNP, indel in-frame, and frameshift mutations. From the file containing the SNPs, four new files were generated, separated by the categories of stop-loss, untranslated region Variation (UTR variation), rare SNPs (Minor Allele Frequency $<5\%$), and common SNPs (Minor Allele Frequency $\geq 5\%$) [34]. For each type of mutation, Perl scripts were implemented to create the new proteoforms according to each mutation type described. All files processed from dbSNP contain the RefSeq identifier of the protein (NP Accession), the reference amino acid, the mutated amino acid, the position of the mutation, and the SNP identifier (RefSNPs, rs).

C. CREATION OF THE PROTEOGENOMICS BASE

The first step in creating the variant database is to generate a multi-fasta file containing the proteoforms according to the dbSNP information and then extract the variant tryptic peptides. To generate the multi-fasta file, six scripts were developed in the Perl language, referring to each type of mutation: rare SNPs, common SNPs, indel in-frame, indel frameshift, stop-loss, and UTR variation. The multi-fasta files were used as input to a second script that performs the search for the tryptic peptides variant of each proteoform. The third script generates a file in the multi-fasta format, concatenating all the variant tryptic peptides for each protein. In this process, we discard variant peptides that had less than or equal to 7 amino acid residues and that were greater than or equal to 35 residues [35]. Due to post-translational processes, a protein can undergo internal cleavages, such as the removal of the initial methionine (Met), to generate a mature product of smaller size [36]. Therefore, for any N-terminal tryptic peptides, we report a duplicate with and without the initial Met. Similarly, in cases where the protein contains a C-terminal peptide with over one mutation or had more than one nonsense mutation, we generate a new entry in the fasta file for any additional peptide variation. Otherwise, the identification software could interpret them as a single peptide, as the C-terminal peptide could lack a tryptic cleavage site (Arg or Lys). For the peptides in which the mutation has removed or added a cleavage site, we verify whether the number of residues remained within the range established in our approach.

The SNPs detected from a particular protein-coding region by NGS technologies in a tumor tissue sample are possibly derived from heterogeneous cells [37]. Furthermore, non-random genetic variations tend to occur together, as haplotypes inherited from a single parent are linked on the same chromosome [37]. Therefore, peptides that showed more than one mutation were reported according to the number of possible combinations. This criterion is made only for peptides SNP-type mutations with a common allelic frequency (Minor Allele Frequency $\geq 5\%$) to reduce computational complexity. Using this criterion only for common mutations also allows us to reduce the search space and to avoid undesirable combinations with other types of mutations. The number of combinations that a peptide can present is according to the formula $2^n - 1$, where n is the number of mutations. For instance, for a peptide that has 12 mutations, the number of combinations will be $2^{12} - 1 = 4{,}095$ peptides that will be generated. This process is described in Fig. S3.

For frameshift and stop-loss mutations, we developed a script that performs the mutation in the mRNA and translates it to the respective protein. After this process, we report the variant peptide sequence that starts at the point where the mutation occurred. Missense mutation can also alter the mRNA's untranslated region (UTR variation) by replacing the base in the initial codon. In this way, the translation machinery scans up to the next initial ATG nucleotide codon triplet [38]. In this case, we report the tryptic peptide corresponding to the new start of the translation of the protein.

We also implemented scripts to extract information such as the protein identifier (NP accession), the SNP identifier (RefSNP, rs), the position of the mutation in the sequence protein, the reference tryptic peptide, and the mutated peptide. This information was useful for missed cleavage analysis and for classifying the peptides according to their mutation type.

D. MS IDENTIFICATION AND ANALYSIS OF THE IDENTIFIED PEPTIDES

The LC-MS data in RAW format was analyzed by MaxQuant (version 1.6.14), using previously described parameters [39]. The peptide identification process was carried out using a two-stage strategy, where the MS data were sought first using the human proteome Refseq and, later, were searched using the personalized dbPepVar database. After each stage, the

"evidence.txt" files are generated regarding the identification of the peptides according to the search base used. The evidence file combines all information about the identified peptide spectrum matches (PSM) and is usually the only file needed for processing the results.

Initially, we remove false positives and contaminants identifications. For data provided from the Super-SILAC quantification protocol (Breast and Prostate cancer), peptides with an Intensity value of L = 0 (i.e., only identified in the reference Super-SILAC cells) are removed, as it shows that the peptide was not detected in the non-reference sample. The next step is to analyze the quality of the peptide scores and select those that have the best value. All peptides identified in dbPepVar with a score lower than 50 were removed as they indicate low-quality identification.

We compared the two evidence files obtained from the two RefSeq and dbPepVar databases to check if there are two different sequences identified to the same MS spectrum. For this, the fields of the evidence file Raw File and MS/MS Scan Number are used to generate a unique identifier of the PSM. A Perl script was implemented to check and add to a file the peptides that were identified on both databases. In cases of a conflict, the script selects the variant peptide only when its score is 20% higher than the score of the RefSeq sequence identified for the same spectrum.

During the proteolytic digestion process, the enzyme can fail to cleave in one or more tryptic sites. Therefore missed cleavages are often considered during peptide identification. In the evidence file obtained from dbPepVar, we verify whether the peptides that showed missed cleavages were possible false positives. We considered false-positive variant tryptic peptides with missed cleavage whose location contradicted the actual position in the original protein. This occurs because the way our database is created, mutated peptides are concatenated even though they are not necessarily neighbouring peptides in the reference protein. For this analysis, we used the field Missed cleavages in the evidence table, where values greater than 0 indicate the presence of a missed cleavage. The registration file was used to check the position of the peptide in the source protein. In this filtering process, we also discard the variant peptides that were also present in the RefSeq base. Moreover, for mutations of the UTR variation type, we discard peptides that had a cleavage site before methionine, avoiding an erroneous identification of an enzymatic cleavage as a false new start of translation. We then classify the variant peptides according to the type of mutation. The information was obtained from the log files generated in the creation of the dbPepVar database. As an output, we extract two new fields to the evidence files regarding the type of mutation and the dbSNP code. To visualize the mass spectra of the identified peptides, we used a tool called Proteogenomics Viewer [40]. It is a web tool that collects the identification of peptides by mass spectrometry, indexes a sequence of genetic structure, attributes the use of the exon, and relates to isoforms of proteins. Thus, to suit the data to the tool, we generate a protein sequence base, replacing the reference sequence or peptides used by the MS for each type of cancer. In cases where different variant peptides are used for the same protein and position, a new sequence is generated for each variation.

III. RESULTS

A. BUILDING dbPepVar AND IDENTIFICATION OF VARIANT PEPTIDES BY MS

The dbSNP data was processed by selecting and categorizing types of mutations according to genomic coding region and impact on the protein sequence. Mutations where there was an uncertainty of the actual biological event, e.g.: when mutations were represented by a question-mark, were discarded. A total of 10,490,264 SNPs were selected from dbSNP, 10,417,131 with minor allele frequency (MAF) $<$ 5% and 73,133 with MAF $\geq$ 5% [34]. The other selected types of mutations were indels (194,056), stop-loss (10,753), translational start sites (24,195), and frameshifts (367,348). In the end, 11,086,616 mutations were considered. These data were used as input for the construction phase of the dbPepVar database, generating a total of 7,747,637 tryptic peptides. For comparison, this database is approximately seven times larger than the Refseq (1,174,168 tryptic peptides). Fig. 1 shows the workflow used to create the dbPepVar database. Fig. S1 shows the process used to report the variant peptides. As additional data to dbPepVar, a file was generated containing information about the mutated peptide, such as its SNP identifier, the location of the mutation in the protein, and the sequence of the reference peptide (see Fig. S2). This information allows the validation of genomic variants at the proteomic level, the location of the type of mutation that affects the peptide, the association of the variant peptide with the corresponding SNP, and the analysis of heterozygosity, through the screening of samples that show the identification of mutated peptides and variants.

Peptide identifications were carried out using publicly available MS data, by performing MS/MS searches in each dbPepVar and Refseq databases separately. The results obtained were submitted to a filtration step that resulted in the data shown in Table S1. The table shows the number of unique peptides identified for the respective types of cancer and the search database used. A removal criterion based on identification scoring was rigorously applied to peptides derived from the dbPepVar database to guarantee the reliability of the identification. The identification of the variant peptide was only considered if: i) the MS spectrum was exclusively identified in the dbPepVar database; ii) the same spectrum providing conflicting identifications in each database, the dbPepVar result must have a score value higher than 20% compared to the identification derived from Refseq. For example, the peptides sequences TEIQGIIGQIDEVSIEK (dbPepVar; NP_004864.1) and AAAAVSESWPEIEIAER (Refseq; NP_065761.1), were identified for the same MS spectrum (MS/MS Scan Number = 89008 and Raw File = 20131021_EXQ3_FaCo_SA_CL_ME180_B) in the ovary sample CL-ME180, with score values of 107.4 and 96.1,

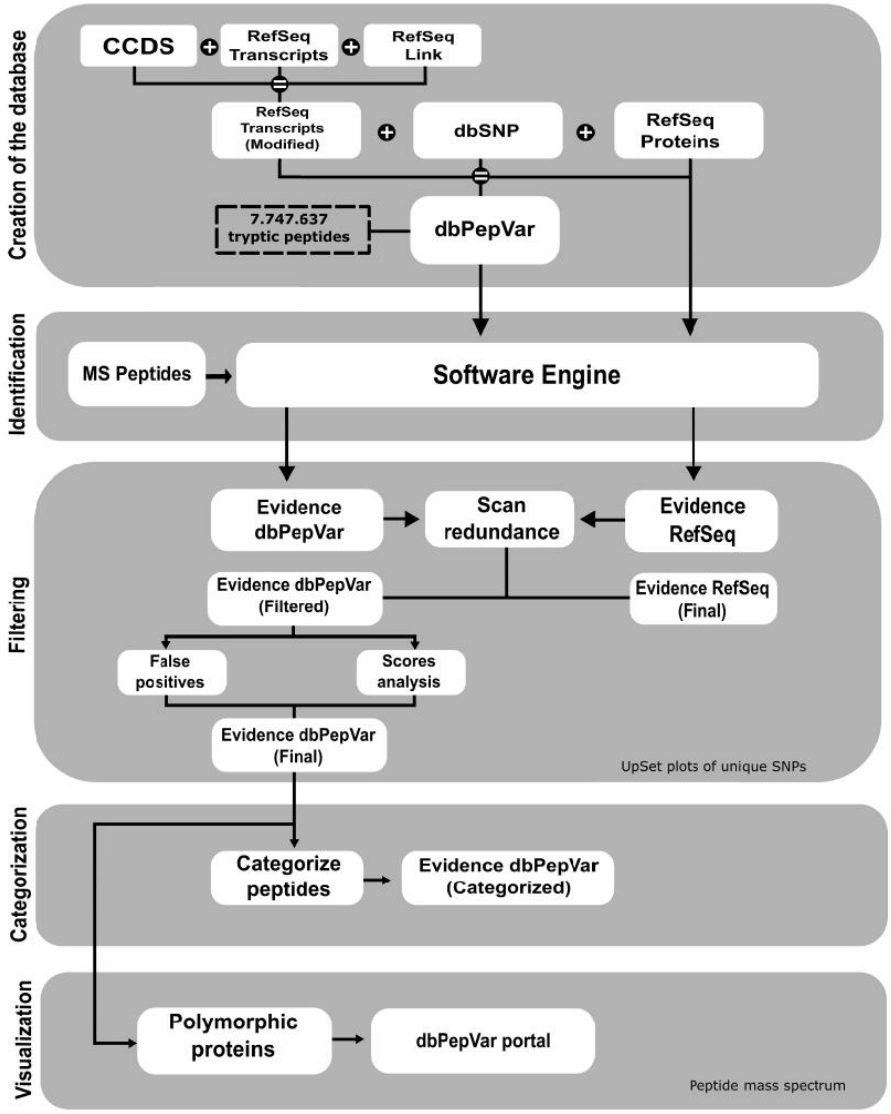


FIGURE 1. Workflow of dbPepVar creation and the analysis of the results. Database creation step: The first step consists of generating a multi-fasta transcript file containing information about the position of the beginning of the reading frame, the transcript identifier, and the reference protein. This information was obtained from CCDS (Consensus Coding Sequence) and RefSeq Link files, which contain the association between CCDS identifiers, Refseq protein, and transcribed RefSeq. The modified transcript Refseq was used to generate the sequences with frameshift and stop-loss mutations. **Peptide identification step:** The search process uses each base individually (dbPepVar and RefSeq) to identify the peptides.**Filtering step:** In this step, the identified peptides are checked in both bases (dbPepVar and RefSeq) and verified if they have the same MS spectrum. In a redundancy case, the variant peptide with a higher score was kept. The variant peptides with scores less than 50 and false positives were removed. The false positives are the variant peptides with enzymatic cleavage error whose position differs from the reference protein. **Classification step:** In this step, variant peptides are classified according to the type of mutation. **Visualization step:** In the last step, the evidence tables of each cancer analyzed are available at the portal dbPepVar, and the data is integrated with the mass spectra visualization tool, Proteogenomics Viewer.

respectively. In this case, the reference peptide was kept and the variant peptide was deleted from the analysis. The adopted criterion provides an additional level of certainty in the identification of the variants present in the samples. A range of 220,405 to 341,906 variant identified peptides were found to have conflicting MS scans with reference sequences by cancer type. After analyzing the scores, 1,429 to 9,735 variant peptides remained. To remove experimental error introduced by the concatenation of variant peptides necessary to the database construction, peptides with missed tryptic cleavage

TABLE 1. Number of unique peptides identified for the types of cancer: ovary (OvCa), prostate (PrCa), breast (BrCa), and colorectal (CrCa).

Cancer Samples	SNP	INDEL	Frameshift	Stop Loss	UTR variation
OvCa	3.460	197	34	16	19
PrCa	2.340	139	29	10	26
BrCa	2.501	117	20	6	18
CrCa	2.266	104	22	8	11

sites and whose location differed on the reference proteins were regarded as false positives and excluded. Variant peptides found on the reference database were discarded as well. All identified peptides in RefSeq and dbPepVar and their MS identification features (score, mass, mass error, and others) are available on the portal by accessing the "evidence tables" tab. When also considering peptide spectrum matches (PSMs) that were identified only in the dbPepVar search, the following number of variations were detected: 3,726 in the ovarian cancer samples, 2,543 in the prostate cancer samples, 2,661 in the breast cancer samples, and 2,411 in the colorectal cancer samples (Table S1). We estimated the number of peptides specific to the types of mutations used in the construction of dbPepVar to verify the proportion concerning the different types of cancer. As expected, most mutations identified are missense SNPs (Table 1), but there are also peptides with small in-frame indels (3-4% avg), frameshift indels (1% avg), and a few other characterized as UTR variation, stop-loss and c-terminal peptides derived from premature termination codons (average $< 0.5\%$).

After classifying and counting the identified variants, those were organized as unique or shared between samples. This last step required the use of the SNP identifier (rs, reference sequence) as a unique key to each mutation. The shared and unique counts according to the SNP identifier can be seen in Fig. 2. Fig. 2a shows that ovarian cancer has most of the identified mutations with 3,684 SNPs (horizontal bar graph), of which 2,281 were unique to the sample (vertical bar graph). From all SNPs identified, there are 365 shared by all selected colorectal, prostate, breast, and ovarian samples, as shown by the connecting dots at the bottom of Fig. 2a (sixth bar from the left). Prostate and breast cancer samples share the highest number of common SNPs, with 437 entities. Prostate and colorectal cancer samples have the least, with 81 SNPs in common. There are 248 entities shared for three types of cancers: ovarian, prostate, and breast. The prostate samples have less exclusive SNPs, but share most SNPs with other cancers. A similar pattern arises among less frequent mutations ($< 5\%$) (Fig. 2B). While through this type of analysis it is not possible to discriminate specific cancer mutations from those that were already present in the donors genomic background, comparing unique and shared SNPs in such samples might raise interesting hypotheses about the clinical condition under study.

B. THE WEB PORTAL

The data built into dbPepVar offers a wide range of potential opportunities for data mining and analysis. Our database, built using Shiny R, is available at https://bioinfo.imd.ufrn.br/dbPepVar/ and can be used by life science researchers who do not have command line experience that may benefit from a guided-tour of each section and tab of the main page. Here, we present a glimpse of the potential that dbPepVar has for the discovery of new data (Fig. 3 and 4). However, this paper does not cover the full extent of the data or all potential applications of the platform, which is available as an open resource for the researcher to use in their investigations.

The first menu ("dbPepVar") contains a summary of the data accessible through the portal (Fig. 3). The graphical displays were separated by section according to the type of data and analysis that can be performed. Broadly, the initial section reports the distribution of samples, peptide sequences, and unique polymorphisms filtered by cancer type or by variant type. The latter sections summarize different aspects of the database in graphical and table format. More specifically, dbPepVar users can view graphs of the distribution of peptides and SNPs by cancer type and mutation classification (SNPs graph only). In the second section, users can explore and visualize the count of the most mutated genes, segregated by cancer type and with a responsive table explicitly showing the displayed data. As with all graphs in the portal, Plotly tools (i.e. lasso or box select) are available and allow comparing data, filtering by cancer type and gene groups from a threshold that can be defined by counting SNPs identified per sample. The responsive table also allows to filter and visualize the number of samples that have a mutation in a specific gene according to the type of cancer. Similar analysis can be done with the graph and table provided in the following sections.

The third section of the first menu exhibits the number of SNPs per gene, which may be used to build a mutational panel for each cancer type and gene of interest. The fourth and fifth sections are dedicated to amino acid change counts by sample and by SNP, respectively. In this way, it is possible to observe, at the proteomic level, the most frequent amino acid exchanges for different cancers and SNPs, which may help understand which mutations propagate from the genome to the proteome. Two additional sections summarizing other layers of integrated information are then displayed, without tables: one with chemical property changes of amino acids sorted by cancer type, where 'Multiple' refers to samples with frame-shift mutations, and another showing the distribution of mutated genes by chromosomal location. Thus, users can interactively perform two tasks: (i) filter and visualize the most frequent changes in amino acids according to cancer type, and (ii) filter and visualize the common exchanges between chemical groups of amino acids.

