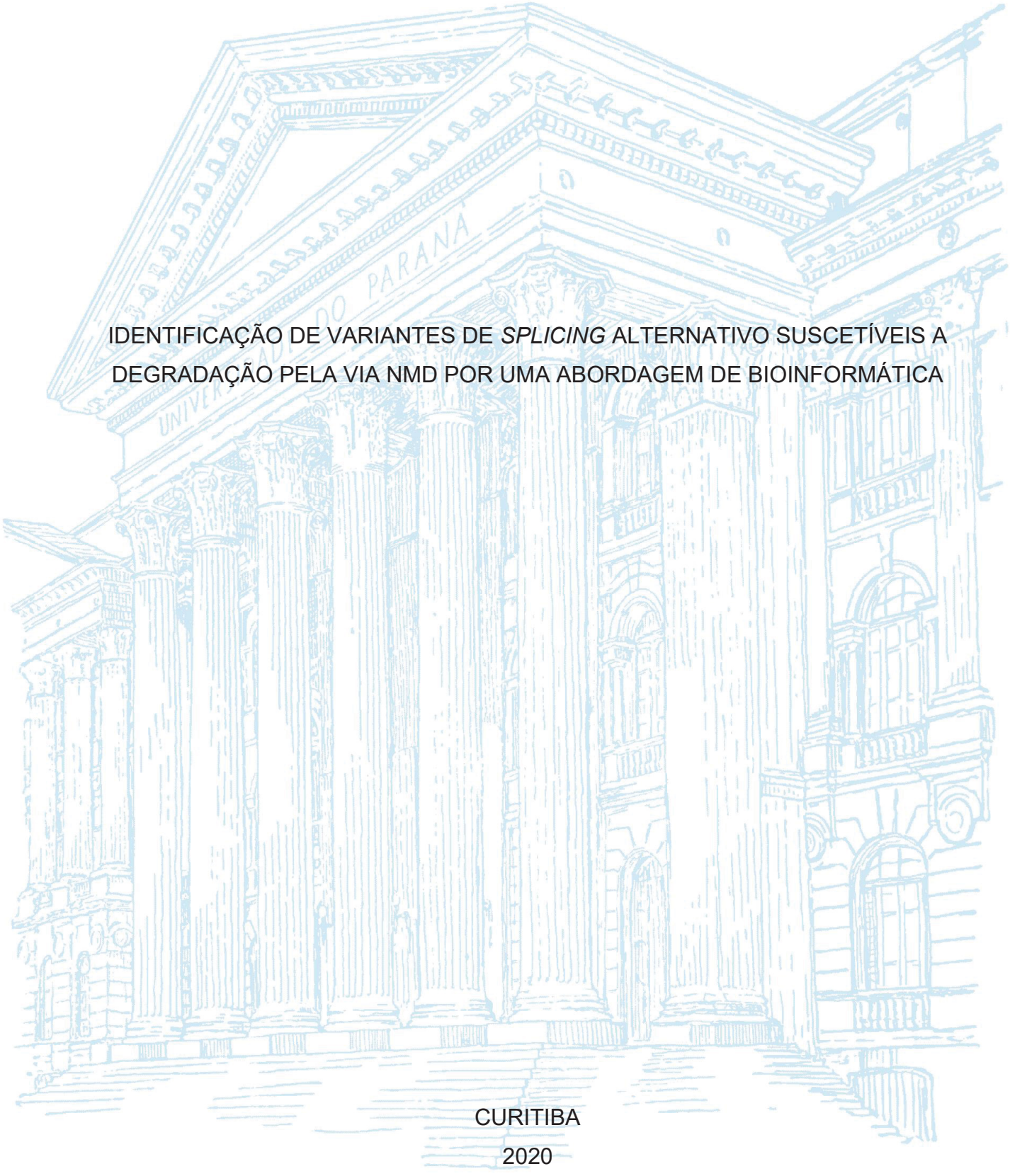
UNIVERSIDADE FEDERAL DO PARANÁ

VINÍCIUS DA SILVA COUTINHO PARREIRA



IDENTIFICAÇÃO DE VARIANTES DE *SPLICING* ALTERNATIVO SUSCETÍVEIS A DEGRADAÇÃO PELA VIA NMD POR UMA ABORDAGEM DE BIOINFORMÁTICA

CURITIBA

2020

VINÍCIUS DA SILVA COUTINHO PARREIRA

IDENTIFICAÇÃO DE VARIANTES DE *SPLICING* ALTERNATIVO SUSCETÍVEIS A DEGRADAÇÃO PELA VIA NMD POR UMA ABORDAGEM DE BIOINFORMÁTICA

Dissertação apresentada ao curso de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Bioinformática.