The second menu ("Variants") shows the actual dataset in an interactive format, where users can perform data mining and generate insights for their research (Fig. 4a). This action can be done by selecting all or single rows with up to 27 columns that describe each mutation. The table includes links to GeneCards, NCBI protein, and dbSNP. Users can filter on any of the provided columns using plain text and

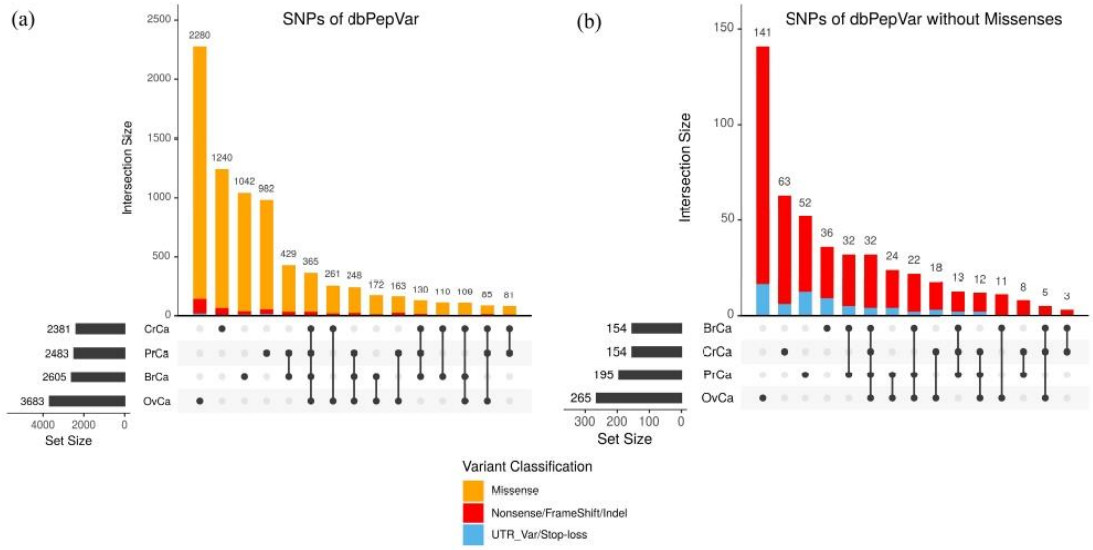


FIGURE 2. Variant distribution by sample typesand frequency of unique and shared variants from dbPepVar found in ovary (OvCa), prostate (PrCa), breast (BrCa) and colorectal (CrCa) cancer samples. The vertical bars show the number of identified variants for each sample combination, which is indicated by the vertical line and dots in the lower portion of the figure. (a) Data where missense mutations (orange) are considered. (b) Missense mutations are removed from the data, red bars indicate indel observations and blue bars indicate UTR and stop loss mutations.

regular expressions. Recovered results can be downloaded as CSV or PDF formatted files (all pages or current page only). The third menu (“Evidence Tables”) is constructed by parsing the evidence files, which combine all information about the peptides identified by MS and is normally the only information needed for processing the results (Fig. 4b). It is from the evidence file that the other results presented on the portal are generated. Each type of cancer has an evidence file that can be accessed in its respective tab (BrCa, PrCa, OvCa, CrCa). Every file contains peptide information such as its amino acid sequence, post-translational modifications, the number of enzyme missed cleavages, its mass/charge ratio, identification scores, intensity, gene and protein names where it belongs, and more. The displayed columns can be changed by selecting specific columns. By default, unique rows are displayed, but all rows may be selected. It is also possible to download filtered information in PDF or CSV format (all pages or current page only).

Next, the “Proteogenomics Viewer” menu [40] integrates genomic and proteomic data, providing a genetic view of peptides in a sliding panel with their respective Peptide Spectrum and Peptide Expression. The search is performed by the name of the gene of interest and, after selecting it, the identified variant peptide sequences and its exonic location are shown. Finally, the “Download data” menu contains the files referring to the multi-fasta containing the mutated protein sequences and the log files containing information about SNPs identifiers, proteins, the position of the peptide in the protein, and mutated peptides and reference. It is also possible to obtain a detailed description of the information in each file and its respective construction process.

Different proteogenomic approaches have developed web portals for data availability and analysis, using different criteria. Therefore, we listed the major databases for variant proteins and compared them with dbPepVar to highlight the unique features of our approach. The result of this comparison can be seen in Table 2.

In recent years, many new biological databases of mutant proteins have been developed and published. However, all published databases have distinct and particular scopes, and to our knowledge, no databases have been published reporting the variants for cancer proteogenomics data using our reverse engineering methodology, i.e. identifying genetic mutations from altered proteins. In particular, the dbPepVar uses more refined criteria to detect peptides that accurately represent the actual peptide, such as changes in cleavage sites, peptide size, and peptides with combined mutations.

IV. DISCUSSION

The characterization of genetic mutations in their protein products is a key step to understanding their role in diseases such as cancer. However, MS-based approaches do not routinely allow the identification of polymorphisms in samples of interest. In this study, a database of variant peptides (dbPepVar) to be used in proteomics was created combining information of proteins from dbSNP and RefSeq. The dbPepVar identified genetic changes at the protein level in MS samples from four different types of cancer. In proteomics, the identification of genetic variants depends on the presence of such variants in the database used during MS spectrum matches. Many publications had suggested diverse approaches to improve such identification coverage.

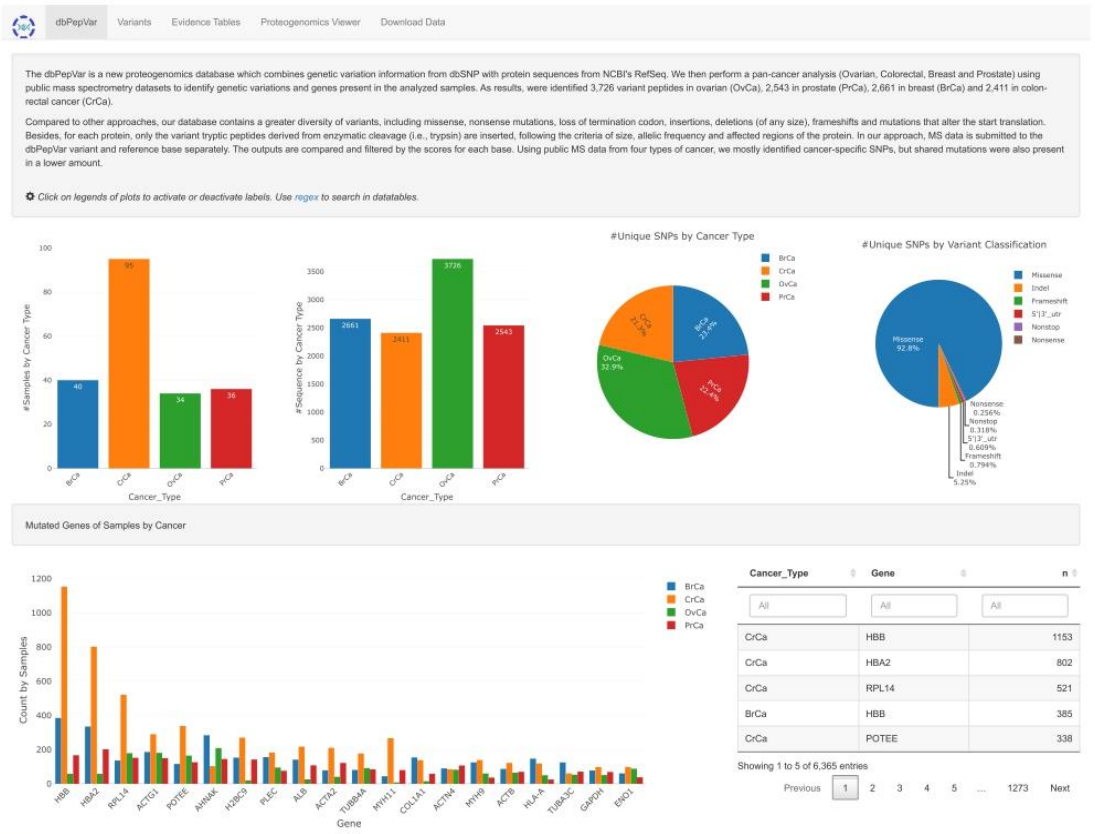


FIGURE 3. The dbPepVar portal. Main Page - View of initial page containing the summary information of the samples, peptide sequences, and unique polymorphisms.

This includes adding variant tryptic peptides concatenated to the reference sequence entry, such as observed in CanProVar [21], [22], Swiss-CanSAAVs [23] and dbSAP [23], [25], which may lead to false-positive identifications. Adding variant and reference peptides to the same fasta file can increase the probability of matching an inappropriate, but high-scoring peptide among the large number of available sequences. For instance, a variant peptide may be correctly assigned to an isobaric reference peptide, according to the spectrum, but still correspond to a different reference protein. Also, a peptide variant can be mismatched to a spectrum because the change in mass caused by a mutation coincides with the change in mass associated with a post-translational change in a different peptide. In dbPepVar, the variant peptides are incorporated in a single fasta file and the search is performed separately, allowing to distinguish between mutated and reference peptides during the identification process. In addition, a set of filters based on MS scores, removal of redundant sequences, and analysis of cleavage errors were developed to ensure that the identified peptides match the reported protein. The database built by Alfaro and coworkers [26] presents a similar approach. However, they only consider the minimum size of the peptide incorporated in its variant base (7 residues); dbSAP also considers only the minimum size of the peptide (10 residues). Mass spectrometry-based proteomics has some limitations, including the difficulty of identifying peptides that have very small or large sizes. For the peptides to be identified with greater precision, they must have a size in the range of 7-35 amino acid residues [44]. To avoid losing the identification of variants, dbPepVar also considered this interval as a parameter to determine the number of amino acid residues of the peptides present in the base. Swiss-CanSAAVs and the database proposed by Alfaro and coworkers have peptides with two missing cleavage sites, while dbPepVar has only fully tryptic peptides. Including peptides with missing cleavages in protein quantification does not produce significant differences in precision, accuracy, specificity, and sensitivity compared to the use of fully tryptic peptides [45].

The dbPepVar also performs the N-terminal methionine processing for peptides that have two or more mutations, as the sequences are generated by the concatenation of the peptides. This process is done to ensure the identification of variant peptides from proteins where the N-terminal

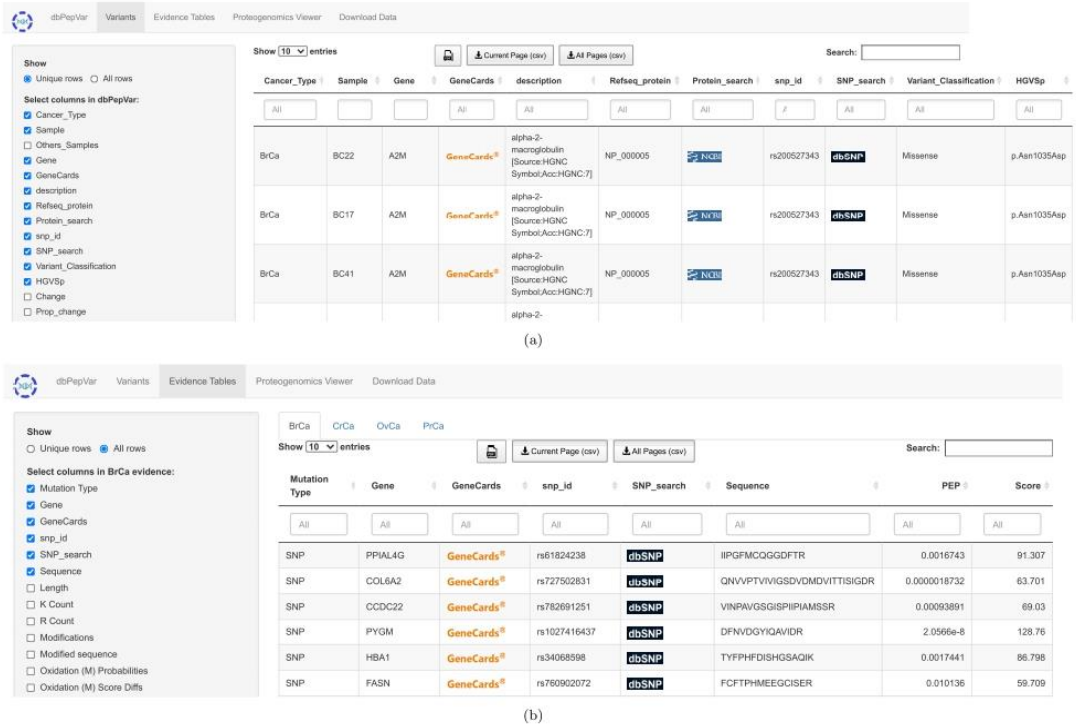


FIGURE 4. **Variant and Evidence menus. (a) The variant menu presents a table with 27 filterable columns to describe each mutation, 11 of which are shown by default. (b) The evidences menu has the actual data used to extract information available at dbPepVar. The parsed data is shown in tabulated format and can be filtered by users through 75 columns, eight of which are presented by default.**

methionine has been cleaved by co-translation by the enzyme methionine aminopeptidase [46]. The approaches presented in Table 2 assume that all digested peptides cannot have more than a single mutation. These features reduce search time, avoiding the exhaustive search for all possibilities, but naturally prevent coverage of all possible variant peptides at the same time [47]. A key advantage of dbPepVar lies in its ability to identify multiple combinatorial variants, considering all possible mutations contained in the same peptide. To avoid increasing the search space, combinations were made only for mutations with an allelic frequency greater than 5%. It is known that non-random genetic variations of a haplotype tend to occur together [37]. Therefore, the discovery of peptides with multiple mutations can be interpreted as a disease-associated haplotype, because the altered phenotypes often result from a combination of multiple factors [48]. For instance, peptides with multiple variations have been reported in ovarian and lung cancer samples [8], [37].

dbPepVar was also customized to add mutation types that affect coding regions, such as SNPs, indels, variation of the translation initiation codon, and stop-loss. SwissCanSAAVs and dbSAP have only missense mutations; CanProVar and the database developed by Alfaro *et al.* have almost the same types of mutations incorporated as dbPepVar, except for gene fusion (Alfaro *et al.*) and splice site (CanProVar) mutations. The dbPepVar differs from these approaches by the addition of the UTR-variation/start codon variation mutations. This type of mutation affects the initial methionine, generating changes in the translation start and the untranslated region of the protein [49]. Thus, in the approach proposed by dbPepVar, the peptides were generated from the search for a new alternative translation start methionine. Clinical genetic testing has identified two variants related to endometrial and breast cancers likely to affect native translational initiation on the MLH1 and BRCA2 genes [49]. Although few peptides generated by this type of mutation have been identified, this finding highlights the existence of isoforms that are being expressed by the cell in diverse cancerous environments.

Another advantage of dbPepVar is the possibility of an association between mutations identified in cancer and genetic variations in populations, which can be made from information available in public databases, such as the TCGA. In this way, this information can be used to investigate the predisposition of individuals to the disease and how this variation propagates over generations. This data expands the scope of investigation of an individual's predisposition to cancer development, given their genetic makeup. Recently, a study was conducted showing that the genotypes of patients with congenital heart disease may be responsible for the increased risk of cancer [50]. Thus, the recognition that genotypes

TABLE 2. Comparison of dbPepVar with the characteristics of databases proposed by literature.

Databases	Separate peptide files	Mutation types	Chemically guided filtering	Online Database	Data sources
CanProVar	No	Missense, insertions, and deletions of single amino acid, frameshifts, stop-loss, splice-site	Not reported	Yes	COSMIC, HPI, TCGA, OMIM, BIOMART, Sjoblom et al.[41], Greenman et al. [42], Ding et al. [43]
Swiss-CanSAAVs	No	Missense	Excluded Ile / Leu-Leu/Ile (post-processing), variant tryptic peptides with two missed cleavage sites.	Yes	UniProtKB/ Swiss-Prot, CanProVar, humsavar
dbSAP	No	Missense	Peptide length (> 9 Aa), redundant sequence excluded.	Yes	Uniprot, PMD, HPMD, MSIPI, COSMIC, dbSNP, Ensembl, CanProVar
Alfaro et al. [2017]	Yes	Missense, insertions, deletions, frameshifts, stop-loss, and fusions.	Peptide length (>6Aa), variant tryptic peptides with two missed cleavage sites, redundant sequence excluded (pre-processing)	No	Uniprot, COSMIC, dbSNP, Exome set, RNA set
dbPepVar	Yes	Missense, UTR variation, stop-loss, frameshift, insertions, deletions.	Peptide length (> 6 Aa and <36 Aa), redundant sequences excluded (post-processing), fully tryptic variant peptides, in silico cleavage of the initial methionine into peptides with two or more mutations, combinatorial analysis of mutations in peptides with more than one variant.	Yes	dbSNP, RefSeq/NCBI

influence cancer risk can promote early clinical care and interventions and further promote lifelong health in patients.

The dbPepVar portal contains all the information presented in this work but is not limited to these findings. Each researcher can use it according to their research needs. The results described can be found by navigating to the "variants" tab and selecting the fields referring to the type of search intended. The direct link with dbSNP makes it possible to verify (i) whether the mutation identified in dbPepVar is associated with other congenital or acquired diseases throughout life and (ii) the frequency of a specific variant in different populations. dbPepVar's variant menu also has a field indicating the remaining percentage of the protein sequence due to amino acid loss in a protein with a premature termination codon (PTC). For example, by selecting the field "PTC gene" and filtering by "TRUE", it is possible to obtain the variant peptide sequences that cause protein shortening, as well as information associated with the quality of identification by MS and the relationship with other databases such as GeneCards, NCBI and dbSNP. Thus, this information may be useful in investigating the impact of the mutation concerning the reduction of the protein's polypeptide chain and its relationship with some disease or alteration of its biological function. CanProVar and dbSAP present a portal with some similar characteristics to dbPepVar. CanProVar has the option to visualize the alteration of KEGG biological pathways in cancer and links that direct the user to information on genetic ontology and protein-protein interaction networks. Swiss-CanSAAVs presents in its article a link that directs to the portal, but it is inactive. dbSAP has an exclusive tab for viewing post-translational modifications and another for viewing variant and reference peptide sequences according to tissue type or cell lineage. In dbPepVar, it is also possible to visualize the PTM through the evidence tables resulting from the identification by MS. These tables also served as input for the construction of theoretical protein sequences used to visualize the information presented in the Proteogenomics Viewer. Integrating this platform to dbPepVar is unique to our approach, so the user will access the expression of the peptides and their respective mass spectra, besides being able to visualize their exonic location in the genome. All sections on the main page of dbPepVar are presented in a guided-tour, as well as the tabs described above. This feature allows users to get quickly familiarized with the portal in a first encounter. To the best of our knowledge, none of the other databases ease first-user experience with any similar approaches.

dbPepVar presents a proteomics overview for several samples from different types of cancer, allowing researchers to search for information on the set of mutations that affect specific groups of samples, analyze the most frequent mutations and changes in amino acid residues, and have direct access to information regarding each type of mutation. Approaches based on mass spectrometry gain their limitations of the technique, for example, the absence of the mutant peptide in the identification due to size. In that case, the mutant tryptic peptide may be relatively small (e.g., less than six

amino acids) and therefore difficult to reliably match the corresponding MS/MS spectrum.

V. CONCLUSION

This work presents a new proteogenomic approach for building a database of variant peptides that helps identify genetic protein variations with mass spectrometry. The dbPepVar reports missense, nonsense, frameshift, indel, stop loss, and UTR variation mutations absent in major protein databases such as RefSeq/NCBI and UNIPROT. Furthermore, the peptides available in dbPepVar were obtained upon careful consideration of the number of amino acids in the sequence, alterations in cleavage sites, and post-translational modifications, which are essential biological characteristics for the reliability of the identification. We have also developed a web portal https://bioinfo.imd.ufrn.br/dbPepVar/ for the database in which provide information on samples of four cancer types: ovarian, colon-rectal, breast, and prostate. Our portal has different filters that help the user search for information on the genetic variations identified for each type of cancer. We also integrate our data into a platform to visualize mass spectrometry-based peptide data and the corresponding genome alignments.

In the future, we aim to expand our database by adding other types of mutations and integrating them with other databases of genetic variations. Forthcoming research may investigate the relationship of the variants reported to the different types of cancer. Finally, we intend to add new features that will allow users to submit their own data for analysis visualization.

VI. THE dbPepVar WEB PORTAL

The web portal https://bioinfo.imd.ufrn.br/dbPepVar/ was implemented with the R package 'shiny' (v1.6.0). The packages required are 'tidyverse' (v1.3.1) for data preprocessing, 'vroom' (v1.5.5) for efficient reading data, 'plotly' (v4.9.4.1) for interactive visualizations, and the 'DT' (v0.19) for building interactive tables.

VII. DATA AVAILABILITY

Publicly available datasets were analyzed in this study. Code used for analyses and to produce the figures is publicly available at: https://github.com/terrematte/dbPepVar. A containerized version of the web portal is also available at GitHub, with instructions for building the image. Users may also download the container image at: https://hub.docker.com/r/fiuzatayna/dbpepvar.

VIII. ABBREVIATIONS

dbPepVar: new proteogenomics database; MS: Mass spectrometry; OvCa: ovary cancer of dbPepVar; PrCa: prostate cancer of dbPepVar; BrCa: breast cancer of dbPepVar; CrCa: colorectal cancer of dbPepVar; TCGA: The Cancer Genome Atlas;

IX. AUTHOR CONTRIBUTIONS

Lucas Marques da Cunha: conceptualization. Lucas Marques da Cunha, Patrick Terrematte: code implementations. Lucas Marques da Cunha, Patrick Terrematte, Vandeclécio Lira da Silva, José Eduardo Kroll: methodology. Lucas Marques da Cunha, Patrick Terrematte, Tayná da Silva Fiúza: composition and draft preparation. Lucas Marques da Cunha, Patrick Terrematte, Tayná da Silva Fiúza, Gustavo Antônio de Souza: composition, reviewing, and editing. Sandro José de Souza, Gustavo Antônio de Souza: supervision, funding, and infrastructure. All authors contributed to the article and approved the submitted version.

X. CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGMENT

The data processing was conducted with computational resources of the High-Performance Computing Center (NPAD) (https://npad.ufrn.br) and the Bioinformatics Multidisciplinary Environment (BioME) at the UFRN (https://bioinfo.imd.ufrn.br).

REFERENCES

[1] G. M. Sheynkman, M. R. Shortreed, A. J. Cesnik, and L. M. Smith, "Proteogenomics: Integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation," *Annu. Rev. Anal. Chem.*, vol. 9, no. 1, pp. 521–545, Jun. 2016, doi: 10.1146/annurev-anchem-071015-041722.

[2] A. Végvári, "Mutant proteogenomics," in *Proteogenomics* (Advances in Experimental Medicine and Biology), vol. 926, Sep. 2016, pp. 77–91. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-42316-6_6, doi: 10.1007/978-3-319-42316-6_6.

[3] S. Renuse, R. Chaerkady, and A. Pandey, "Proteogenomics," *Proteomics*, vol. 11, no. 4, pp. 620–630, Feb. 2011, doi: 10.1002/pmic.201000615.

[4] G. Menschaert and D. Fenyö, "Proteogenomics from a bioinformatics angle: A growing field," *Mass Spectrometry Rev.*, vol. 36, no. 5, pp. 584–599, Sep. 2017, doi: 10.1002/mas.21483.

[5] K. V. Ruggles, K. Krug, X. Wang, K. R. Clauser, J. Wang, S. H. Payne, D. Fenyö, B. Zhang, and D. R. Mani, "Methods, tools and current perspectives in proteogenomics," *Mol. Cellular Proteomics*, vol. 16, no. 6, pp. 959–981, Jun. 2017, doi: 10.1074/mcp.MR117.000024.

[6] A. I. Nesvizhskii, "Proteogenomics: Concepts, applications and computational strategies," *Nature Methods*, vol. 11, no. 11, pp. 1114–1125, Nov. 2014, doi: 10.1038/NMETH.3144.

[7] N. Castellana and V. Bafna, "Proteogenomics to discover the full coding content of genomes: A computational perspective," *J. Proteomics*, vol. 73, no. 11, pp. 2124–2135, Oct. 2010, doi: 10.1016/j.jprot.2010.06.007.

[8] S. Woo, S. W. Cha, S. Na, C. Guest, T. Liu, R. D. Smith, K. D. Rodland, S. Payne, and V. Bafna, "Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data," *Proteomics*, vol. 14, nos. 23–24, pp. 2719–2730, Dec. 2014, doi: 10.1002/pmic.201400206.

[9] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, and S. R. Davies, "Proteogenomic characterization of human colon and rectal cancer," *Nature*, vol. 513, no. 7518, pp. 382–387, Sep. 2014, doi: 10.1038/nature13438.

[10] S. Schandorff, J. V. Olsen, J. Bunkenborg, B. Blagoev, Y. Zhang, J. S. Andersen, and M. Mann, "A mass spectrometry-friendly database for cSNP identification," *Nature Methods*, vol. 4, no. 6, pp. 465–466, 2007, doi: 10.1038/nmeth0607-465.

[11] M. K. Bunger, B. J. Cargile, J. R. Sevinsky, E. Deyanova, N. A. Yates, R. C. Hendrickson, and J. L. Stephenson, "Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data," *J. Proteome Res.*, vol. 6, no. 6, pp. 2331–2340, Jun. 2007, doi: 10.1021/pr0700908.

[12] H. Pawar, S. Renuse, S. N. Khobragade, S. Chavan, G. Sathe, P. Kumar, K. N. Mahale, K. Gore, A. Kulkarni, T. Dixit, R. Raju, T. S. K. Prasad, H. C. Harsha, M. S. Patole, and A. Pandey, "Neglected tropical diseases and omics science: Proteogenomics analysis of the promastigote stage of Leishmania major Parasite," *OMICS, A J. Integrative Biol.*, vol. 18, no. 8, pp. 499–512, Aug. 2014, doi: 10.1089/omi.2013.0159.

[13] E. M. G. da Silva, L. G. C. Santos, F. S. de Oliveira, F. C. D. P. Freitas, V. D. S. C. Parreira, H. G. dos Santos, R. Tavares, P. C. Carvalho, A. G. D. C. Neves-Ferreira, A. S. Haibara, P. S. de Araujo-Souza, A. A. M. Dias, and F. Passetti, "Proteogenomics reveals orthologous alternatively spliced proteoforms in the same human and mouse brain regions with differential abundance in an Alzheimer's disease mouse model," *Cells*, vol. 10, no. 7, p. 1583, Jun. 2021, doi: 10.3390/cells10071583.

[14] W. C. Hahn and R. A. Weinberg, "Rules for making human tumor cells," *New England J. Med.*, vol. 347, no. 20, pp. 1593–1603, Nov. 2002, doi: 10.1056/nejmra021902.

[15] J. Whitworth and E. Maher, "Cancer genetics and genomics," in *Medical and Health Genomics*, 1st ed., Jun. 2016, pp. 261–284. [Online]. Available: https://www.elsevier.com/books/medical-and-health-genomics/kumar/978-0-12-420196-5, doi: 10.1016/B978-0-12-420196-5.00020-4.

[16] D. M. Eccles, N. Li, R. Handwerker, T. Maishman, E. R. Copson, L. T. Durcan, S. M. Gerty, L. Jones, D. G. Evans, L. Haywood, and I. Campbell, "Genetic testing in a cohort of young patients with HER2-amplified breast cancer," *Ann. Oncol.*, vol. 27, no. 3, pp. 467–473, Mar. 2016, doi: 10.1093/annonc/mdv592.

[17] F. Adi-Kusumo and A. Wiraya, "Mathematical modeling of the cells repair regulations in nasopharyngeal carcinoma," *Math. Biosci.*, vol. 277, pp. 108–116, Jul. 2016, doi: 10.1016/j.mbs.2016.04.007.

[18] J. A. Alfaro, A. Sinha, T. Kislinger, and P. C. Boutros, "Onco-proteogenomics: Cancer proteomics joins forces with genomics," *Nature Methods*, vol. 11, no. 11, pp. 1107–1113, Nov. 2014, doi: 10.1038/nmeth.3138.

[19] J. E. Phay and J. F. Moley, "Genetics of cancer," in *Surgery: Basic Science and Clinical Evidence*, 2nd ed. New York, NY, USA: Springer, 2008, pp. 1901–1924. [Online]. Available: https://www.amazon.com.br/Surgery-Basic-Science-Clinical-Evidence/dp/0387308008

[20] N. Deng, H. Zhou, H. Fan, and Y. Yuan, "Single nucleotide polymorphisms and cancer susceptibility," *Oncotarget*, vol. 8, no. 66, pp. 110635–110649, Dec. 2017, doi: 10.18632/oncotarget.22372.

[21] J. Li, D. T. Duncan, and B. Zhang, "CanProVar: A human cancer proteome variation database," *Hum. Mutation*, vol. 31, no. 3, pp. 219–228, Mar. 2010, doi: 10.1002/humu.21176.

[22] M. Zhang, B. Wang, J. Xu, X. Wang, L. Xie, B. Zhang, Y. Li, and J. Li, "CanProVar 2.0: An updated database of human cancer proteome variation," *J. Proteome Res.*, vol. 16, no. 2, pp. 421–432, Feb. 2017, doi: 10.1021/acs.jproteome.6b00505.

[23] C. Song, F. Wang, K. Cheng, X. Wei, Y. Bian, K. Wang, Y. Tan, H. Wang, M. Ye, and H. Zou, "Large-scale quantification of single amino-acid variations by a variation-associated database search strategy," *J. Proteome Res.*, vol. 13, no. 1, pp. 241–248, Jan. 2014, doi: 10.1021/pr400544j.

[24] Z. Tan, S. Nie, S. P. McDermott, M. S. Wicha, and D. M. Lubman, "Single amino acid variant profiles of subpopulations in the MCF-7 breast cancer cell line," *J. Proteome Res.*, vol. 16, no. 2, pp. 842–851, Feb. 2017, doi: 10.1021/acs.jproteome.6b00824.

[25] R. Cao, Y. Shi, S. Chen, Y. Ma, J. Chen, J. Yang, G. Chen, and T. Shi, "DbSAP: Single amino-acid polymorphism database for protein variation detection," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D827–D832, Jan. 2017, doi: 10.1093/nar/gkw1096.

[26] J. A. Alfaro, A. Ignatchenko, V. Ignatchenko, A. Sinha, P. C. Boutros, and T. Kislinger, "Detecting protein variants by mass spectrometry: A comprehensive study in cancer cell-lines," *Genome Med.*, vol. 9, no. 1, p. 62, Dec. 2017, doi: 10.1186/s13073-017-0454-9.

[27] C. S. Lane, "Mass spectrometry-based proteomics in the life sciences," *CMLS Cellular Mol. Life Sci.*, vol. 62, nos. 7–8, pp. 848–869, Apr. 2005, doi: 10.1007/s00018-005-5006-6.

[28] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, "Predicting the functional effect of amino acid substitutions and indels," *PLoS ONE*, vol. 7, no. 10, Oct. 2012, Art. no. e46688, doi: 10.1371/journal.pone.0046688.

[29] J. Hu and P. C. Ng, "Predicting the effects of frameshifting indels," *Genome Biol.*, vol. 13, no. 2, p. R9, 2012, doi: 10.1186/gb-2012-13-2-r9.

[30] Z. Bánfai, K. Hadzsiev, E. Pál, K. Komlósi, M. Melegh, L. Balikó, and B. Melegh, "Correction to: Novel phenotypic variant in the MYH7 spectrum due to a stop-loss mutation in the C-terminal region: A case report," *BMC Med. Genet.*, vol. 18, no. 1, p. 150, Dec. 2017, doi: 10.1186/s12881-017-0510-8.

[31] F. Coscia, K. M. Watters, M. Curtis, M. A. Eckert, C. Y. Chiang, S. Tyanova, A. Montag, R. R. Lastra, E. Lengyel, and M. Mann, "Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status," *Nature Commun.*, vol. 7, no. 1, p. 12645, Nov. 2016, doi: 10.1038/ncomms12645.

[32] D. Iglesias-Gato, P. Wikström, S. Tyanova, C. Lavallee, E. Thysell, J. Carlsson, C. Hägglöf, J. Cox, O. Andrén, P. Stattin, L. Egevad, A. Widmark, A. Bjartell, C. C. Collins, A. Bergh, T. Geiger, M. Mann, and A. Flores-Morales, "The proteome of primary prostate cancer," *Eur. Urol.*, vol. 69, no. 5, pp. 942–952, May 2016, doi: 10.1016/j.eururo.2015.10.053.

[33] S. Tyanova, R. Albrechtsen, P. Kronqvist, J. Cox, M. Mann, and T. Geiger, "Proteomic maps of breast cancer subtypes," *Nature Commun.*, vol. 7, no. 1, p. 10259, Apr. 2016, doi: 10.1038/ncomms10259.

[34] D. M. Altshuler *et al.*, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, no. 7311, pp. 52–58, 2010, doi: 10.1038/nature09298.

[35] L. D. Fricker, "Limitations of mass spectrometry-based peptidomic approaches," *J. Amer. Soc. Mass Spectrometry*, vol. 26, no. 12, pp. 1981–1991, Dec. 2015, doi: 10.1007/s13361-015-1231-x.

[36] T. Strachan and A. Read, *Genética Molecular Humana*. Porto Alegre, Brazil: Artmed, 2013. [Online]. Available: http://cnpj.info/Artmed-Editora-Ltda-Artmed

[37] W.-K. Choong, J.-H. Wang, and T.-Y. Sung, "MinProtMaxVP: Generating a minimized number of protein variant sequences containing all possible variant peptides for proteogenomic analysis," *J. Proteomics*, vol. 223, Jul. 2020, Art. no. 103819, doi: 10.1016/j.jprot.2020.103819.

[38] A. G. Hinnebusch, I. P. Ivanov, and N. Sonenberg, "Translational control by 5'-untranslated regions of eukaryotic mRNAs," *Science*, vol. 352, no. 6292, pp. 1413–1416, Jun. 2016, doi: 10.1126/science.aad9868.

[39] K. C. T. Machado, S. Fortuin, G. G. Tomazella, A. F. Fonseca, R. M. Warren, H. G. Wiker, S. J. de Souza, and G. A. de Souza, "On the impact of the pangenome and annotation discrepancies while building protein sequence databases for bacteria proteogenomics," *Frontiers Microbiol.*, vol. 10, p. 1410, Jun. 2019, doi: 10.3389/fmicb.2019.01410.

[40] J. E. Kroll, V. L. da Silva, S. J. de Souza, and G. A. de Souza, "A tool for integrating genetic and mass spectrometry-based peptide data: Proteogenomics viewer: PV: A genome browser-like tool, which includes MS data visualization and peptide identification parameters," *BioEssays*, vol. 39, no. 7, Jul. 2017, Art. no. 1700015, doi: 10.1002/bies.201700015.

[41] G. Parmigiani, J. Lin, S. M. Boca, T. Sjöblom, S. Jones, L. D. Wood, D. W. Parsons, T. Barber, P. Buckhaults, S. D. Markowitz, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, and V. E. Velculescu, "Response to comments on 'The consensus coding sequences of human breast and colorectal cancers,'" *Science*, vol. 317, no. 5844, p. 1500, Sep. 2007, doi: 10.1126/science.1138773.

[42] C. Greenman, P. Stephens, and M. R. Stratton, "Patterns of somatic mutation in human cancer genomes," *Nature*, vol. 446, no. 7132, pp. 153–158, 2007, doi: 10.1038/nature05610.

[43] L. Ding, G. Gets, and R. K. Wilson, "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, no. 7216, pp. 1069–1075, Oct. 2008, doi: 10.1038/nature07423.

[44] D. L. Swaney, C. D. Wenger, and J. J. Coon, "Value of using multiple proteases for large-scale mass spectrometry-based proteomics," *J. Proteome Res.*, vol. 9, no. 3, pp. 1323–1329, Mar. 2010, doi: 10.1021/pr900863u.

[45] C. Chiva, M. Ortega, and E. Sabidó, "Influence of the digestion technique, protease, and missed cleavage peptides in protein quantitation," *J. Proteome Res.*, vol. 13, no. 9, pp. 3979–3986, Sep. 2014, doi: 10.1021/pr500294d.

[46] P. T. Wingfield, "N-terminal methionine processing," *Current Protocols Protein Sci.*, vol. 88, no. 1, pp. 6–14, Apr. 2017, doi: 10.1002/cpps.29.

[47] S. Choi and E. Paek, "MutCombinator: Identification of mutated peptides allowing combinatorial mutations using nucleotide-based graph search," *Bioinformatics*, vol. 36, no. 1, pp. I203–I209, Jul. 2020, doi: 10.1093/BIOINFORMATICS/BTAA504.

[48] S. Na, N. Bandeira, and E. Paek, "Fast multi-blind modification search through tandem mass spectrometry," *Mol. Cellular Proteomics*, vol. 11, no. 4, Apr. 2012, Art. no. M111.010199, doi: 10.1074/mcp.M111.010199.

[49] M. T. Parsons, P. J. Whiley, J. Beesley, M. Drost, N. de Wind, B. A. Thompson, L. Marquart, J. L. Hopper, M. A. Jenkins, M. A. Brown, K. Tucker, L. Warwick, D. D. Buchanan, and A. B. Spurdle, "Consequences of germline variation disrupting the constitutional translational initiation codon start sites of MLH1 and BRCA2: Use of potential alternative start sites and implications for predicting variant pathogenicity," *Mol. Carcinogenesis*, vol. 54, no. 7, pp. 513–522, Jul. 2015, doi: 10.1002/mc.22116.

[50] C. E. Seidman, A. Shimamura, P. E. Newburger, A. R. Opotowsky, D. Quiat, A. C. Pereira, S. C. Jin, M. Gurvitz, M. Brueckner, W. K. Chung, and Y. Shen, "Association of damaging variants in genes with increased cancer risk among patients with congenital heart disease," *JAMA Cardiol.*, vol. 6, no. 4, pp. 457–462, 2021, doi: 10.1001/jamacardio.2020.4947.

LUCAS MARQUES DA CUNHA received the B.S. degree in information systems from the University Center of Patos (UNIFIP), Paraíba, Brazil, and the M.S. degree in computer science from the Federal University of Paraiba (UFPB), Paraíba, in 2016. He is currently pursuing the Ph.D. degree in bioinformatics with the Federal University of Rio Grande do Norte (UFRN), Brazil.

He is also working as an Assistant Professor with the Faculty of Computer Science, Federal University of Rondônia (UNIR), Porto Velho, Rondônia. He is an expert on biologicals database development and forensic analysis on digital images. He developed a method for detecting tampering in digital images for different compression ratios. He worked in remote learning at the Federal University of Paraíba and supervised several researches on technologies applied to education. He works with teaching programming for mobile devices for android using the Java language. He has over six years of research and teaching experience. His research interests include artificial intelligence, explainable machine learning, bioinformatics, proteogenomics, and computational modeling.