Orientador: Dr. Fabio Passetti

CURITIBA

2020

Catalogação na publicação
Sistema de Bibliotecas UFPR
Biblioteca de Educação Profissional e Tecnológica

Parreira, Vinícius da Silva Coutinho
Identificação de variantes de splicing alternativo suscetíveis a degradação pela via NMD por uma abordagem de Bioinformática / Vinícius da Silva Coutinho Parreira. – Curitiba, 2020.
78 p.: il.

Dissertação (Mestrado) – Universidade Federal do Paraná, Setor Educação Profissional e Tecnológica, Programa de Pós-Graduação em Bioinformática, 2020.
Orientador: Fabio Passetti

1. Genômica. 2. Proteômica. 3. Biologia computacional. 4. Bioinformática. I. Passetti, Fábio. II. Título. III. Universidade Federal do Paraná.

Elaboração: Angela Pereira de Farias Mengatto - CRB 9/1002

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **VINÍCIUS DA SILVA COUTINHO PARREIRA** intitulada: **"Identificação de variantes de *splicing* alternativo suscetíveis a degradação pela via NMD por uma abordagem de Bioinformática"**, sob orientação do Prof. Dr. FABIO PASSETTI, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.
A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 20 de fevereiro de 2020

DR FÁBIO PASSETTI
Presidente/Instituto Carlos Chagas-FIOCRUZ
Programa de Pós-graduação em Bioinformática-UFPR

DR RICARDO LEHTONEN RODRIGUES DE SOUZA
Avaliador Externo/Departamento de Genética-UFPR

DR MAURO ANTONIO ALVES CASTRO
Avaliador Interno/Programa de Pós-graduação em Bioinformática-UFPR

Dedico esta dissertação em especial à minha tia Elaine Tavares, cujos dias aqui na Terra tiveram por foco tornar a via de outros mais felizes. Uma pessoa que me deu muito apoio e que sei que me amava um tanto que não consigo mensurar.

AGRADECIMENTOS

- Ao Dr. Fabio Passetti por ter aceitado me orientar, por ter confiado em mim e por ter me ensinado a caminhar pela bioinformática. Também por todo o apoio além da vida acadêmica. Por ser uma excelente pessoa além de pesquisador, por estar atento e me guiar nos próximos passos.

- Ao Laboratório de Regulação da Expressão Gênica e o setor de Bioinformática do Instituto Carlos Chagas por compartilhar a estrutura para o desenvolvimento deste trabalho

- Ao Instituto Carlos Chagas, Fiocruz – Paraná, por fornecer a estrutura e financiamento para o projeto, e valorizar seus colaboradores.

- Ao Instituto Oswaldo Cruz, Fiocruz – Rio de Janeiro, por disponibilizar o servidor necessário para realizar algumas análises deste trabalho.

- Aos pesquisadores Dr. Raphael Tavares e Dra. Nicole Scherer pelas versões anteriores do SpliceProt.

- À Doutoranda Letícia Santos pelo grande auxílio no desenvolvimento deste trabalho, por trabalhar junto comigo e me compartilhar sua experiência em bioinformática.

- Ao pesquisador Dr. Helisson Faoro, por apoiar e auxiliar nos primeiros passos para ingresso na Pós-Graduação em Bioinformática da Universidade Federal do Paraná

- Ao programa de Pós-Graduação em Bioinformática da Universidade Federal do Paraná pela oportunidade de realização do curso de mestrado.

- À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pela concessão da bolsa de mestrado e pelo apoio financeiro para a realização desta pesquisa.

- À Deus, por sempre me mostrar o caminho, me dando forças, principalmente para encarar o desconhecido e tornar a minha jornada mais fácil.

- Aos meus pais, Marcos e Mariene, por me apoiarem à cima de todas as coisas. Por colocarem a minha felicidade à frente de tudo. Por estarem ao me lado mesmo que distante. Por terem me criado e acreditado na pessoa que me tornei

- À professora Myrian Cardoso Coutinho, minha avó, à quem vou sempre agradecer pelo incentivo acadêmico e sentimental. Por te me mostrado que o amor é um sentimento imensurável. Por permanecer comigo, no meu coração, mesmo após ter partido.

- À minha avó Emerenciana e ao meu avô Armando por todo apoio e carinho em toda minha vida, por sempre me incentivarem, por me fazerem me sentir uma das pessoas mais amadas desse mundo. Por me ajudarem a alcançar qualquer objetivo que eu possa almejar.

- Aos meus tios, Junior e Luzimar, que foram exemplos de generosidade e estiveram ao meu lado sempre me ajudando de coração puro.