PATRICK TERREMATTE received the B.S. degree in systems analysis from the Federal Institute of Education, Science and Technology of Rio Grande do Norte (IFRN), Rio Grande do Norte, Brazil, in 2011, the M.S. degree in systems and computation from the Federal University of Rio Grande do Norte (UFRN), Rio Grande do Norte, in 2013, and the Ph.D. degree in bioinformatics from the Bioinformatics Multidisciplinary Environment (BioME), UFRN, in 2022.

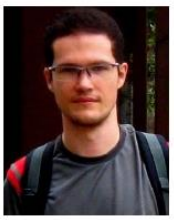
From 2016 to 2022, he was an Assistant Professor of computing engineering at the Federal Rural University of Semi-arid, Department of Engineering and Technology (DETEC), Brazil. He is currently an Assistant Professor at the Metropolis Digital Institute (IMD), UFRN. He has published on survival prediction for clear cell renal cell carcinoma. His research interests include artificial intelligence, explainable machine learning, information theory, feature selection, bioinformatics, survival analysis, systems biology, computational modeling, and logic.

TAYNÁ DA SILVA FIÚZA received the B.S. degree in biotechnology from the Federal University of Ceara (UFC), Ceará, Brazil, in 2017, and the M.S. degree in bioinformatics from the Federal University of Rio Grande do Norte (UFRN), Rio Grande do Norte, Brazil, in 2019, where she is currently pursuing the Ph.D. degree in bioinformatics.

She worked with isolating proteins and carbohydrates from marine algae, evaluating their biological activities in animal models, animal cell culture and transfection. Her research interest includes the development of computational approaches for the identification of potential vaccine candidates in the genomes of pathogenic microorganisms and applying immunoinformatics tools for neglected diseases.

VANDECLÉCIO LIRA DA SILVA received the B.S. degree in computer science from the State University of Rio Grande do Norte (UERN), Rio Grande do Norte, Brazil, the M.S. degree in genetics from the Federal University of Pará (UFPA), Pará, Brazil, and the Ph.D. degree in bioinformatics from the Federal University of Rio Grande do Norte (UFRN), Rio Grande do Norte.

He was a Bioinformatics Specialist at Duna Bioinformatics (2019–2020) and a Postdoctoral Researcher at the Laboratory of Human Genetic Diversity (LDGH), Federal University Minas Gerais (2020–2022). He is currently a Postdoctoral Researcher at the Beneficência Portuguesa Hospital, Sao Paulo, Brazil. He has experience in computer science, focusing on databases. He has experience with cloud computing and AWS. His research interests include piRNAs, gastric tissues, microarray, and drosha.



JOSÉ EDUARDO KROLL received the B.S. degree in pharmacy and the M.S. degree in biotechnology from the University of Mogi das Cruzes, Sao Paulo, Brazil, in 2006, and the Ph.D. degree in bioinformatics from the Federal University of Sao Paulo, Sao Paulo.

He was a Senior Bioinformatician at the Institute for the Heart (2020–2021), a Chief Technology Officer at Duna Bioinformatics (2019–2020), and a Postdoctoral Researcher at the Brain Institute (2014–2018). He currently works as a Bioinformatics Specialist at a Hospital & Health Care Company (DASA), Sao Paulo. His research interests include diagnostics, genomic profiling for disease risk, proteogenomics, and cancer bioinformatics.

SANDRO JOSÉ DE SOUZA received the degree in biology from the Federal University of Parana, Parana, Brazil, in 1989, and the Ph.D. degree in biochemistry from the University of Sao Paulo, Sao Paulo, Brazil, in 1993.

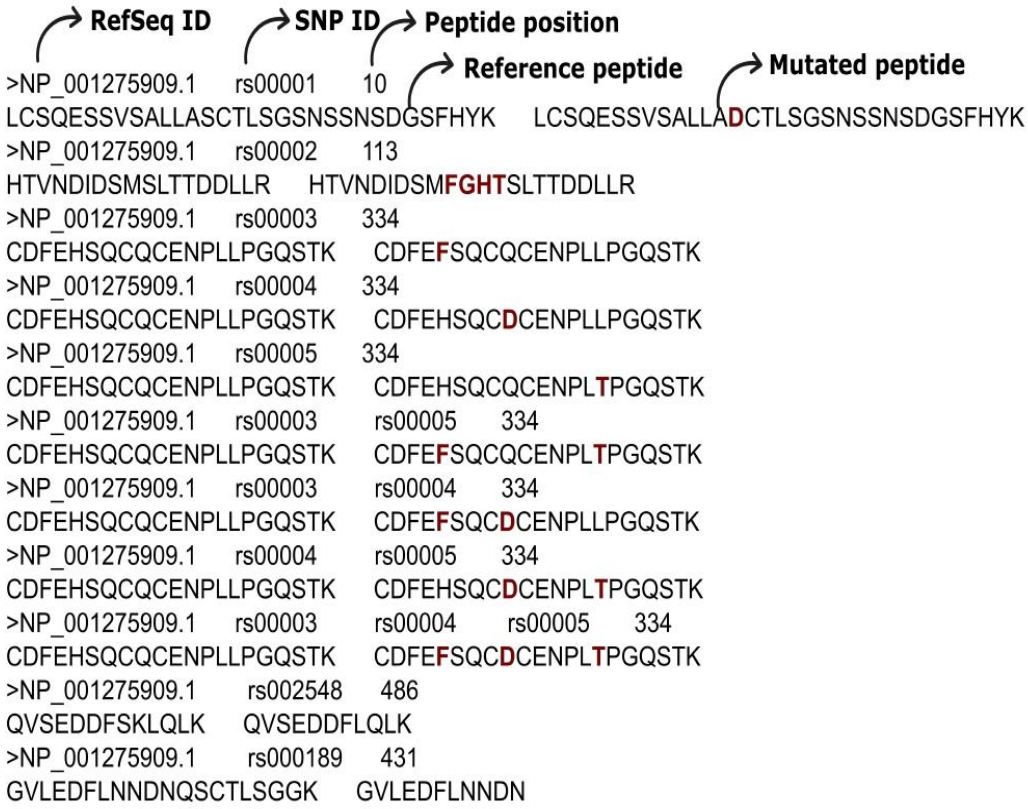
From 1995 to 1998, he was a Pew Latin American Fellow at Harvard University. He was one of the pioneers of genomics and bioinformatics in Brazil. He was an Associate Member of the Ludwig Institute for Cancer Research, from 1999 to 2012. He was elected by the World Economic Forum as a Young Global Leader, in 2009. He was a Tinker Visiting Professor at the University of Chicago, in 2011. He is currently a Full Professor at the Brain Institute, UFRN. He is also a Researcher 1B at the Brazilian National Council for Scientific and Technological Development (CNPq).

GUSTAVO ANTÔNIO DE SOUZA received the B.S. degree in biology from the Federal University of Parana, Parana, Brazil, in 1999, and the M.S. and Ph.D. degrees in molecular and cell biology from the University of Sao Paulo, Sao Paulo, Brazil, in 2004. From 2005 to 2007, he was a Postdoctoral Collaborator at the Max-Planck-Institut für Biochemie, Munich, Germany. From 2010 to 2016, he worked at the Oslo University Hospital, Oslo, Norway. He is currently affiliated with the Department of Biochemistry, Federal University of Rio Grande do Norte, Rio Grande do Norte, Brazil. His research interest includes developing computational approaches for the analysis of proteomic data from diverse organisms.

1 Supplementary Material

1.1 Supplementary Figures

Supplementary Figure 1: **The dbPepVar construction process.** (A) Initially, the reference protein is mutated according to dbSNP information. The mutated peptides are then located on the generated protein. (B) A list containing the mutated peptides for each protein present in RefSeq is generated. (C) Final fasta file is generated by concatenating the mutated peptides of each protein, generating a new theoretical sequence.



Supplementary Figure 2: **The dbPepVar provides a log file containing information about mutated peptides.** The header fields are the protein identifier (RefSeq), the SNP identifier, and the position of the peptide in the reference protein. A tab delimits the fields. Each entry has the sequence of reference and the mutated peptide. Each type of mutation is in separate files, and the missense and nonsense mutations are available in the Minor Allele Frequency (MAF) files.

1.2 Supplementary Table

Supplementary Table 1: Total peptides identified in the RefSeq and dbPepVar bases.

Cancer Samples	Peptides of RefSeq	Peptides of dbPepVar
OvCa	117.214	3.726
PrCa	104.328	2.543
BrCa	113.699	2.661
CrCa	7.987	2.411

4 APLICAÇÃO DO DBPEPVAR EM AMOSTRAS DE CÂNCER

Esta seção descreve a aplicação do banco de dados desenvolvido (dbPepVar) em um conjunto de amostras de câncer. Para isso, foram utilizados dados de MS de câncer de ovário, próstata, colorretal e mama. O objetivo é demonstrar como os pesquisadores podem utilizar a abordagem desenvolvida para obter dados sobre novos mecanismos biológicos. Para este fim, o estudo dos dados foi direcionado para o mecanismo de degradação de RNA mensageiro com códon de parada prematura (PTC).

4.1 MATERIAIS E MÉTODOS

O processo de construção do banco de dados utilizado para identificação das variações genéticas nas amostras de câncer está descrito na seção 3 deste trabalho. Assim, será apresentado apenas as informações de obtenção dos dados para validação da abordagem computacional proposta.

4.1.1 Fonte de dados

Para realizar a identificação de variantes em diferentes tipos de câncer, usamos dados experimentais de espectrometria de massa derivados de amostras de quatro estudos: câncer de ovário (OvCa) (COSCIA et al., 2016), câncer de próstata (PrCa) (IGLESIAS-GATO et al., 2016), câncer colorretal (CrCa) (ZHANG et al., 2014), e câncer de mama (BrCa) (TYANOVA et al., 2016). Todos os dados brutos do MS estão disponíveis no repositório ProteomeXchange (https://www.proteomexchange.org).

4.1.2 Seleção de proteoformas com um códon de terminação prematura (PTC)

Filtramos peptídeos identificados na posição C-terminal da proteína derivada de mutação nonsense, indel *in-frame* e *frameshift*. Para cada proteoforma, um valor de proporção de aminoácidos (APV) é calculado como o tamanho proporcional da mutação para a região C-terminal na proteína de referência, de acordo com a fórmula:

$$APV = \frac{SVP}{SRP} * 100$$

Onde SVP é o tamanho da proteína variante e SRP é o tamanho da proteína de referência. Selecionamos proteoformas com valor APV < 95% como critério para PTC em mutações indel *frameshift* e *inframe*.

Para amostras com proteoformas encurtadas, analisamos a presença de mutações deletérias em proteínas pertencentes à maquinaria NMD. Utilizou-se a ferramenta PROVEAN (*Protein Variation Effect Analyzer*) (CHOI; CHAN, 2015) para prever o impacto (neutro ou deletério) da mutação na função biológica de uma proteína.

4.1.3 Análise de Enriquecimento

A análise de enriquecimento foi realizada pelos pacotes R 'clusterProfiler' (v3.16.1) (YU et al., 2012) e 'ReactomePA' (v1.32.0) (YU; HE, 2016) para desenhar os gráficos de pontos e o mapa de calor da rede gene-conceito. Comparamos a lista de todos os genes mutados (n=2421) em amostras com PTC versus a lista de todos os genes mutados em amostras sem PTC (n=1039). Usamos a função ‘enrichPathway’ para obter os termos enriquecidos, com um valor de p abaixo de 0,05 para ser estatisticamente significativo e aplicamos a correção de Benjamin-Hochberg para reduzir o falso positivo.

4.2 ANÁLISE PAN-CÂNCER DE MUTAÇÕES NA MAQUINARIA DE *NONSENSE MEDIATED DECAY* (NMD) E PTC

Diferentes tipos de câncer podem ter um conjunto de genes mutantes em comum. Por exemplo, o risco de câncer de próstata é aumentado em portadores de mutação BRCA1 e BRCA2 com história familiar de câncer de mama (CUCCHIARA et al., 2018). Considerando essas informações, foi construído um painel mutacional para analisar os genes mais afetados por mutações e observar sua frequência de mutação em diferentes tipos de câncer. Essa abordagem permite a visualização rápida de genes mutados comumente compartilhados entre as amostras e pode revelar padrões, seja um único gene mutado na maioria das amostras ou um conjunto de genes biologicamente relacionados entre os mais mutados.

Desse modo, um gráfico do tipo Oncoplot foi empregado para visualizar um painel mutacional das amostras para genes específicos (Figura 14). As colunas representam amostras e as linhas representam os 20 principais genes mais mutantes encontrados em todos os quatro tipos de câncer analisados. As amostras foram agrupadas de acordo com o tipo de câncer. O objetivo é verificar no conjunto de dados a frequência mutacional por gene, amostra, tipo de mutação e como essa distribuição

afeta cada tipo de câncer. Em relação à frequência mutacional, observou-se que os genes ACTG1, POTEE e ECHS1 aparecem frequentemente mutados nas amostras para os quatro tipos de câncer, com valores variando entre 98% - 99% (gráfico de barras horizontal - Figura 14). Os genes ACTG1, HBB, SPTAN1 e H2BC9 apresentam vários tipos de mutação (cor preta) em várias amostras. Os genes RPL14 e SPTAN1 têm uma grande proporção de mutações INDEL in-frame (cor azul). A mutação de variação UTR aparece em menor grau no gene H2BC9 (cor roxa). As amostras com maior número de mutações são aquelas de câncer de mama, ovário e próstata; são amostras TU-LOV (728 SNPs), CL-IOSE397 (658), TU-ROV (598); BC07 (577), BC12 (563 SNPs), BC33 (555), PC09 (657), PC03 (577) e PC24 (558). Essas informações também podem ser encontradas no portal na “aba variantes”: https://bioinfo.imd.ufrn.br/dbPepVar/.

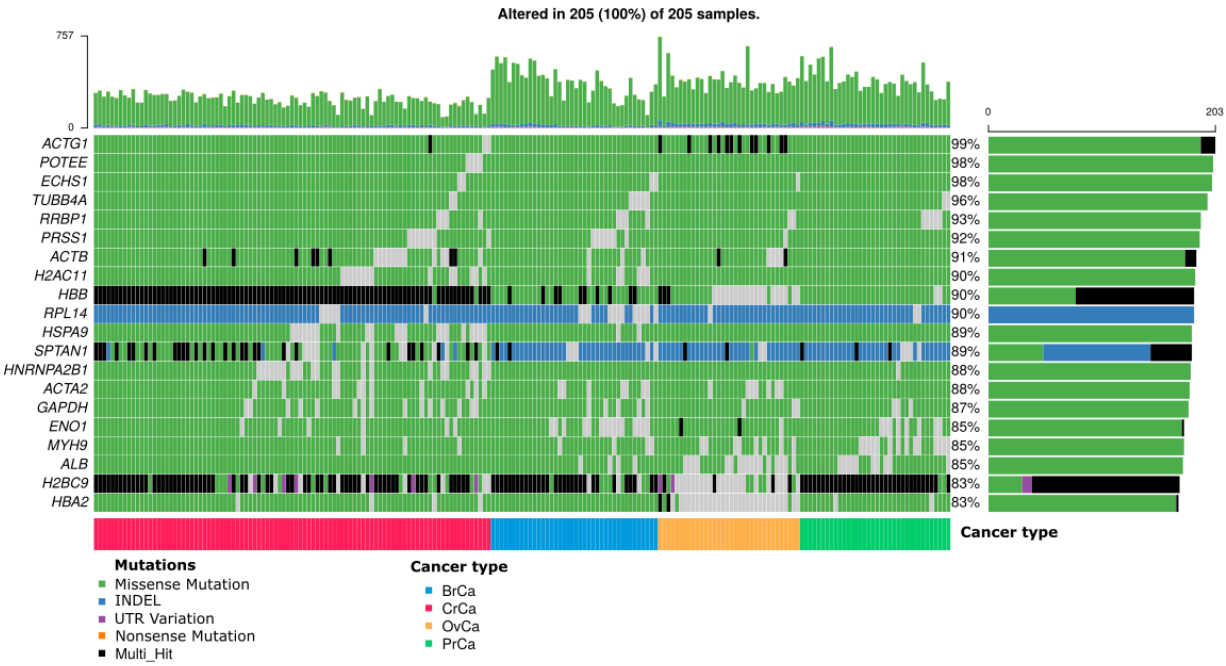


Figura 14. Análise oncoplot dos 20 principais genes mutados encontrados em amostras de câncer. O histograma no topo representa a frequência de mutações para todas as 205 amostras. No lado direito, a porcentagem de amostras afetadas e seu tipo de mutação para cada gene. A área cinza representa uma amostra sem mutação no respectivo gene. A anotação abaixo exibe o tipo de câncer para cada amostra.

Curiosamente, as mutações que resultaram em PTCs foram frequentemente observadas em 3 dos 20 principais genes da Figura 2. Essa característica de PTC pode estar ligada a três genes: ACTG1, ACTB e HBB. O gene ACTG1 aparece em 99% das amostras com mutações PTC missense e indel-generated. Os genes ACTB e HBB apresentam mutações missense e nonsense em 91% e 89% das amostras,

respectivamente (gráfico de barras horizontal - Figura 14). O portal “aba variantes” foi utilizado para confirmar a presença de mutações que geram PTC. Investigamos ainda mais as amostras para mutações que poderiam levar a proteínas aberrantes com um códon de terminação prematuro originado por mutações *nonsense*, *frameshift* e indel.

Para verificar se as mutações PTC eram frequentes em diferentes tipos de câncer, apenas as mutações PTC causadas por mutações *frameshift*, indel ou *nonsense* foram examinadas. Esses casos estão presentes em 171 das 205 amostras, conforme mostrado na Figura 15. Para os 20 principais genes mais mutados, 162 amostras tiveram mutações em pelo menos um deles detectados por MS. Comparando a anotação de frequência de mutações e o tipo de câncer na parte inferior da Figura 15, percebe-se que o câncer colorretal possui as amostras mais frequentemente mutadas. Os genes HBB e LDHA têm mais mutações PTCs por amostras, em 49% e 25%, respectivamente e, com mutações frameshift no gene LDHA nos quatro tipos de câncer. Os genes HBB, SATL1, DES e SIGLEC1 possuem mutações nonsense e estão presentes apenas em pacientes com câncer colorretal. Para este tipo de câncer, foram identificadas mutações de frameshift para os genes CFAP65 (frequência de 15%), MTCH1 (frequência de 4%), AP3B1 (frequência de 4%), PITRM1 (frequência de 3%) e DHFR2 (frequência de 2%). O gene ACTB possui mutações *nonsense* identificadas nos tipos de câncer colorretal e ovariano. No câncer de ovário, amostras com mutações nonsense nos genes HSP90AA1 e H2AC20 apresentaram frequência mutacional de 8% e 4%, respectivamente. Os genes SPRR2F e GRK7 têm mutações do tipo frameshift e foram identificados exclusivamente nos tipos de câncer de mama e próstata. Nesses dois tipos de câncer, mutações do tipo indel no gene PDCD6 também foram identificadas com uma frequência mutacional de 5%. O gene ACTG1 foi identificado com mutações do tipo indel em amostras de câncer de ovário e colorretal e uma frequência mutacional de 8%. O gene RPL29 também tem a mesma mutação com uma frequência mutacional de 11% e foi identificado em todos os quatro tipos de câncer.

Os dados apresentados mostram evidências de peptídeos que mostram a expressão de proteínas encurtadas em diferentes amostras de câncer. Em alguns casos, é possível notar uma grande predominância desse traço em um gene específico e câncer, como o HBB no câncer colorretal. Essa evidência nos leva a acreditar que as vias responsáveis pela NMD estão alteradas nestas amostras e que

elas também podem ter sofrido mutações que poderiam levar a uma redução em sua atividade.

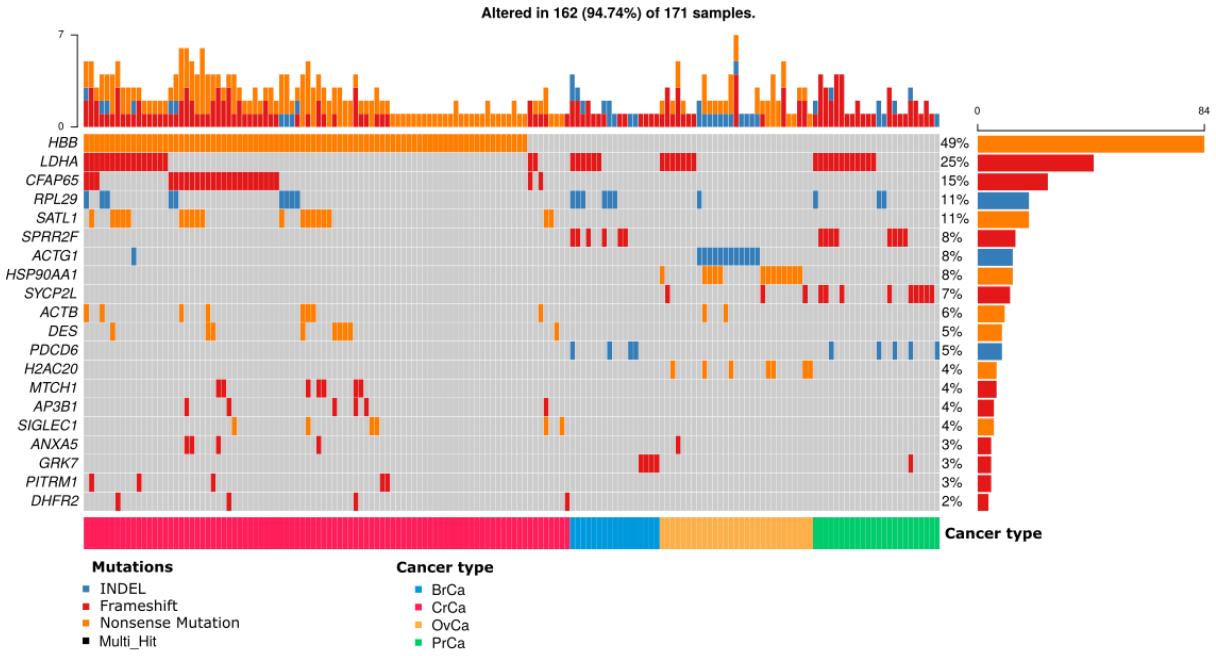


Figura 15. Análise oncoplot dos 20 principais genes mais mutados identificados por MS contendo um PTC.
Todos os tipos de câncer têm peptídeos que indicam ocorrências de PTC e todas as 205 amostras têm genes alterados nesses tipos de mutação. O gene HBB é mais frequentemente mutado no câncer colorretal com mutações indel e *nonsense*. O gene LDHA é mais frequentemente mutado com mutação *frameshift*. As mutações indel e frameshift são comuns em amostras de câncer de ovário e próstata.

4.3 ANÁLISE DE ENRIQUECIMENTO E POSSÍVEL IMPACTO FUNCIONAL DE MUTAÇÕES

Devido à alta frequência de evidências peptídicas derivadas de mutações PTC, uma análise de enriquecimento foi realizada para comparar termos enriquecidos de dois conjuntos de genes: todos os genes mutados (n=2421) em amostras com mutações PTC versus todos os genes mutados (n=1039) de amostras sem mutações PTC. Em particular, o termo NMD é enriquecido (p-ajustado < 0,01) apenas para o conjunto de amostras com PTC. Além disso, três processos biológicos de NMD foram encontrados - a própria via NMD, bem como a NMD aprimorada pelo *Exon Junction Complex* e as vias NMD independentes do *Exon Junction Complex*. Esse mesmo conjunto foi usado para buscar enriquecimento em um subgrupo de vias relacionadas à tradução de proteínas, que compartilha genes com a via NMD. O processamento de

rRNA no núcleo e na via do citosol tem o ajuste de p mais significativo (< 0,01) - seguido pela via NMD (ver Figura 16A). A relação entre genes e termos é mostrada na Figura 16B. Além disso, as 60 principais vias enriquecidas de 400 genes mais frequentemente mutados em amostras com PTC estão relacionadas a processos como sinalização por Rho GTPases, respostas ao estresse, ativação da matriz extracelular, metabolismo de carboidratos ou processos relacionados ao sistema imunológico, como degranulação de neutrófilos, degranulação plaquetária, "doença infecciosa" (Figura 2 suplementar - Apêndice B).

A via de vigilância NMD é um mecanismo de controle de qualidade de mRNA pós-transcricional responsável pela degradação de mRNAs contendo PTC que levaria à produção de proteínas encurtadas com efeitos deletérios para o organismo se permanecessem intactas (HUG et al., 2016). O alto número de amostras com proteínas encurtadas sugere uma falha na maquinaria NMD, principal mecanismo de reparo que atua degradando mRNAs com códons de parada prematuros. Conjecturamos que mutações e falhas nos genes da maquinaria NMD poderiam explicar por que proteoformas encurtadas foram observadas nos dados de MS. Assim, foram selecionadas amostras identificadas com proteínas encurtadas e que também apresentavam mutações em genes pertencentes à maquinaria NMD. Todas as 171 amostras têm pelo menos uma mutação (*missense* e indel *inframe*) em um dos 117 genes classificados como parte da maquinaria NMD. As observações mais frequentes foram para genes de famílias de RPS e RPL. O gene RPL14 é mutado exclusivamente por indels e em todos os quatro tipos de câncer (91%). Como visto no lado esquerdo da Figura 17, esta mutação não tem efeito deletério sobre a proteína. Os genes EIF4G1 e RPL29 também apresentam mutações deletérias missense e INDEL, mas apresentam menor frequência mutacional, com 15% e 12%, respectivamente. Os genes RPS2, RPS27L, RPL9 têm uma mutação missense e têm frequências mutacionais de 65%, 58% e 48%, respectivamente. Apenas o gene RPS27L sofre mutação com efeito deletério. Esses genes também estão presentes em todos os quatro tipos de câncer. Os genes UPF1 estão mutados em 8% das amostras. Os genes UPF1 e EIF4G1 interagem com as vias das famílias de genes RPS e RPL. Os genes RPL10L e RPLP1 estão presentes na maioria das amostras de câncer de mama e próstata e apresentam mutações com impacto neutro e deletério. O gene GSPT1 foi identificado com maior frequência no câncer de ovário e apresenta mutações com impacto neutro. Comparando a anotação da frequência de mutações na parte superior

e o tipo de câncer na parte inferior da Figura 17, amostras de câncer de mama e câncer de ovário têm as mutações mais frequentes na maquinaria NMD. As amostras de câncer colorretal têm a maioria de suas mutações de NMD nos 8 principais genes mais frequentes mutados (RPL14 até RPL13).

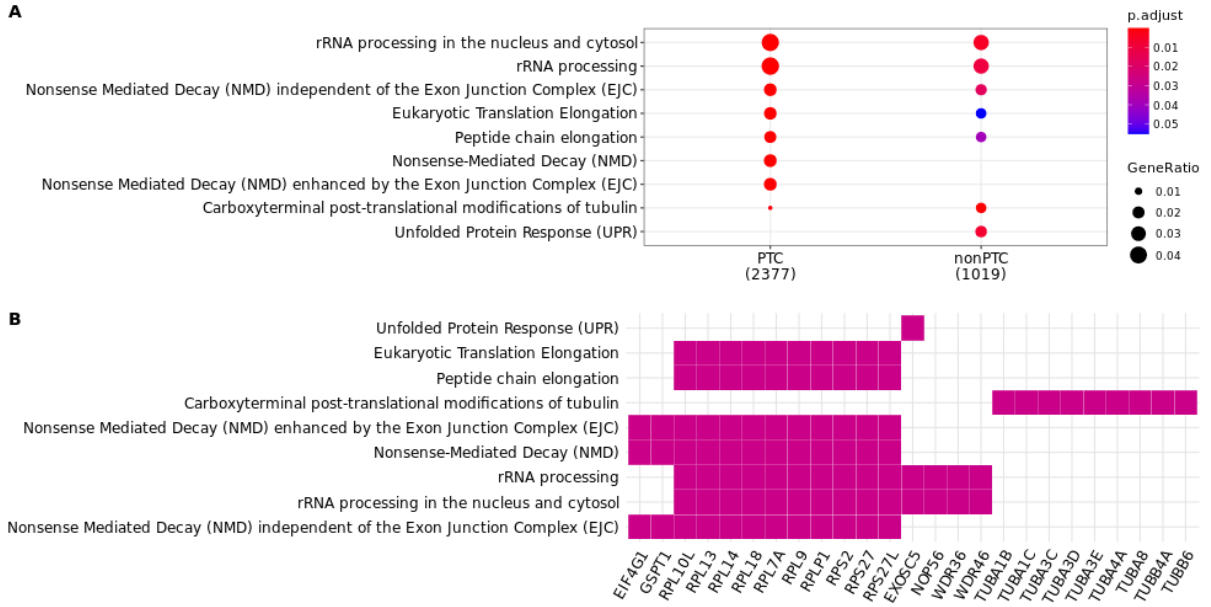


Figura 16. Análise de enriquecimento de todos os genes mutados em amostras com PTC em comparação com o conjunto de genes mutados em amostras sem PTC de dbPepVar.
A) Gráfico de pontos de vias de Reactome enriquecidas de termos selecionados. O tamanho dos pontos representa a proporção de genes, dada pela proporção de genes associados a cada uma das vias Reactome e o total de genes mutados em amostras com PTC, e a cor dos pontos representa os valores de p ajustados pela taxa de falsa descoberta. B) gráfico de mapa de calor de genes mutados e termos enriquecidos de lista de genes de amostras com PTC.

Uma vez que mutações *missense* identificadas em genes de maquinaria NMD ainda podem resultar em uma proteína funcional, o impacto previsto de todas as mutações na função proteica de tais produtos gênicos foi avaliado. O PROVEAN foi usado para prever o impacto funcional de mutações em genes que são encontrados apenas em vias associadas com NMD (Figura 16B). A maioria das amostras (152) tem mutações deletérias, perfazendo 88% de todas as 171 amostras. A porcentagem de mutações deletérias para os 20 principais genes de maquinaria NMD mutados pode ser vista no lado esquerdo da Figura 17 e as mutações neutras ou deletérias para cada gene estão disponíveis na Tabela Suplementar 1.

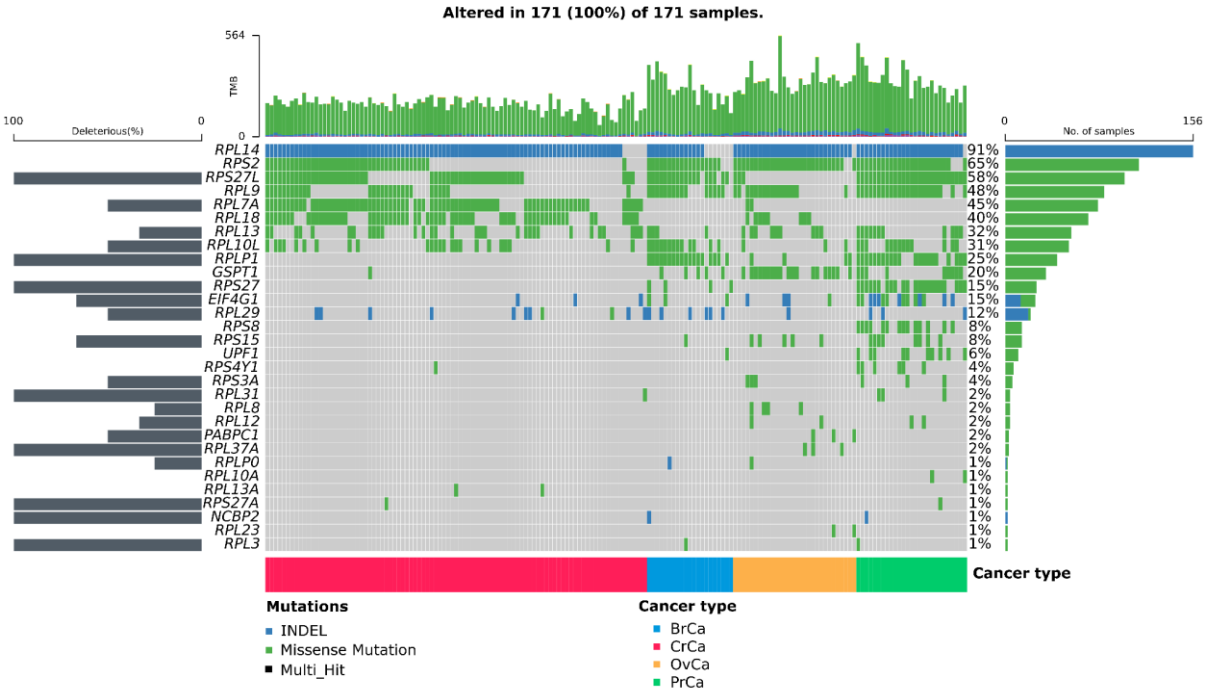


Figura 17. Análise oncoplot dos 20 principais genes mais mutados da maquinaria NMD com mutações deletérias identificadas com PROVEAN Choi; Chan (2015) em dbPepVar.
A família de genes RPL e RPS são mais frequentes. As amostras de câncer de mama e câncer de próstata apresentam as mutações mais frequentes na maquinaria NMD.

4.4 ANÁLISE DOS RESULTADOS SOBRE APLICAÇÃO DO DBPEPVAR EM AMOSTRAS DE CÂNCER

Ao usar dados de dbPepVar para MS coletados de diferentes tipos de câncer, vários marcadores gênicos de reconhecida relevância para o câncer foram identificados, validando o potencial da abordagem para detectar variantes de interesse: ACTG1, POTEE e ECHS1 (Figura 14). O gene POTEE é expresso em próstata normal e neoplásica e testículo, ovário e placenta normais (BERA et al., 2002). Este gene é descrito em alguns estudos como um oncogene para tumorigênese e progressão de cânceres colorretais, ovarianos, de mama, próstata e pâncreas, e um possível novo marcador molecular para diagnóstico clínico e tratamento (BARGER et al., 2018; BERA et al., 2004; CINE et al., 2014; HAO et al., 2020; SHEN et al., 2019; XU et al., 2020). A superexpressão de POTEE pode promover comportamento agressivo das células, indicam estadiamento tumoral avançado e mau prognóstico em pacientes com câncer de pâncreas e colorretal (HAO et al., 2020; XU et al., 2020), e também tem sido observado como um marcador de mau prognóstico em câncer de ovário, onde foi descrito como um potencial alvo terapêutico (BARGER et al, 2018).

Esta expressão aumentada também foi detectada em cânceres de próstata e mama (BERA et al., 2004; e CINE et al., 2014).

O gene ECHS1 atua na segunda etapa da via de beta-oxidação de ácidos graxos mitocondriais (NAIR et al., 2016). A desregulação do metabolismo dos ácidos graxos foi observada em muitos tipos de câncer, incluindo carcinoma de células renais, câncer de mama, câncer de próstata, câncer de ovário e câncer colorretal (LI, N. et al., 2021; LIN et al., 2007; LI, R. et al., 2021; NAIR et al., 2016; QU et al., 2020; SHI et al., 2015). A inibição da proteína ECHS1 reduz a proliferação celular em câncer de mama (SHI et al., 2015). A superexpressão desta proteína está associada à diferenciação, metástase e mau prognóstico no câncer colorretal (LI, R. et al., 2021; XIE, J.-P. et al., 2014). Com dbPepVar, mutações na proteína codificada pelo gene ECHS1 foram identificadas para câncer de ovário, mama, próstata e colorretal. A mutação Pro163Leu (rs371582393), encontrada em câncer de próstata pelo dbPepVar, é relatada em ClinVar associada à deficiência de ECHS1 e, no banco de dados COSMIC, está associada a câncer de próstata e pulmão. Segundo o PROVEAN, essa alteração tem efeito deletério sobre a proteína. Sua presença no COSMIC valida a identificação feita pelo dbPepVar e também corrobora a literatura que apresenta essa proteína como alvo terapêutico para esta doença. Outra mutação de interesse é Ala158Thr (rs960738876) que é relatado em COSMIC associado a linfoma maligno. Utilizando o dbPepVar, verificou-se que esta mutação está presente nos cânceres de mama e próstata e essa alteração tem efeito deletério sobre a proteína, segundo o PROVEAN.

A análise cuidadosa, guiada por hipóteses, das informações disponíveis no dbPepVar pode auxiliar na definição dos próximos passos da pesquisa e na obtenção de ideias sobre os mecanismos que atuam na manutenção do processo de tradução. Proteínas truncadas codificadas pelos genes HBB, LDHA e CFAP65 aparecem frequentemente mutadas nas amostras de câncer descritas neste trabalho. Na célula, existem mecanismos de controle de erros que evoluíram para degradar preferencialmente RNAs aberrantes ou não funcionais e evitar erros na biogênese e função do RNA, porém, esses mecanismos podem falhar gerando proteínas aberrantes. Proteínas não funcionais aberrantes são frequentemente tóxicas para as células e resultam em muitas doenças humanas (POLI et al., 2018; LINDEBOOM et al., 2019; LINDEBOOM et al., 2016). A introdução de PTCs em tais genes codificadores pode ter várias consequências, incluindo *exon skipping* e diminuição da

estabilidade do mRNA, bem como encurtamento de proteínas (MORT et al., 2008), e alguns desses genes mutados têm uma relação de causa ou consequência estabelecida com a etiologia ou progressão do câncer. Apesar de menos frequente, o gene LDHA também possui uma mutação frameshift e foi identificado nos quatro tipos de câncer. O LDHA codifica uma importante enzima, que mantém a glicólise e outras atividades metabólicas, sendo regulado positivamente em cânceres humanos e associado à agressividade tumoral (XIE, H. et al., 2014). O silenciamento do gene LDHA no câncer de mama inibe a migração e a invasão via regulação negativa da glicólise no receptor do fator de crescimento epidérmico 2 (ErbB2) (HE et al., 2019). A proteína do gene CFAP65 foi identificada com mutações frameshift presentes no câncer colorretal. O gene CFAP65 codifica uma proteína associada a cílios e flagelos, que é altamente expressa em testículos (ZHANG et al., 2019). Mutações nesta proteína estão associadas à infertilidade masculina e astenospermia (LI et al., 2020; WANG et al., 2019).