- À minha tia Elaine por me mostrar que essa vida é passageira e que a verdadeira felicidade se encontra em compartilhar esse sentimento com todos a nossa volta.

- Aos meus primos Lucas, Luara e Luan, que me fazem sentir repleto de felicidade apenas com um sorriso. Por serem meus irmãos mais novos e acreditarem em mim.

- À Elaine Cristina Lima dos Santos, que mesmo distante é uma excelente amiga. Por compartilhar as mesmas risadas sinceras a cada reencontro. E por sempre me apoiar e incentivar.

- À Alice, minha afilhada, que me motiva a ser uma pessoa melhor a cada dia. Que me deixa feliz a cada sorriso sincero, que recupera minha infância a cada visita e que me surpreende em todas as atitudes.

- Aos meus amigos do Rio de Janeiro, em especial Vitor, Gabriel, Jonathan, Vilma e Vitória e que mesmo distantes permanecem comigo e me ensinaram que amizade não é algo que nem a distância nem o tempo podem consumir.

- Aos novos amigos da bioinfo, Aruana, Willian, Hugo, Alexandre, Letícia e Saloe por terem me ensinado mais da bioinformática e da amizade. Por terem me acolhido e tornarem os meus dias mais prazerosos, por estarem presentes além do trabalho e serem amigos verdadeiros.

- Aos amigos do RPG que já não são mais só do RPG, são pessoas que posso contar além de qualquer coisa, que estão ao meu lado e se importam comigo. Pessoas que conheço pouco, mas pelas quais já me sinto acolhido.

- Pelos amigos que o Kung fu me deu. Principalmente Geórgia e Lucas que fazem realmente Curitiba ser um lar, que mostraram para mim a verdadeira vida, que alcançaram meu coração muito rápido e me ajudaram a eu mesmo conhece-lo melhor.

- Aos demais amigos que a vida me deu, Alex, Jô e Renato que me permitiram seguir em frente mais forte a cada dia, que tornaram dias difíceis apenas dias diferentes e me fazem acreditar mais em mim e no mundo.

Pois há que aqueça o zero absoluto, há que transforme o breu em clareira, há que faça brotar da rocha seca. Há que congele o fogo, há que conforte a luz ofuscante, há que faça valer a vida curta. Pois se há impossível, não mais haverá.

RESUMO

. A estatística mais recente do GENCODE para a espécie humana, identificou 19.965 genes codificadores de proteína e um número de transcritos oriundos destes genes superior a 83.000. Esse fato está associado principalmente ao processamento do pré-mRNA após a transcrição denominado *splicing* alternativo. O *splicing* alternativo do pré-mRNA promove a formação de diferentes transcritos oriundos do mesmo gene devido ao reconhecimento alternativo de íntrons e éxons. Além disso, esses eventos podem alterar o potencial codificador da variante, como no caso da inserção de PTC (do inglês, Premature Termination Code) que pode levar a degradação da variante pela via de NMD (do inglês, *Nonsense Mediated Decay*). Em experimentos com ênfase em proteogenômica, é relevante a utilização de bancos de dados personalizados de sequências de proteínas que possuam informações de variantes por *splicing* alternativo e também do potencial codificador dessas variantes. Com esse objetivo, o nosso grupo de pesquisa desenvolveu o repositório SpliceProt para identificar peptídeos de proteoformas derivadas de variantes por *splicing* alternativo presentes no repositório RefSeq e dbEST. Entretanto, o repositório Unigene foi descontinuado em junho de 2019. O objetivo deste trabalho foi, então, atualizar o fluxo de construção do SpliceProt para receber dados do projeto Ensembl e além disso, adicionar uma etapa de filtragem para variantes por *splicing* alternativo que poderiam ser degradados pela via de NMD. Para tal, foram elaborados dois bancos de dados contendo dados de alinhamento e anotação do projeto Ensembl (Homo sapiens GRCh38 – versão 96) utilizando o padrão TSL (do inglês, *Transcript Support Level*) e tamanho da sequência como parâmetros de confiabilidade. Foi verificado discordâncias quanto o número de variantes em cada base de dados indicando que os dados disponíveis no projeto Ensembl não são correspondentes. Ao avaliar variantes preditas como alvo da via de NMD pelo software NMDClassifier, foi verificado que há variantes que não estão de acordo com a anotação do projeto Ensembl, o que pode refletir a necessidade de reformular como os dados estão anotados. Alguns transcritos já descritos na literatura como alvo da via de NMD foram preditos exclusivamente pelo programa NMDClassifier, outros, anotados exclusivamente pelo projeto Ensembl. Sendo assim, sugerimos utilizar as duas fontes de classificação para definir alvos da via de NMD, com destaque para anotação de transcritos codificadores de proteína com valor de TLS igual a 1. Essa etapa se mostrou de grande relevância devido a presença de variantes que eram potenciais alvo da via de NMD no conjunto de dados após tradução in silico. O parâmetro TSL foi avaliado como compatível com a esta abordagem. Entretanto alguns ajustes necessitam ser feitos para a otimização do repositório SpliceProt. Ainda assim, nossa abordagem permitiu identificar mais de 35.000 proteínas hipotéticas associadas à variantes por *splicing* alternativo que não estavam classificadas no projeto Ensembl como “codificadoras de proteína”. Nosso método pode então ser utilizado para a criação de um banco de proteínas hipotéticas e contribuir com a identificação de peptídeos em abordagens de proteômica.

Palavras-chave: *Splicing* alternativo, Proteogenômica, Bioinformática, Códon de Terminação Precoce, NMD.

ABSTRACT

The Last GENCODE statistics for Homo sapiens show that there is more than 19,900 protein-coding gene associated with more than 83,000 transcripts. This difference is associated with the post-transcription pre-mRNA processing, named alternative splicing. The alternative splicing is associated to non-canonical recognition of exons and introns and the formation of different transcripts from a single gene. Moreover, the alternative splicing may be associated to the alteration of coding capacity of the variant. For example, a PTC (Premature Termination Code) can be introduced in the transcript and making it sensitive to NMD (Nonsense Mediated Decay). For many proteogenomics research is very important the utilization of personal protein sequences databases with information about coding potential of splicing variants. In this manner, our research group have developed a repository named SpliceProt. This repository were designed to in silico identification of peptides derived from proteins associated with splicing variants in the RefSeq and dbEST databases. However, the Unigene database was retired at July, 2019. The aim of this project was update the SpliceProt pipeline to receive the Ensembl project data and introduce a step for NMD-target identification. We have created two databases, one with alignment information and other with annotation information available at Ensembl project (Homo sapiens GRCh – version 96). The TSL (Transcript Support Level) and the transcript sequence length was used to select all the trustworthy coordinates. We found some discordances about the number of variants in each database, indicating that the data available in the Ensembl project are not corresponding. When we evaluate variants predicted as a NMD-target by the NMDClassifier software, it was found that some variants are not in accordance with the annotation of the Ensembl project. This may reflect the necessity of reformulate how this data is annotated. Some transcripts that have been already described in the literature as NMD-target were predicted exclusively by the NMDClassifier. On the other way, some validated transcripts are exclusively annotated as NMD-target by Ensembl project. Therefore, we suggest using the two classification sources to define NMD-targets, with emphasis on the Ensembl annotation of protein coding transcripts with a TLS value 1. This step proved to be very relevat due to the presence of potential NMD-target variants in the data set after in silico translation. The TSL parameter was evaluated as compatible with this approach. However, we have to perform some adjustments to optimize the SpliceProt repository. Our approach was able to identify more than 35,000 hypothetical proteins associated with alternative splicing variants that were not classified in the Ensembl project as “protein-coding”. Our method can be used to create a hypothetical protein database and contribute to peptides identification in proteomic research.

Keywords: Alternative *Splicing*, Proteogenomics, Bioinformatics, Premature Termination Code, NMD.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

RNA – Ácido Ribonucleico (do inglês*: Ribonucleic acid*).

mRNA - RNAs mensageiro (do inglês: *Messenger RNA*).

EST - do inglês: *Expressed Sequence Tag*

PTC - Códon De Terminação Precoce (do inglês: *Premature Termination Code*).

NMD – Degradação Mediada Por Mutação Sem Sentido (do inglês: *Nonsense Mediated Decay*).

SS – Sítio de *Splicing.*

snRNAs - Pequenos RNAs Nucleares (do inglês: *Small Nuclear RNA*).

U2 - Fator Auxiliar de U2 (do inglês: *U2 Auxiliary Fator).*

SF1 – Fator de *Splicing* 1, (do inglês: *splicing fator 1*).

ESE – Sítio intensificador de *splicing* no éxon

ESS – Sítio silenciador de *splicing* no éxon.

ISE – Sítio intensificador de *splicing* no íntron.

ISS – Sítio silenciador de *splicing* no íntron.

EJC – Complexo de Junção entre os Éxons (do inglês: *Éxon Junction Complex*).

RBP - Proteínas de Ligação ao RNA (do inglês: *RNA-binding Protein*).

ClC-1 – Canal de cloreto músculo específico (do inglês: *muscle-specific chloride channel*).

UPF(1-3) - do inglês, *UP frameshift Protein*(1 -3).

SMG(2-4) – do inglês, *Suppressor with Morphogenetic effect on Genitalia*(2-4).