De acordo com BioXpress (DINGERDISSEN et al, 2018), a proteína CFAP65 é super expressa em cânceres de cabeça e pescoço, tireóide e pulmão; e com baixa expressão em câncer de próstata e mama. A mutação Asn1325Metfs (rs773843180) não está associada a doenças congênitas e também não há estudos mostrando sua relação com câncer, sendo relatada exclusivamente em dbPepVar e apenas em amostras colorretais. A proteína mutada identificada tem 68,83% em tamanho em comparação com a proteína de referência, faltando regiões importantes, incluindo os sítios de fosforilação Ser1715 e Ser1736, ambos relatados no iPTMnet e Phosphosite relacionados ao câncer de fígado e mama. Assim, é possível que a perda desses sítios, assim como a redução do tamanho da proteína, possa estar relacionada ao tipo de câncer investigado neste trabalho.

Outras proteínas de interesse são SATL1, H2AC20 e AP3B1. A alta expressão de SATL1 tem sido associada a uma maior taxa de sobrevida no carcinoma ovariano (LIN et al., 2020). O Projeto Achilles, que cataloga genes essenciais para diferentes linhagens celulares (TSHERNIAK et al., 2017) descobriram que o silenciamento de SATL1 leva ao aumento da aptidão em linhagens celulares de ovário e intestino grosso e diminui a aptidão para linhagens celulares de mama e intestino delgado, como apontado pelo painel EnrichR (CHEN et al., 2013; KULESHOV et al., 2016). Existem mutações *nonsense* no gene SATL1 em 20% das amostras de câncer colorretal analisadas. H2AC20 é um regulador mestre da expressão gênica dependente de

receptores de estrogênio alfa e é especialmente expresso em receptores de estrogênio positivos (ER+) tecidos de câncer de mama. Tem sido sugerido como um alvo para intervenção no câncer, pois os oncogenes são regulados positivamente por H2AC20 através do recrutamento de um ativador (SU et al., 2014). O dbPepVar identifica H2AC20 com mutações *nonsense* e *missense* (Ala->Thr, Ala-> Val) em amostras de câncer de ovário e mutações *missense* (Met->Thr) e mutações *start-loss* (Met->Ile) em amostras de câncer de próstata. Assim, os resultados encontrados nesse trabalho são coerentes com outros resultados encontrados na literatura e são evidências de genes envolvidos na progressão tumoral.

Além disso, proteínas truncadas podem ser traduzidas, mas geralmente são menos abundantes na célula devido à degradação por mecanismos de reparo de mRNA mutado (VICENTE-CRESPO; PALÁCIOS, 2010). Esse processo de degradação do mRNA é uma forma de evitar erros na síntese proteica (HUG et al., 2016; LINDEBOOM et al., 2016). Uma via de vigilância de RNA bem conhecida é a maquinaria NMD, que atua na degradação de mRNAs que possuem PTCs (LINDEBOOM et al., 2016; PECCARELLI; KEBAARA, 2014), conforme mostrado na análise de enriquecimento (ver Figura 16). Em alguns casos, a maquinaria NMD pode ser ineficiente, permitindo a síntese de proteínas curtas aberrantes (HUG et al., 2016; PECCARELLI; KEBAARA, 2014; VICENTE-CRESPO; PALACIOS, 2010). Os processos que envolvem esse mecanismo ainda são desconhecidos, mas sabe-se que a localização do PTC em relação ao *exon junction complex* (EJC) pode permitir que alguns mRNAs não sejam degradados pela via NMD (HUG et al., 2016; PECCARELLI; KEBAARA, 2014; VICENTE-CRESPO; PALACIOS, 2010). Os genes mutados do maquinário NMD em amostras com mutações que inserem PTC foram o foco de uma análise com o objetivo de reunir evidências de uma hipótese que possa explicar a ineficiência da maquinaria.

Quando um ribossomo detecta um PTC localizado acima de 50 nucleotídeos a montante de um EJC, fatores de terminação ligados ao ribossomo e o EJC recrutam UPF1, que se liga a SMG1, eRF1 e eRF3 para formar o complexo SURF, marcando o mRNA para degradação (MAQUAT et al., 2010; POPP; MAQUAT, 2013). Nesse contexto, nossos dados apontam para 14 amostras de PrCa com mutações em UPF1 e quatro amostras de BrCa com as mesmas mutações. Houve também uma mutação no gene codificador de proteínas do complexo eRF3 GSPT1, que medeia a terminação da tradução em eucariotos (SALAS-MARCO; BEDWELL, 2004) Outras

mutações no EIF4G1, que codifica uma proteína que faz parte do complexo EIF4F, foram identificadas. Este complexo facilita o recrutamento de mRNA para o ribossomo durante a fase de iniciação da síntese proteica (LEJEUNE et al., 2004). Todas as 205 amostras analisadas tiveram proteínas com PTC detectadas também tinham genes de maquinaria NMD mutados (Figuras 15 e 17), o que pode tornar um mecanismo de atuação menos eficiente, e os genes citados acima não são apenas importantes na via NMD, mas em outros processos relacionados ao desenvolvimento do câncer (GOUDARZI; LINDSTRÖM, 2016). Os resultados obtidos podem ser explorados em outros estudos de biologia de sistemas, verificando se mutações em proteínas afetam a rede de interação proteína-proteína da via NMD e como essas alterações afetam a proliferação tumoral.

Entre os genes da maquinaria NMD relatados em dbPepVar (ver Figura 17), RPL14 teve o maior número de mutações, principalmente indels. Este gene codifica uma proteína ribossomal que possui uma região de comprimento altamente polimórfico composta por uma repetição trinucleotídica (GCT), resultando em uma extensão de resíduos de alanina na proteína codificada (LIU et al., 2018). A baixa expressão desta proteína está associada a um mau prognóstico de sobrevida em pacientes com câncer de mama (LIN et al., 2021), enquanto a alta expressão suprime a proliferação, migração, invasão e processo de transição epitelial-mesenquimal no carcinoma nasofaríngeo (ZHANG et al., 2021). Também está relacionado à instabilidade de microssatélites no câncer colorretal (YU et al., 2019). De acordo com o PROVEAN, as mutações relatadas identificadas têm um impacto neutro na proteína. No entanto, as mutações A159_K160insAA, A159_K160insAAA (rs369485042) foram encontradas em linhagens celulares de melanoma e câncer de mama, respectivamente (FAKTOR et al., 2020; YANG e LAZAR, (2014). No dbPepVar, essas mutações foram identificadas em cânceres de próstata, mama, ovário e colorretal. O COSMIC relata a mutação G148_T149insA (rs764850005) associada a câncer de estômago e ovário; no dbPepVar, foi identificada em amostras colorretais. A mutação G148_T149insAAAA (rs764850005) em amostras de câncer de ovário, colorretal e próstata também foi relatada no dbPepVar. O achado experimental dessas variantes em outros estudos específicos para células cancerígenas aumenta a confiança de outras mutações relatadas através da abordagem utilizada para construir dbPepVar.

A interrupção da biogênese ribossômica leva à liberação de proteínas ribossomais livres de ribossomos, que possuem atividades extra-ribossômicas como

regulação da apoptose, ciclo celular, neoplasticidade e outras (XU et al., 2016; ZHOU et al., 2015). A proteína ribossomal RPS27L foi previamente descrita como um gene induzível de p53 de tipo selvagem e alvo direto - nesse contexto, RPS27L promoveu apoptose induzida por etoposídeo (HE e SUN, 2007). Deletando RPS27L causou instabilidade e resultou em linfomagênese em antecedentes heterozigotos Trp53 (XIONG et al., 2014). Um estudo com tecidos de câncer de mama indica *downregulation* do RPS27L como um sinal de sobrevivência para facilitar a formação do tumor (XIONG et al., 2018). Esses resultados apontam para a importância de atividade normal de RPS27L em promover apoptose e dificultar a formação de tumores. Nossa análise mostra mutações em RPS27L em 60% das amostras em todos os tipos de câncer. Todas essas mutações são deletérias, de acordo com PROVEAN, o que indica uma perda da função da proteína na amostra.

As mutações na maquinaria NMD podem ser exploradas como biomarcadores para um prognóstico da doença. Nesse contexto, o gene EIF4G1 é descrito como um possível biomarcador para diferentes tipos de câncer, pois quando inibido juntamente com os membros da família EIF4G, atuam na clonogenicidade, formação da esfera tumoral e invasão celular, características associadas à progressão tumoral (JAISWAL et al., 2019). Além disso, o aumento do nível de expressão de EIF4G1 também está associado à baixa sobrevida do paciente para cânceres de glioma cerebral, rim, fígado, pulmão, mesotelioma, pâncreas, câncer de próstata, sarcoma e melanoma cutâneo (JAISWAL et al., 2019). No dbPepVar, o EIF4G1 está mutado em 15% das amostras entre todos os tipos de câncer, sendo mais frequente no PrCa. A regulação positiva da proteína RPL3 livre de ribossomos é proposta como um possível promotor da atividade apoptótica independente de TP53 em pacientes com câncer de cólon durante a quimioterapia (PAGLIARA et al., 2016). Da mesma forma, a proteína codificada pelo gene SMG7 também atua na estabilização e atuação da proteína supressora de tumor TP53 (LUO et al., 2016).

5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

A abordagem proteômica refere-se ao conjunto de técnicas adotadas para caracterização das proteínas expressas em tecidos e fluidos distintos. Essa abordagem destaca-se por permitir a quantificação e identificação do proteoma de uma espécie, analisando a dinâmica desse produto gênico e determinando funções e interações proteicas. No processo de identificação de proteínas, geralmente utiliza-se um espectrômetro de massas que mede o valor de massa/carga de fragmentos proteicos ou peptídicos. Nesse processo, é requerido uma base de dados de proteínas teóricas juntamente com um software identificador que realizar o processo de digestão *in silico* e compara os valores obtidos com os fragmentos experimentais.

Banco de dados, como UniProt, geralmente é utilizado para a identificação das proteínas. Embora possua eficiência, esse tipo de base, considerada referência, não contempla variações genéticas que ocorrem no DNA e que modificam a cadeia polipeptídica. Nesse contexto, surge a proteogenômica que utiliza dados proteômicos para a validação genômica. Geralmente, esse processo é feito por meio da criação de banco de dados personalizados que incorporam variações genéticas ou variações específicas a uma espécie. Mesmo havendo êxito nesses processos, geralmente os bancos de dados desenvolvidos não contemplam uma quantidade de variações que permitam identificar eficientemente os peptídeos e/ou tem o aumento significativo na base criada.

Desse modo, a pesquisa apresentada teve como objetivo desenvolver abordagens computacionais para criação de uma base de dados contendo peptídeos variantes, de modo que, houvesse a adição de vários tipos de mutações e que fosse possível associar à patologias. Assim, foram incorporadas à base mutações *missense*, *nonsense*, *indels*, *frameshift*, perda de códon final e alteração do códon inicial. A base desenvolvida uniu informações contidas no Refseq de proteínas com dados de polimorfismo genéticos presentes no dbSNP. Além disso, foram consideradas frequências alélicas raras e comuns das mutações presentes em algumas populações. A abordagem proposta difere-se das bases apresentadas na literatura por realizar o processo de identificação submetendo as bases de dados referência e variante paralelamente ao software identificador. Essa prática permite quantificar os peptídeos variante identificados pela nossa abordagem e reduzir o espaço de busca. Além disso, nossa abordagem é complementada pela análise dos scores dos

peptídeos identificados como forma de selecionar aqueles que apresentam maior qualidade.

A abordagem proteogenômica apresentada aqui foi aplicada à análise pan-câncer para identificar variantes e genes comuns e os impactos de mutações na via de reparo de mRNA. Nossos resultados mostram que tipos notavelmente diferentes de câncer possuem elementos semelhantes e compartilham mutações que podem servir tanto para fortalecer as evidências já apresentadas na literatura quanto para direcionar pesquisas para descobrir novos biomarcadores comuns aos tipos de câncer. Em relação à via de reparo do mRNA, os resultados obtidos sugerem que mutações na maquinaria do NMD podem permitir a síntese de proteínas truncadas e, consequentemente, impactar no prognóstico da doença. Essas mutações foram identificadas principalmente em proteínas ribossomais importantes para a estabilidade do ribossomo. Por fim, o dbPepVar apresenta uma visão geral da proteômica para várias amostras de diferentes tipos de câncer, permitindo que pesquisadores busquem informações sobre o conjunto de mutações que afetam grupos específicos de amostras, analisem as mutações e alterações mais frequentes nos resíduos de aminoácidos e tenham acesso direto informações sobre cada tipo de mutação. Abordagens baseadas em espectrometria de massa ganham suas limitações da técnica, por exemplo, a ausência do peptídeo mutante na identificação devido ao tamanho. Nesse caso, o peptídeo tríptico mutante pode ser relativamente pequeno (por exemplo, menos de seis aminoácidos) e, portanto, difícil de corresponder de maneira confiável ao espectro MS/MS correspondente. Isso pode explicar a ausência de algumas mutações na maquinaria NMD das amostras de PTC, pois delimitamos o tamanho do peptídeo entre 7 e 35 resíduos de aminoácidos. Em trabalhos futuros, é importante analisar a expressão gênica combinada com o genótipo das amostras para verificar o nível de abundância de proteínas encurtadas em diferentes tipos de câncer. Os mecanismos e hipóteses discutidos podem ser usados por outros pesquisadores para avançar na compreensão do câncer e o portal desenvolvido está disponível para a pronta descoberta de novos *insights* e hipóteses.

REFERÊNCIAS

ALFARO, J. A.; IGNATCHENKO, A.; IGNATCHENKO, V.; et al. Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines. **Genome medicine**, v. 9, n. 1, p. 62, 2017. Disponível em: <http://dx.doi.org/10.1186/s13073-017-0454-9>.

ANTCZAK, Andrzej, Wojciech Kluźniak, Dominika Wokołorczyk, Aniruddh Kashyap, Anna Jakubowska, Jacek Gronwald, Tomasz Huzarski et al. "A common nonsense mutation of the BLM gene and prostate cancer risk and survival." **Gene 532**, no. 2 (2013): 173-176.

BAEISSA, Hanadi et al. Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. **Oncotarget**, v. 8, n. 13, p. 21290, 2017.

BALLESTÉ, Raquel Nancy. "Proteomics: Technology and Applications." The Use of Mass Spectrometry Technology (MALDI-TOF) in **Clinical Microbiology**. Academic Press, 2018. 1-17.

BARBOSA, Eduardo Buzolin et al. Proteomics: methodologies and applications to the study of human diseases. **Revista da Associacao Medica Brasileira**, v. 58, n. 3, p. 366-375, 2012.

BARGER, C. J.; ZHANG, W.; SHARMA, A.; et al. Expression of the POTE gene family in human ovarian cancer**. Scientific reports**, v. 8, n. 1, 2018. Sci Rep.

BERA, T. K.; HUYNH, N.; MAEDA, H.; et al. Five POTE paralogs and their splice variants are expressed in human prostate and encode proteins of different lengths. **Gene**, v. 337, p. 45–53, 2004.

BERA, T. K.; ZIMONJIC, D. B.; POPESCU, N. C.; et al. POTE, a highly homologous gene family located on numerous chromosomes and expressed in prostate, ovary, testis, placenta, and prostate cancer. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 26, 2002. Proc Natl Acad Sci U S A.

BOGDANOVA, N., Cybulski, C., Bermisheva, M., Datsyuk, I., Yamini, P., Hillemanns, P., ... & Dörk, T. (2009). A nonsense mutation (E1978X) in the ATM gene is associated with breast cancer. **Breast cancer research and treatment**, 118(1), 207-211.

BUHR, Florian et al. Synonymous codons direct cotranslational folding toward different protein conformations. **Molecular cell**, v. 61, n. 3, p. 341-351, 2016.

BUNGER, Maureen K. et al. Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. **Journal of proteome research**, v. 6, n. 6, p. 2331-2340, 2007.

CASTELLANA, Natalie; BAFNA, Vineet. Proteogenomics to discover the full coding content of genomes: a computational perspective. **Journal of proteomics**, v. 73, n. 11, p. 2124-2135, 2010.

CATHERMAN, Adam D.; SKINNER, Owen S.; KELLEHER, Neil L. Top down proteomics: facts and perspectives. **Biochemical and biophysical research communications**, v. 445, n. 4, p. 683-693, 2014.

CAO, R.; SHI, Y.; CHEN, S.; et al. dbSAP: single amino-acid polymorphism database for protein variation detection. **Nucleic acids research**, v. 45, n. D1, p. D827–D832, 2017. Disponível em: <http://dx.doi.org/10.1093/nar/gkw1096>.

CHEN, E. Y.; TAN, C. M.; KOU, Y.; et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. **BMC bioinformatics**, v. 14, p. 128, 2013.

CHEN, Y., LU, H., ZHANG, N., ZHU, Z., WANG, S., & LI, M. PremPS: Predicting the impact of missense mutations on protein stability. **PLoS computational biology**, 16(12), e1008543. (2020).

CHOI, Jae-Hwan et al. A start codon mutation of the FRMD7 gene in two Korean families with idiopathic infantile nystagmus. **Scientific reports**, v. 5, p. 13003, 2015.

CHOI, Y.; CHAN, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. **Bioinformatics** , v. 31, n. 16, p. 2745–2747, 2015.

CHOI, Yongwook et al. Predicting the functional effect of amino acid substitutions and indels. **PloS one**, v. 7, n. 10, p. e46688, 2012.

CINE, N.; BAYKAL, A. T.; SUNNETCI, D.; et al. Identification of ApoA1, HPX and POTEE genes by omic analysis in breast cancer. **Oncology reports**, v. 32, n. 3, p. 1078–1086, 2014.

COSCIA, F. et al. Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status. **Nature communications**, v. 7, p. ncomms12645, 2016.

CUCCHIARA, V.; COOPERBERG, M. R.; DALL'ERA, M.; et al. Genomic Markers in Prostate Cancer Decision Making. **European urology**, v. 73, n. 4, p. 572–582, 2018

CYR, Jennifer L. et al. The predicted truncation from a cancer-associated variant of the MSH2 initiation codon alters activity of the MSH2-MSH6 mismatch repair complex. **Molecular carcinogenesis**, v. 51, n. 8, p. 647-658, 2012.

DHAMIJA, Sonam et al. A pan-cancer analysis reveals nonstop extension mutations causing SMAD4 tumour suppressor degradation. **Nature cell biology**, v. 22, n. 8, p. 999-1010, 2020.

DINGERDISSEN, H. M.; TORCIVIA-RODRIGUEZ, J.; HU, Y.; et al. BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery. Nucleic acids research, v. 46, n. **Database issue**, p. D1128, 2018. Oxford University Press.

DONNELLY DP, RAWLINS CM, DEHART CJ, FORNELLI L, SCHACHNER LF, LIN Z, LIPPENS JL, ALURI KC, SARIN R, CHEN B, LANTZ C. Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. **Nature methods**. 2019 Jul;16(7):587-94.

DRABKIN, Harold J.; RAJBHANDARY, Uttam L. Initiation of protein synthesis in mammalian cells with codons other than AUG and amino acids other than methionine. **Molecular and cellular biology**, v. 18, n. 9, p. 5140-5147, 1998.