SURF - Fator de Liberação Smg1-Ufp1(do inglês: *Smg-1-Upf1-release factor*).

eRF1 e 3 – Fator De Liberação Eucariótico (do inglês, *Eukaryotic Release Fator*).

DCP2 – Proteína de Decapeamento2 (do inglês, *Decapping mRNA 2*).

UTR – Região Não Traduzida (do inglês: *Untranslated Region*).

PABPC1 - Proteína de Ligação à Cauda Poli(A) 1 (do inglês, *poly(A)-binding protein 1*)

DMD – Distrofina.

COL10A1 - cadeia de colágeno alfa-1 (X) (do inglês, *Collagen Type X Alpha 1 Chain*).

FBN1 – Fibrilina 1.

TP53 – do inglês, *Tumor Protein* 53 kDa.

EphB2 - Receptor 2 Da Efrina Tipo B (do inglês, *Ephrin type-B receptor 2*).

NCBI - Centro Nacional de Informação Biotecnológica (do inglês, *National Center for Biotechnology Informatio*n).

PCR – Reação em Cadeia da Polimerase (do inglês*, Polymerase Chain Reaction*).

HAVANA – Análise e Anotação de Humano e Vetebrados (do inglês, *Human and Vertebrate Analysis and Annotation).*

TSL – do inglês, *Transcript Support Level.*

HLA – Antígeno Leucocitário humano (do inglês: *Human Leukocyte Antigen*).

IGV – do inglês, *Integrative Genomics Viewer.*

cDNA – DNA complementar (do inglês, *Complementary DNA*).

ncRNA – RNA não-codificador (do inglês, Non-coding RNA).

BLAT - do inglês, *BLAST-like alignment Tool*.

lincRNAS - Longos RNAS não codificadores integênicos (do inglês: *Long intergenic noncoding RNAs*).

SUMÁRIO

1 INTRODUÇÃO

Com o desenvolvimento das tecnologias de sequenciamento de moléculas biológicas, diversos bancos de dados vêm sendo necessários para agrupar, integrar, evitar redundâncias e facilitar a análise dos dados gerados (HUNT et al., 2018). Por exemplo, em experimentos cujo objetivo é a avaliação do perfil do proteoma de uma amostra por uma abordagem de proteômica *shotgun*, as informações de bancos de dados de peptídeos são importantes para permitir a correta identificação das proteínas (WANG et al., 2012).

É de grande importância que esses bancos de dados contenham informações das diferentes variantes de proteínas, ou seja, proteoformas, geradas por um mesmo gene. Uma forma de produção de proteoformas em células humanas é a partir da tradução de transcritos gerados a partir de transcritos alternativos (BLACK, 2000), sendo o *splicing* alternativo o evento molecular mais abundante (TRESS; ABASCAL; VALENCIA, 2017). Sendo assim, a construção de um banco de dados personalizado que contém informações sobre variantes por *splicing* alternativo e as respectivas proteoformas é fundamental para avaliação de experimentos de proteômica (WANG et al., 2012).

Neste sentido, Tavares e colaboradores (2014) desenvolveram o repositório SpliceProt baseado na metodologia de matrizes ternárias desenvolvida pelo nosso grupo de pesquisa. Esse repositório utilizava dados de sequências de RNAs mensageiros (mRNA) do banco de dados Unigene (PONTIUS; WAGNER; SCHULER, 2003), que agrupa dados dos projetos RefSeq (O'LEARY et al., 2015) e dbEST (BOGUSKI; LOWE; TOLSTOSHEV, 1993) como entrada para predizer variantes por *splicing* alternativo e em seguida realiza a digestão *in silico* dessas variantes (TAVARES et al., 2014). Entretanto, o banco de dados Unigene foi descontinuado em julho de 2019, o que resultou na necessidade da atualização dos dados de entrada para o repositório.

É sabido que eventos de *splicing* alternativo, ainda que fortemente regulados, podem resultar em transcritos com erros nas suas sequências (FAUSTINO; COOPER, 2003). Esses erros podem produzir variantes que ao serem traduzidas, podem produzir proteínas truncadas com funções alteradas e até mesmo prejudiciais ao organismo (FAUSTINO; COOPER, 2003). Um exemplo de *splicing* defeituoso é a inserção de um códon de terminação no transcrito. Esse códon é denominado códon

de terminação precoce (do inglês: *Premature Termination Code*, PTC) quando estiver presente a uma distância maior que 50 nucleotídeos à montante da última junção entre os éxons do transcrito (DA COSTA; MENEZES; ROMÃO, 2017).

Entretanto, transcritos que contém PTC podem ser identificados e degradados pela ação da via de degradação mediada por mutação sem sentido (do inglês: *Nonsense Mediated Decay*, ou NMD), evitando a tradução de uma proteína aberrante (DA COSTA; MENEZES; ROMÃO, 2017; MILLER; PEARCE, 2014). Visto isso, é de grande relevância que esses transcritos não sejam traduzidos e inseridos em bancos de dados que sejam utilizados para a análise de dados de proteômica *shotgun* (CHANG; IMAM; WILKINSON, 2007).

A presença de um PTC pode estar associada a diferentes eventos além do *splicing* alternativo, como por exemplo, mutações, rearranjo gênico e pseudogenes. Ainda que o foco deste trabalho seja a identificação de transcritos que possuem PTC após a predição dos eventos de *splicing* alternativo realizada pelo SpliceProt, é valido destacar que o programa escolhido, NMDClassifier (HSU; LIN; CHEN, 2017), não faz distinção entre a origem do PTC. Ainda mais, o programa NMDClassifier realiza a predição do possível evento que levou a inserção do PTC, tomando como premissa a inserção via evento de *splicing* alternativo.

1.1 OBJETIVOS

1.1.1 Objetivo geral

O objetivo deste trabalho é identificar computacionalmente variantes por *splicing* alternativo no transcriptoma humano que contenham códon de parada prematuro com potencial de serem degradadas pela via de NMD no conjunto de dados do repositório SpliceProt (TAVARES et al., 2014).

1.1.2 Objetivos específicos

1. Atualizar o repositório SpliceProt para receber dados de sequências provenientes do Ensembl.
2. Utilizar a metodologia de matrizes ternárias para identificar variantes por *splicing* alternativo do conjunto de dados do Ensembl (ZERBINO et al., 2018).

3. Identificar entre as variantes preditas pelo SpliceProt, aquelas que contenham PTC e que podem ser suscetíveis a via de NMD.

4. Criar um subconjunto do SpliceProt sem variantes sensíveis a via de NMD para ser utilizado na tradução *in silico*.

2 REVISÃO DE LITERATURA

2.1 PROCESSO DE *SPLICING*

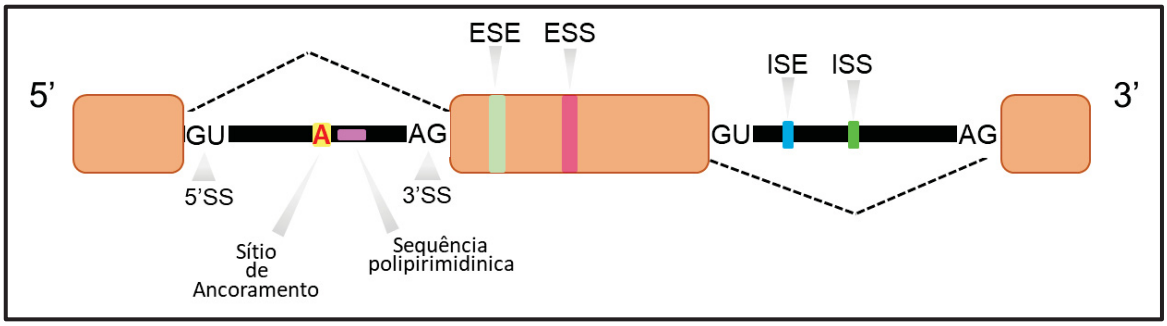
As tecnologias de sequenciamento de alta vazão, associadas à bioinformática, contribuíram para a identificação funcional e estrutural de genes e seus transcritos. Essas abordagens permitiram a avaliação em larga escala da expressão de RNAs em diversos tecidos e organismos (LEE; WANG, 2005). Tal avanço permitiu observar que há diferença entre o número de genes presentes na espécie humana e a diversidade de mRNA que podem ser gerados. A estatística mais recente do GENCODE (FRANKISH et al., 2019) para a espécie humana, identificou 19.965 genes codificadores de proteína e um número de transcritos oriundos destes genes superior a 83.000 (FRANKISH et al., 2019).