FAKTOR, J.; GRASSO, G.; KOKAS, F. Z.; et al. The effects of p53 gene inactivation on mutant proteome expression in a human melanoma cell model. Biochimica et Biophysica Acta (BBA) - **General Subjects**, 2020.

FOLKMAN, Lukas et al. DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. **Bioinformatics**, v. 31, n. 10, p. 1599-1606, 2015.

FRUMKIN, Idan et al. Codon usage of highly expressed genes affects proteome-wide translation efficiency. **Proceedings of the National Academy of Sciences**, v. 115, n. 21, p. E4940-E4949, 2018.

GOUDARZI, K. M.; LINDSTRÖM, M. S. Role of ribosomal protein mutations in tumor development (Review). **International Journal of Oncology**, 2016.

GROSS, Jürgen H. "Tandem mass spectrometry." In **Mass spectrometry**, pp. 539-612. Springer, Cham, 2017.

GUAN, Bin et al. Functional analysis of in-frame indel ARID1A mutations reveals new regulatory mechanisms of its tumor suppressor functions. **Neoplasia**, v. 14, n. 10, p. 986-993, 2012.

GUERREIRO, Rita et al. A nonsense mutation in PRNP associated with clinical Alzheimer's disease. **Neurobiology of aging**, v. 35, n. 11, p. 2656. e13-2656. e16, 2014.

GUSTIN, John P. et al. GATA3 frameshift mutation promotes tumor growth in human luminal breast cancer cells and induces transcriptional changes seen in primary GATA3 mutant breast cancers. **Oncotarget**, v. 8, n. 61, p. 103415, 2017.

HAO, Q.; GAO, L.; NIU, W.; et al. POTEE stimulates the proliferation of pancreatic cancer by activating the PI3K/Akt/GSK-3β/β-catenin signaling. **BioFactors** , v. 46, n. 4, p. 685–692, 2020.

HE, H.; SUN, Y. Ribosomal protein S27L is a direct p53 target that regulates apoptosis. **Oncogene**, v. 26, n. 19, p. 2707–2716, 2007.

HE, Ya-Chao et al. TMEM230 stop codon mutation is rare in parkinson's disease and essential tremor in eastern China. **Movement Disorders**, v. 32, n. 2, p. 301-302, 2017.

HINNEBUSCH, Alan G.; IVANOV, Ivaylo P.; SONENBERG, Nahum. Translational control by 5′-untranslated regions of eukaryotic mRNAs. **Science**, v. 352, n. 6292, p. 1413-1416, 2016.

HOFFMANN, E.; STROOBANT, V. Analysis of biomolecules. **Mass Spectrometry: Principles and Applications**, v. 3, p. 306-402, 2007.

HUG, N.; LONGMAN, D.; CÁCERES, J. F. Mechanism and regulation of the nonsense-mediated decay pathway**. Nucleic Acids Research**, 2016.

HUNT, Ryan C. et al. Exposing synonymous mutations. **Trends in Genetics**, v. 30, n. 7, p. 308-321, 2014.

HUSTOFT, Hanne Kolsrud et al. A critical review of trypsin digestion for LC-MS based proteomics. In: **Integrative Proteomics**. InTech, 2012.

IGLESIAS-GATO, D.; WIKSTRÖM, P.; TYANOVA, S.; et al. The Proteome of Primary Prostate Cancer. **European urology**, v. 69, n. 5, p. 942–952, 2016.

JACKSON, Richard J.; HELLEN, Christopher UT; PESTOVA, Tatyana V. The mechanism of eukaryotic translation initiation and principles of its regulation. **Nature reviews Molecular cell biology**, v. 11, n. 2, p. 113, 2010.

JAFFE, J. D., Berg, H. C., Church, G. M., Proteogenomic mapping as a complementary method to perform genome annotation. **Proteomics** 2004, 4, 59–77.

JAISWAL, P. K.; KOUL, S.; PALANISAMY, N.; KOUL, H. K. Eukaryotic Translation Initiation Factor 4 Gamma 1 (EIF4G1): a target for cancer therapeutic intervention? **Cancer cell international**, v. 19, p. 224, 2019.

JIANG, L., WANG, M., LIN, S., JIAN, R., LI, X., Chan, J., ... & Doherty, J. A. (2020). A quantitative proteome map of the human body. **Cell**, 183(1), 269-283.

KO, Frankie CF; CHOW, King L. A mutation at the start codon defines the differential requirement of dpy-11 in Caenorhabditis elegans body hypodermis and male tail. **Biochemical and biophysical research communications**, v. 309, n. 1, p. 201-208, 2003.

KREBS, J. E.; KILPATRICK, S. T.; GOLDSTEIN, S. E. Lewin's Genes XI Burlington. **MA: Jones & Bartlett Publishers**, 2014.

KROLL, José Eduardo et al. A tool for integrating genetic and mass spectrometry-based peptide data: Proteogenomics Viewer: PV: A genome browser-like tool, which includes MS data visualization and peptide identification parameters. **Bioessays**, v. 39, n. 7, p. 1700015, 2017.

KROLL JE, de Souza SJ, de Souza GA. Identification of rare alternative splicing events in MS/MS data reveals a significant fraction of alternative translation initiation sites. **PeerJ**. 2014 Nov 13;2:e673. doi: 10.7717/peerj.673. PMID: 25405079; PMCID: PMC4232841.

KULESHOV, M. V.; JONES, M. R.; ROUILLARD, A. D.; et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. **Nucleic acids research**, v. 44, n. W1, p. W90–7, 2016.

KUMAR, Ashok. Respiratory Distress. In: PARTHASARATHY, A. **IAP Textbook of pediatrics**. JP Medical Ltd, 2016.

KUPER, Willemijn FE et al. **The c. 1A> C start codon mutation in CLN3 is associated with a protracted disease course**. JIMD reports, v. 52, n. 1, p. 23-27, 2020.

LEE, Sooncheol et al. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. **Proceedings of the National Academy of Sciences**, v. 109, n. 37, p. E2424-E2432, 2012.

LEJEUNE, F.; RANGANATHAN, A. C.; MAQUAT, L. E. eIF4G is required for the pioneer round of translation in mammalian cells. Nature structural & molecular biology, v. 11, n. 10, p. 992–1000, 2004**. Nature Publishing Group**.

LENINGER, M., HER, A. S., & TRAASETH, N. J. Inducing conformational preference of the membrane protein transporter EmrE through conservative mutations. **Elife**, 8. (2019).

LI, Minghui et al. Mutations in DNA-binding loop of NFAT5 transcription factor produce unique outcomes on protein–DNA binding and dynamics. **The Journal of Physical Chemistry B**, v. 117, n. 42, p. 13226-13234, 2013.

LI, R.; HAO, Y.; WANG, Q.; et al. ECHS1, an interacting protein of LASP1, induces sphingolipid-metabolism imbalance to promote colorectal cancer progression by regulating ceramide glycosylation. **Cell death & disease**, v. 12, n. 10, 2021.

LI, W.; WU, H.; LI, F.; et al. Biallelic mutations in CFAP65 cause male infertility with multiple morphological abnormalities of the sperm flagella in humans and mice. Journal of medical genetics, v. 57, n. 2, 2020. **J Med Genet**.

LIN, H.; WANG, J.; WEN, X.; et al. A prognosis-predictive nomogram of ovarian cancer with two immune-related genes: and. **Oncology letters**, v. 20, n. 5, p. 204, 2020.

LIN, J. F.; XU, J.; TIAN, H. Y.; et al. Identification of candidate prostate cancer biomarkers in prostate needle biopsy specimens using proteomic analysis. International journal of cancer, v. 121, n. 12, 2007. **Int J Cancer**.

LIN, Z.; PENG, R.; SUN, Y.; ZHANG, L.; ZHANG, Z. Identification of ribosomal protein family in triple-negative breast cancer by bioinformatics analysis. **Bioscience reports**, v. 41, n. 1, 2021.

LINDEBOOM, R. G. H.; SUPEK, F.; LEHNER, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. **Nature genetics**, v. 48, n. 10, p. 1112–1118, 2016.

LINDEBOOM, R. G. H.; VERMEULEN, M.; LEHNER, B.; SUPEK, F. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. **Nature genetics**, v. 51, n. 11, p. 1645–1651, 2019.

LIU, J.; CHENG, Y.; WANG, X.; CUI, X. Supervised Penalty Matrix Decomposition for Tumor Differentially Expressed Genes Selection. **Chinese Journal of Electronics**, 2018.

LIU, Yansheng; BEYER, Andreas; AEBERSOLD, Ruedi. On the dependency of cellular protein levels on mRNA abundance. **Cell**, v. 165, n. 3, p. 535-550, 2016.

LUCK, K., SHEYNKMAN, G. M., ZHANG, I., & VIDAL, M. (2017). Proteome-scale human interactomics. **Trends in biochemical sciences**, 42(5), 342-354.

LUO, H.; COWEN, L.; YU, G.; JIANG, W.; TANG, Y. SMG7 is a critical regulator of p53 stability and function in DNA damage stress response. **Cell discovery**, v. 2, p. 15042, 2016.

LUO, H., HUANG, Y., STEPANAUSKAS, R., & TANG, J. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. **Nature Microbiology**, 2(8), 1-9. (2017).

MEHLFERBER MM, JEFFERY ED, SAQUING J, et al. Characterization of protein isoform diversity in human umbilical vein endothelial cells via long-read proteogenomics. **RNA Biol.** 2022;19(1):1228-1243. doi:10.1080/15476286.2022.2141938

MAQUAT, L. E.; TARN, W.-Y.; ISKEN, O. The pioneer round of translation: features and functions. **Cell**, v. 142, n. 3, p. 368–374, 2010.

MACHADO K. C. T., Fortuin S, Tomazella GG, Fonseca AF, Warren RM, Wiker HG, de Souza SJ, de Souza GA. On the Impact of the Pangenome and Annotation Discrepancies While Building Protein Sequence Databases for Bacteria Proteogenomics. **Front Microbiol**. 2019 Jun 20;10:1410. doi: 10.3389/fmicb.2019.01410. PMID: 31281302; PMCID: PMC6596428.

MARLIN, Régine et al. Mutation HOXB13 c. 853delT in Martinican prostate cancer patients. **The Prostate**, v. 80, n. 6, p. 463-470, 2020.

MARQUIONI, Vinícius, NUNES, Francis Morais Franco e NOVO-MANSUR, Maria Teresa Marques. "Protein Identification by Database Searching of Mass Spectrometry Data in the Teaching of Proteomics**." Journal of Chemical Education** 98.3 (2021): 812-823.

MERTINS, Philipp et al. Proteogenomics connects somatic mutations to signalling in breast cancer. **Nature**, v. 534, n. 7605, p. 55, 2016.

MILLER, Rachel M., et al. "Improved protein inference from multiple protease bottom-up mass spectrometry data." **Journal of proteome research** 18.9 (2019): 3429-3438.

MORADIAN, Annie et al. The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. **Proteomics**, v. 14, n. 4-5, p. 489-497, 2014.

MORT, M.; IVANOV, D.; COOPER, D. N.; CHUZHANOVA, N. A. A meta-analysis of nonsense mutations causing human genetic disease. **Human mutation**, v. 29, n. 8, p. 1037–1047, 2008.

NAIR, P.; HAMZEH, A. R.; MOHAMED, M.; et al. Novel ECHS1 mutation in an Emirati neonate with severe metabolic acidosis**. Metabolic Brain Disease**, 2016.

NELSON, David L.; COX, Michael M. **Princípios de bioquímica de Lehninger**. Artmed Editora, 2022.

NESVIZHSKII, Alexey I. Proteogenomics: concepts, applications and computational strategies. **Nature methods**, v. 11, n. 11, p. 1114, 2014.

NIEHAUS, M., et al. "Transmission-mode MALDI-2 mass spectrometry imaging of cells and tissues at subcellular resolution." **Nature methods** 16.9 (2019): 925-931.

PAGLIARA, V.; SAIDE, A.; MITIDIERI, E.; et al. 5-FU targets rpL3 to induce mitochondrial apoptosis via cystathionine-β-synthase in colon cancer cells lacking p53. **Oncotarge**t, v. 7, n. 31, p. 50333–50348, 2016.

PAGNAMENTA, Alistair T. et al. A homozygous variant disrupting the PIGH start-codon is associated with developmental delay, epilepsy, and microcephaly. **Human mutation**, v. 39, n. 6, p. 822-826, 2018.

PECCARELLI, M.; KEBAARA, B. W. Regulation of natural mRNAs by the nonsense-mediated mRNA decay pathway. **Eukaryotic cell**, v. 13, n. 9, p. 1126–1135, 2014.

PERUTKA, Zdeněk, and ŠEBELA, Marek. "Pseudotrypsin: A Little-Known Trypsin Proteoform." **Molecules** 23, no. 10 (2018): 2637.

POLI, M. C.; EBSTEIN, F.; NICHOLAS, S. K.; et al. Heterozygous Truncating Variants in POMP Escape Nonsense-Mediated Decay and Cause a Unique Immune Dysregulatory Syndrome**. American journal of human genetics**, v. 102, n. 6, p. 1126–1142, 2018.

Poluri, K.M., Gulati, K., Sarkar, S. **Structural and Functional Properties of Proteins.** In: Protein-Protein Interactions. Springer, Singapore. https://doi.org/10.1007/978-981-16-1594-8_1, 2021.

POPP, M. W.-L.; MAQUAT, L. E. Organizing principles of mammalian nonsense-mediated mRNA decay**. Annual review of genetics**, v. 47, p. 139–165, 2013.

PROKOFYEVA, D., Bogdanova, N., Dubrowinskaja, N., Bermisheva, M., Takhirova, Z., Antonenkova, N., ... & Dörk, T. (2013). Nonsense mutation p. Q548X in BLM, the gene mutated in Bloom's syndrome, is associated with breast cancer in Slavic populations. **Breast cancer research and treatment**, 137(2), 533-539.

QU, Y. Y.; ZHAO, R.; ZHANG, H. L.; et al. Inactivation of the AMPK-GATA3-ECHS1 Pathway Induces Fatty Acid Synthesis That Promotes Clear Cell Renal Cell Carcinoma Growth. **Cancer research**, v. 80, n. 2, 2020. Cancer Res.

RENUSE, Santosh; CHAERKADY, Raghothama; PANDEY, Akhilesh. Proteogenomics. **Proteomics**, v. 11, n. 4, p. 620-630, 2011.

REUTER, Kerstin et al. PreTIS: a tool to predict non-canonical 5'UTR translational initiation sites in human and mouse. **PLoS computational biology**, v. 12, n. 10, p. e1005170, 2016

SALAS-MARCO, J.; BEDWELL, D. M. GTP hydrolysis by eRF3 facilitates stop codon decoding during eukaryotic translation termination. **Molecular and cellular biology**, v. 24, n. 17, p. 7769–7778, 2004.

SCHANDORFF, Søren et al. A mass spectrometry–friendly database for cSNP identification. **Nature methods**, v. 4, n. 6, p. 465, 2007.

SHEN, Z.; FENG, X.; FANG, Y.; et al. POTEE drives colorectal cancer development via regulating SPHK1/p65 signaling. Cell death & disease, v. 10, n. 11, 2019. **Nature Publishing Group**.

SHEYNKMAN, Gloria M. et al. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. **Journal of proteome research**, v. 13, n. 1, p. 228-240, 2013.

SHI, Y.; QIU, M.; WU, Y.; HAI, L. MiR-548-3p functions as an anti-oncogenic regulator in breast cancer. **Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie**, v. 75, 2015.

SMIRNOVA, Victoria V., et al. "Ribosomal leaky scanning through a translated uORF requires eIF4G2." **Nucleic acids research** 50.2 (2022): 1111-1127.

SMITH, L., KELLEHER, N. & The Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. **Nat Methods** 10, 186–187 (2013). https://doi.org/10.1038/nmeth.2369

SONG, Chunxia et al. Large-scale quantification of single amino-acid variations by a variation-associated database search strategy. **Journal of proteome research**, v. 13, n. 1, p. 241-248, 2013.

SORTINO, Katherine, BRIANNA L. Tylec, RUNPU Chen, and YIJUN Sun. "Conserved and transcript-specific functions of the RESC factors, RESC13 and RESC14, in kinetoplastid RNA editing." **RNA 28**, no. 11 (2022): 1496-1508.

STEEN, Hanno; MANN, Matthias. The ABC's (and XYZ's) of peptide sequencing. **Nature reviews Molecular cell biology**, v. 5, n. 9, p. 699, 2004.

STEFL, Shannon et al. Molecular mechanisms of disease-causing missense mutations. **Journal of molecular biology**, v. 425, n. 21, p. 3919-3936, 2013.

SU, C.-H.; TZENG, T.-Y.; CHENG, C.; HSU, M.-T. An H2A histone isotype regulates estrogen receptor target genes by mediating enhancer-promoter-3'-UTR interactions in breast cancer cells. **Nucleic acids research**, v. 42, n. 5, p. 3073–3088, 2014.

SUZEK, Baris E. et al. UniRef: comprehensive and non-redundant UniProt reference clusters. **Bioinformatics**, v. 23, n. 10, p. 1282-1288, 2007.

TAN, Zhijing et al. Single amino acid variant profiles of subpopulations in the MCF-7 breast cancer cell line. **Journal of proteome research**, v. 16, n. 2, p. 842-851, 2017.

TSHERNIAK, A.; VAZQUEZ, F.; MONTGOMERY, P. G.; et al. Defining a Cancer Dependency Map**. Cell**, v. 170, n. 3, p. 564–576.e16, 2017.

TYANOVA, Stefka; TEMU, Tikira; COX, Juergen. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. **Nature protocols**, v. 11, n. 12, p. 2301, 2016.

UVERSKY, Vladimir N. et al. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. **BMC genomics**, v. 10, n. 1, p. S7, 2009.

VÉGVÁRI, Ákos. Mutant Proteogenomics. In: **Proteogenomics**. Springer, Cham, 2016. p. 77-91.

VICENTE-CRESPO, Marta; PALACIOS, Isabel M. Nonsense-mediated mRNA decay and development: shoot the messenger to survive?. 2010.

VOGEL, Christine; MARCOTTE, Edward M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. **Nature Reviews Genetics**, v. 13, n. 4, p. 227, 2012.

WANG, W.; TU, C.; NIE, H.; et al. Biallelic mutations in CFAP65 lead to severe asthenoteratospermia due to acrosome hypoplasia and flagellum malformations. **Journal of medical genetics**, v. 56, n. 11, 2019. J Med Genet.

WANG, D., ERASLAN, B., WIELAND, T., HALLSTRÖM, B., HOPF, T., ZOLG, D.P., ZECHA, J., ASPLUND, A., Li, L.H., MENG, C. and FREJNO, M., 2019. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. **Molecular systems biology**, 15(2), p.e8503.

What kinds of gene variants are possible? National Library of Medicine. Disponível em: <https://medlineplus.gov/genetics/understanding/mutationsanddisorders/possiblemutations/> Acesso em 20 out. 2022.

WONG, Jason WH; CAGNEY, Gerard. An overview of label-free quantitation methods in proteomics by mass spectrometry. In: **Proteome bioinformatics**. Humana Press, 2010. p. 273-283.

XIE, H.; HANAI, J.-I.; REN, J.-G.; et al. Targeting lactate dehydrogenase--a inhibits tumorigenesis and tumor progression in mouse models of lung cancer and impacts tumor-initiating cells. **Cell metabolism**, v. 19, n. 5, p. 795–809, 2014.

XIE, J.-P.; ZHU, X.-S.; DAI, Y.-C.; et al. Expression of enoyl coenzyme A hydratase, short chain, 1, in colorectal cancer and its association with clinicopathological characteristics. **Molecular and Clinical Oncology**, v. 2, n. 6, p. 1081, 2014. Spandidos Publications.

XIONG, X.; ZHAO, Y.; TANG, F.; et al. Ribosomal protein S27-like is a physiological regulator of p53 that suppresses genomic instability and tumorigenesis. **eLife**, v. 3, p. e02236, 2014.

XU, Q.; CHEN, J.; PENG, M.; et al. POTEE promotes colorectal carcinoma progression via activating the Rac1/Cdc42 pathway. **Experimental cell research**, v. 390, n. 1, 2020. Exp Cell Res.

XU, X.; XIONG, X.; SUN, Y. The role of ribosomal proteins in the regulation of cell proliferation, tumorigenesis, and genomic integrity. **Science China. Life sciences**, v. 59, n. 7, p. 656–672, 2016.

YANG, X.; LAZAR, I. M. XMAn: a Homo sapiens mutated-peptide database for the MS analysis of cancerous cell states. **Journal of proteome research**, v. 13, n. 12, p. 5486–5495, 2014.

YU, C.; HONG, H.; ZHANG, S.; et al. Identification of key genes and pathways involved in microsatellite instability in colorectal cancer. **Molecular medicine reports**, v. 19, n. 3, p. 2065, 2019. Spandidos Publications.

YU, G.; HE, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. **Molecular bioSystems**, v. 12, n. 2, p. 477–479, 2016.

YU, G.; WANG, L.-G.; HAN, Y.; HE, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. **Omics: a journal of integrative biology**, v. 16, n. 5, p. 284–287, 2012.

ZHANG, Taylor et al. Identification of a single base-pair mutation of TAA (Stop codon)→ GAA (Glu) that causes light chain extension in a CHO cell derived IgG1. In: **mabs**. Taylor & Francis, 2012. p. 694-700.

ZHANG, X.; SHEN, Y.; WANG, X.; et al. A novel homozygous CFAP65 mutation in humans causes male infertility with multiple morphological abnormalities of the sperm flagella. **Clinical genetics**, v. 96, n. 6, p. 541–548, 2019

ZHANG, Z.; ZHANG, Y.; QIU, Y.; MO, W.; YANG, Z. Human/eukaryotic ribosomal protein L14 (RPL14/eL14) overexpression represses proliferation, migration, invasion and EMT process in nasopharyngeal carcinoma. **Bioengineered**, v. 12, n. 1, 2021. Bioengineered.

ZHANG, Yaoyang et al. Protein analysis by shotgun/bottom-up proteomics. **Chemical reviews**, v. 113, n. 4, p. 2343-2394, 2013.

ZHANG, Zheng, Meghan Burke, Yuri A. Mirokhin, Dmitrii V. Tchekhovskoi, Sanford P. Markey, Wen Yu, Raghothama Chaerkady, Sonja Hess, and Stephen E. Stein. Reverse and random decoy methods for false discovery rate estimation in high mass accuracy peptide spectral library searches**. Journal of proteome research** 17, no. 2 (2018): 846-857.

ZHENG, Siyuan; KIM, Hoon; VERHAAK, Roel GW. Silent mutations make some noise. **Cell**, v. 156, n. 6, p. 1129-1131, 2014.

ZHOU, X.; LIAO, W.-J.; LIAO, J.-M.; LIAO, P.; LU, H. Ribosomal proteins: functions beyond the ribosome. **Journal of molecular cell biology**, v. 7, n. 2, p. 92–104, 2015.

APÊNDICES

APÊNDICE A - TRABALHO PUBLICADO COMO COLABORADOR.

Coelho *et al. BMC Medical Genomics* (2020) 13:30
https://doi.org/10.1186/s12920-020-0694-1

BMC Medical Genomics

SOFTWARE Open Access

neoANT-HILL: an integrated tool for identification of potential neoantigens

Ana Carolina M. F. Coelho[1], André L. Fonseca[1], Danilo L. Martins[1], Paulo B. R. Lins[1], Lucas M. da Cunha[1,2] and Sandro J. de Souza[1,3,4*]

Abstract

Background: Cancer neoantigens have attracted great interest in immunotherapy due to their capacity to elicit antitumoral responses. These molecules arise from somatic mutations in cancer cells, resulting in alterations on the original protein. Neoantigens identification remains a challenging task due largely to a high rate of false-positives.

Results: We have developed an efficient and automated pipeline for the identification of potential neoantigens. neoANT-HILL integrates several immunogenomic analyses to improve neoantigen detection from Next Generation Sequence (NGS) data. The pipeline has been compiled in a pre-built Docker image such that minimal computational background is required for download and setup. NeoANT-HILL was applied in The Cancer Genome Atlas (TCGA) melanoma dataset and found several putative neoantigens including ones derived from the recurrent RAC1:P29S and SERPINB3:E250K mutations. neoANT-HILL was also used to identify potential neoantigens in RNA-Seq data with a high sensitivity and specificity.

Conclusion: neoANT-HILL is a user-friendly tool with a graphical interface that performs neoantigens prediction efficiently. neoANT-HILL is able to process multiple samples, provides several binding predictors, enables quantification of tumor-infiltrating immune cells and considers RNA-Seq data for identifying potential neoantigens. The software is available through github at https://github.com/neoanthill/neoANT-HILL.

Keywords: Neoantigens, Cancer, Immunogenomic analyses

Background

Recent studies have demonstrated that T cells can recognize tumor-specific antigens that bind to human leukocyte antigens (HLA) molecules at the surface of tumor cells [1, 2]. During tumor progression, accumulating somatic mutations in the tumor genome can affect protein-coding genes and result in mutated peptides [1]. These mutated peptides, which are present in the malignant cells but not in the normal cells, may act as neoantigens and trigger T-cell responses due to the lack of thymic elimination of autoreactive T-cells (central tolerance) [3–5]. As result, these neoantigens appear to represent ideal targets attracting great interest for cancer immunotherapeutic strategies, including therapeutic vaccines and engineered T cells [1, 6].

In the last few years, advances in next-generation sequencing have provided an accessible way to generate patient-specific data, which allows the prediction of tumor neoantigens in a rapid and comprehensive manner [7]. Several approaches have been developed, such as pVAC-Seq [8], MuPeXI [9], TIminer [10] and TSNAD [11], which predict potential neoantigens produced by non- synonymous mutations. However, none of these proposed tools considers tumor transcriptome sequencing data (RNA-seq) for identifying somatic mutations. Moreover, only one of these tools provides quantification of the fraction of tumor-infiltrating immune cell types (Supplementary: Table S1).

Here we present a versatile tool with a graphical user interface (GUI), called neoANT-HILL, designed to identify potential neoantigens arising from cancer somatic mutations. neoANT-HILL integrates complementary features to prioritizing mutant peptides based on predicted binding affinity and mRNA expression level (Fig. 1). We used datasets from GEUVADIS

* Correspondence: sandro@imd.ufrn.br
[1]Bioinformatics Multidisciplinary Enviroment (BioME), Institute Metropolis Digital, Federal University of Rio Grande do Norte, UFRN, Natal, Brazil
[3]Brain Institute, Federal University of Rio Grande do Norte, UFRN, Natal, Brazil
Full list of author information is available at the end of the article

Figura Suplementar 1. Artigo publicado como colaborador na revista BMC Medical Genomics ISSN: 1755-8794, Qualis A3.

APÊNDICE B - ANÁLISE DE ENRIQUECIMENTO

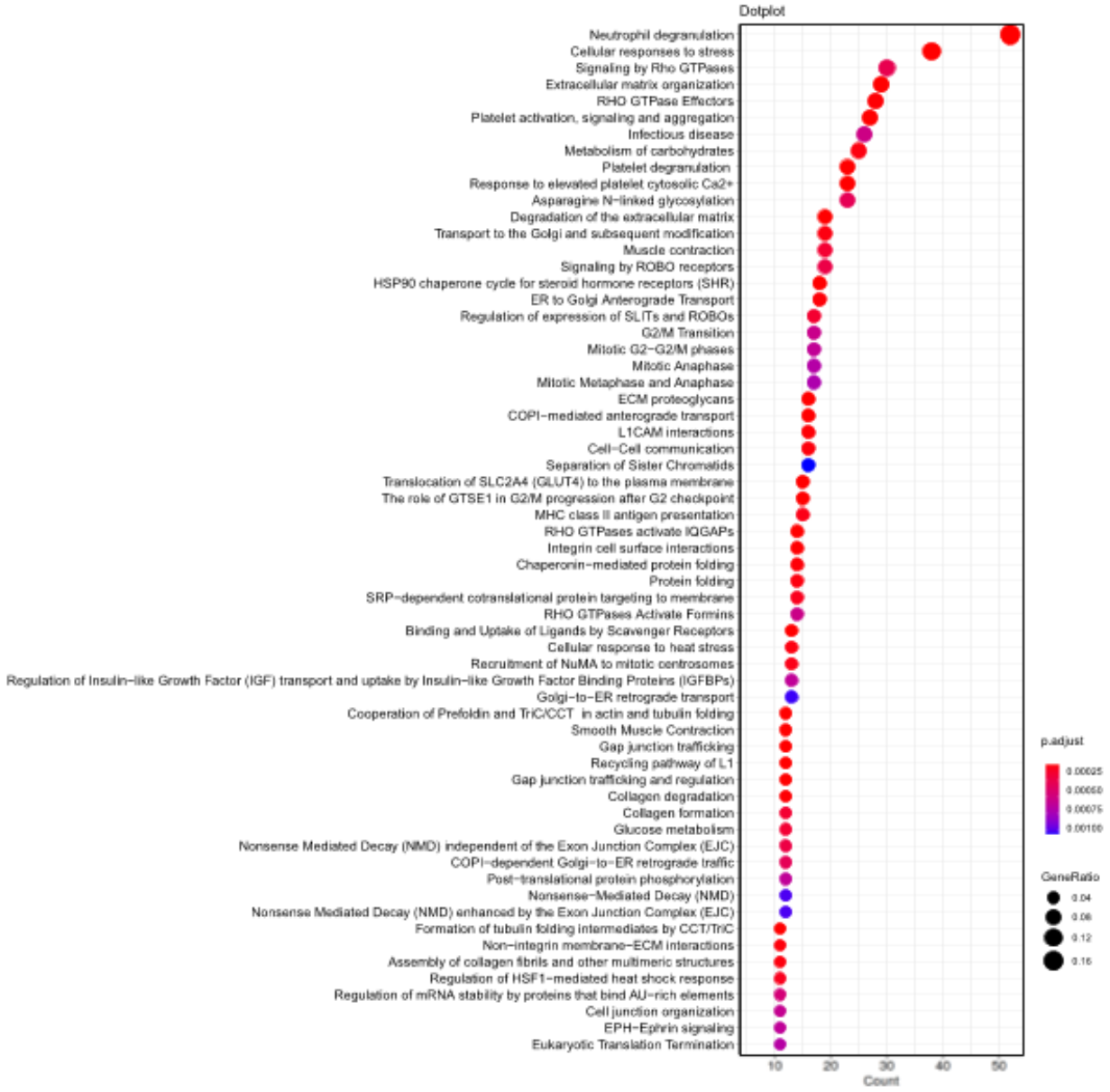


Figura Suplementar 2. Análise de enriquecimento dos 400 genes mais mutados de amostras com PTC de dbPepVar. Gráfico de pontos de vias de Reactome enriquecidas de 60 vias de Reactome ordenadas pelo número de contagens de genes enriquecidos. O tamanho dos pontos representa a proporção de genes, dada pela proporção de genes associados a cada uma das vias Reactome e o total de 400 genes mais mutados em amostras com PTC, e a cor dos pontos representa os valores ajustados a P (BH).

APÊNDICE C - TABELA SUPLEMENTAR 1: ANÁLISE DE IMPACTO DA FUNÇÃO BIOLÓGICA DE UMA PROTEÍNA USANDO PROVEAN (NMD).

Gene	Refseq protein	Mutation	SNP	Functional impact
EIF4G1	NP_937884	N1257D	rs1035168505	Deleterious
EIF4G1	NP_937884	I1250V	rs573974788	Neutral
EIF4G1	NP_937884	K524del	rs568141854	Deleterious
GSPT1	NP_002085	G101C	rs185937624	Neutral
GSPT1	NP_002085	G102C	rs771724893	Neutral
GSPT1	NP_002085	G92C	rs11544193	Neutral
GSPT1	NP_002085	P96S	rs370239120	Neutral
GSPT2	NP_060564	V63I	rs782057159	Neutral
NCBP2	NP_001294965	K60del	rs756862806	Deleterious
NCBP2	NP_031388	K78del	rs756862806	Deleterious
PABPC1	NP_002559	L126V	rs747729489	Deleterious
PABPC1	NP_002559	S4N	rs778812386	Neutral
PPP2CA	NP_002706	E9D	rs775616921	Neutral
RPL10A	NP_009035	A89S	rs11553986	Neutral
RPL10L	NP_542784	E145K	rs751834362	Neutral
RPL10L	NP_542784	H51R	rs367951246	Deleterious
RPL12	NP_000967	D153E	rs576415871	Neutral
RPL12	NP_000967	V139L	rs770575415	Neutral
RPL12	NP_000967	V73M	rs143660046	Deleterious

RPL13	NP_001230059	A93T	rs9930567	Neutral
RPL13	NP_001230059	M136T	rs148355340	Deleterious
RPL13	NP_001230059	V138I	rs781576247	Neutral
RPL13A	NP_001257420	A93D	rs150697570	Neutral
RPL13A	NP_036555	A154D	rs150697570	Neutral
RPL13A	NP_036555	G40D	rs199838441	Neutral
RPL14	NP_003964	A152_A153insTA	rs758279993	Neutral
RPL14	NP_003964	A155_A156insTAAAA	rs758279993	Neutral
RPL14	NP_003964	A159_K160insA	rs369485042	Neutral
RPL14	NP_003964	A159_K160insAA	rs369485042	Neutral
RPL14	NP_003964	A159_K160insAAA	rs369485042	Neutral
RPL14	NP_003964	A159_K160insAAAAA	rs369485042	Neutral
RPL14	NP_003964	A159_K160insAAAAA	rs758279993	Neutral
RPL14	NP_003964	A159_K160insAAAAAA	rs764283379	Neutral
RPL14	NP_003964	A159_K160insAKA	rs764283379	Neutral
RPL14	NP_003964	G148_T149insA	rs764850005	Neutral
RPL14	NP_003964	G148_T149insAAAA	rs764850005	Neutral
RPL14	NP_003964	G148_T149insAAAAA	rs764850005	Neutral
RPL14	NP_003964	G148_T149insAAAAAA	rs764850005	Neutral
RPL14	NP_003964	G148_T149insAAAAAAAA	rs764850005	Neutral
RPL14	NP_003964	G148_T149insPAA	rs764850005	Neutral
RPL14	NP_003964	G148_T149insPAAAA	rs764850005	Neutral

RPL14	NP_003964	T149del	rs111899316	Neutral
RPL17	NP_001186273	R3L	rs761827697	Deleterious
RPL18	NP_001257419	N11D	rs11554942	Neutral
RPL21	NP_000973	Q69E	rs200767103	Neutral
RPL22	NP_000974	E30D	rs758745546	Neutral
RPL23	NP_000969	A34G	rs554988536	Neutral
RPL23A	NP_000975	N93D	rs763483485	Deleterious
RPL27A	NP_000981	L78V	rs115721984	Neutral
RPL29	NP_000983	A137P	rs769488393	Neutral
RPL29	NP_000983	K128_A135del	rs761953149	Deleterious
RPL3	NP_001029025	H273R	rs767797371	Deleterious
RPL30	NP_000980	I97V	rs776163436	Neutral
RPL31	NP_000984	T112N	rs148028338	Deleterious
RPL37A	NP_000989	A68S	rs745839935	Deleterious
RPL3L	NP_005052	I180V	rs926901161	Neutral
RPL5	NP_000960	E82D	rs771331683	Deleterious
RPL6	NP_001307071	T167M	rs199644294	Neutral
RPL7	NP_000962	K9del	rs556746526	Neutral
RPL7A	NP_000963	N38D	rs781890308	Deleterious
RPL7A	NP_000963	K212R	rs781888936	Neutral
RPL7A	NP_000963	V207L	rs781795890	Neutral
RPL8	NP_150644	I98V	rs11539893	Neutral

RPL8	NP_150644	L96V	rs11539889	Neutral
RPL8	NP_150644	P103L	rs775973617	Deleterious
RPL8	NP_150644	V104L	rs770494770	Neutral
RPL9	NP_001020092	N42D	rs751277956	Neutral
RPL9	NP_001020092	D17E	rs763721363	Neutral
RPLP0	NP_444505	A292V	rs757288238	Neutral
RPLP0	NP_444505	D157E	rs754247040	Neutral
RPLP0	NP_444505	P280_A284del	rs572058538	Deleterious
RPLP0	NP_444505	V276L	rs752906137	Neutral
RPLP1	NP_000994	G45A	rs147484893	Deleterious
RPLP2	NP_000995	A87V	rs757586394	Neutral
RPS12	NP_001007	A47S	rs916600302	Deleterious
RPS14	NP_005608	A77S	rs772136334	Deleterious
RPS15	NP_001009	D82N	rs1011629191	Neutral
RPS15	NP_001009	Q32L	rs980644616	Deleterious
RPS15	NP_001009	Y97F	rs771534816	Deleterious
RPS16	NP_001308040	V38L	rs370701690	Neutral
RPS19	NP_001308413	T55M	rs147508369	Neutral
RPS2	NP_002943	N134D	rs11543098	Neutral
RPS2	NP_002943	D254E	rs374773804	Neutral
RPS2	NP_002943	V274I	rs753906824	Neutral
RPS24	NP_001135756	D80E	rs199598396	Deleterious

RPS26	NP_001020	V58I	rs769953745	Neutral
RPS27	NP_001021	P10L	rs146810634	Deleterious
RPS27A	NP_002945	S20L	rs368266147	Deleterious
RPS27A	NP_002945	T22M	rs765215187	Deleterious
RPS27L	NP_057004	N29D	rs370019223	Deleterious
RPS3	NP_001243731	R76G	rs11546227	Deleterious
RPS3A	NP_000997	N99D	rs1025970589	Deleterious
RPS3A	NP_000997	I204V	rs1015733074	Neutral
RPS4Y1	NP_000999	V46I	rs777332771	Neutral
RPS6	NP_001001	Q13E	rs538652085	Neutral
RPS7	NP_001002	A124V	rs761211921	Deleterious
RPS8	NP_001003	N138D	rs944331163	Neutral
RPSA	NP_001291217	A67V	rs781714830	Deleterious
UPF1	NP_001284478	I466V	rs769976755	Neutral