O aumento na diversidade dos transcritos está associado ao processamento do RNA após a transcrição, denominado *splicing* alternativo, que será mais detalhado no item 2.2 (FRANKISH et al., 2019). O processo de *splicing* consiste na remoção dos íntrons de um pré-RNA através da identificação de sinais no transcrito e posterior agrupamento dos éxons. Em eucariotos, esse processo é coordenado por um complexo ribonucleoprotéico denominado spliceossomo (WANG et al., 2015). O spliceossomo é formado por diferentes proteínas e pequenos RNAs nucleares (do inglês: Small Nuclear RNA, ou snRNAs). Duas estruturas são reconhecidas como spliceossomo, o spliceossomo principal (do inglês: *Major*) e spliceossomo secundário (do inglês: *Minor*). Ainda que responsáveis por executar funções semelhantes, essas estruturas variam em sua composição de snRNAs e consequentemente o sítio de reconhecimento no pré-RNA. Dentre outras diferenças, o spliceossomo secundário está geralmente associado a processamentos no citoplasma tendo como alvo uma restrita classe de íntrons/RNAs, enquanto os spliceossomo principal está envolvido no processamento do pré-RNA e produção de diferentes isoformas (VAN DEN HOOGENHOF; PINTO; CREEMERS, 2016; WANG et al., 2015).

A atuação do spliceossomo é coordenada pela ação de elementos *cis* e *trans*. Os elementos *cis* incluem sequências intensificadoras e silenciadoras presentes nos éxons e nos íntrons de um pre-RNA. Elementos *trans*, como proteínas de ligação ao RNA (do inglês: *RNA-binding Protein*, ou RBP), reconhecem sítios *cis* na estrutura do pré-RNA e recrutam o spliceossomo para a região. A interação de elementos *trans* com elementos *cis* é responsável pela regulação do processo de *splicing* e determinação dos sítios de *splicing* (SS) (VAN DEN HOOGENHOF; PINTO; CREEMERS, 2016; WANG et al., 2015) (FIGURA 1).

Os SS são sequências de dinucleotídeos conservadas na espécie humana, são denominadas sítio doador de *splicing* 5’ (GU) e sítio aceptor de *splicing* 3’ (AG). Os SS estão presentes nas extremidades dos íntrons e são responsáveis por definir os éxons do transcrito. Em suma, a ação do spliceossomo inicia com o reconhecimento e ligação do snRNA U1 no sítio doador 5’ e do fator de *splicing* SF1 à sequência de ramificação no íntron (CHEN; MANLEY, 2009). A sequência de ramificação (do inglês: *branch sequence*) é uma sequência que geralmente contém uma alanina e é anterior a uma sequência polipirimidínica, estas sequências estão associadas com a excisão a extremidade 5’ do íntron e formação de uma estrutura em laço do íntron removido (KASTNER et al., 2019). Em seguida há a ligação do fator auxiliar U2 (do inglês: *U2 Auxiliary Fator*, ou U2AF) com o sítio aceptor 3’. Então, SF1 é substituído pelo snRNA U2 na sequência de ramificação e em seguida são atraídas as demais subunidades do spliceossomo até que se forme a estrutura catalítica completa. A sucessão do processo de *splicing* resulta na remoção do íntron formando uma estrutura em laço (CHEN; MANLEY, 2009; VAN DEN HOOGENHOF; PINTO; CREEMERS, 2016). As extremidades 3’ e 5’ dos éxons adjacentes são aproximadas e unidas definindo uma região de junção entre os éxons. Ao final do processo, o spliceossomo se dissocia e os seus componentes são reciclados (VAN DEN HOOGENHOF; PINTO; CREEMERS, 2016).

FIGURA 1 – ESQUEMA REPRESENTATIVO DE UM PRÉ-RNA INDICANDO OS ELEMENTOS DE REGULAÇÃO DO PROCESSO DE *SPLICING*



FONTE: O autor (2020).

LEGENDA: Os íntrons estão representados pela linha preta contínua. Éxon estão representados pelos retângulos preenchidos em laranja. 5'SS – Sítio de *splicing* 5'. 3'SS – Sítio de *splicing* 3'. ESE – Sítio intensificador de *splicing* no éxon. ESS – Sítio silenciador de *splicing* no éxon. ISE – Sítio intensificador de *splicing* no íntron. ISS – Sítio silenciador de *splicing* no íntron. O nucleotídeo Adenosina em vermelho representa o sítio de ancoramento que está seguido da sequência polipirimidínica (retângulo rosa). As linhas descontínuas representam o processo de *splicing* o pre-RNA indicado. (Adaptado de (CHEN; MANLEY, 2009; WANG; BURGE, 2008).

O complexo de junção entre os éxons (do inglês: *Éxon Junction Complex*, ou EJC) é um complexo proteico depositado pelo spliceossomo aproximadamente a 20 nucleotídeos à montante da junção entre os éxons no mRNA nascente (IDEUE et al., 2007). Esse complexo é formado por diversas proteínas, sendo quatro principais e centrais, eIF4A3, MAGOH, Y14 e MLN51. O EJC participa na interação com demais eventos pós-transcricionais como exportação do RNA, tradução e localização e na estabilidade de eventos de controle como o a via de NMD (ZHENG et al., 2018). No momento que ocorre a tradução, a interação do ribossomo com as proteínas do EJC resulta da dissociação desse complexo do RNA (ZHENG et al., 2018).

2.2 *SPLICING* ALTERNATIVO

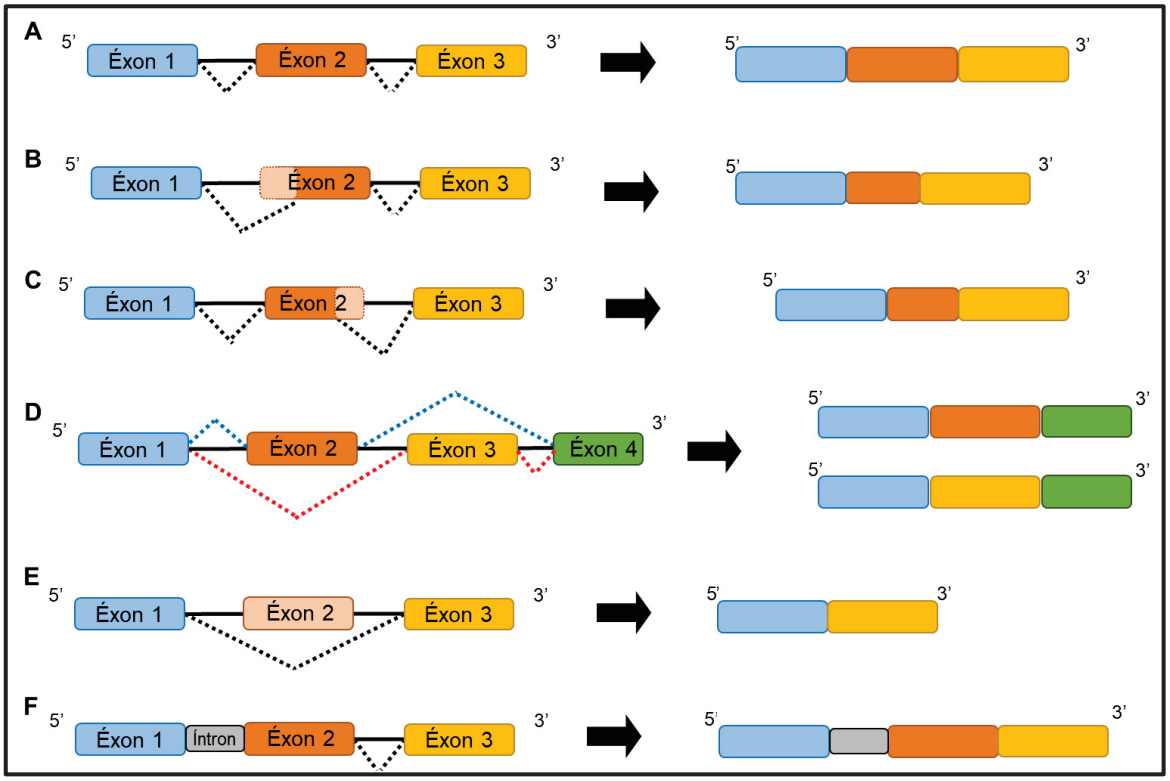
Como dito no item 1.1, o *splicing* é regulado por elementos *cis*, como sequências intensificadoras e silenciadoras de *splicing*. A interação de diferentes proteínas com diferentes desses elementos resulta na utilização de SS não canônicos, ou seja, alternativos. A utilização de SS alternativos durante o processo de *splicing* promove a formação de variantes de RNA para um mesmo gene. Tal processo é dito

como *splicing* alternativo (CHEN; MANLEY, 2009). Estima-se que cerca de 90% dos genes com mais de um éxon sofram *splicing* alternativo (CHEN; MANLEY, 2009).

Existem diferentes tipos de *splicing* alternativo (FIGURA 2), o que resulta na produção de diferentes transcritos, que podem ou não ser traduzidos e produzir diferentes proteoformas para um mesmo gene. Dentre os eventos de *splicing* alternativo podemos citar: uso alternativo de éxon, quando um éxon é completamente removido do transcrito; éxons mutualmente exclusivos, quando a presença/ausência de um éxon é dependente da ausência/presença de outro éxon; uso de SS 5' alternativo, quando há um aumento ou encurtamento do éxon associado ao sítio SS 5'; uso de SS alternativo 3', quando há um aumento ou encurtamento do éxon associado ao sítio SS 3'; retenção de íntron, quando um íntron é retido no transcrito; uso alternativo de éxon inicial e uso alternativo de éxon final (FIGURA 2) (WANG; BURGE, 2008). É importante destacar que um mesmo transcrito pode sofrer diferentes eventos de *splicing* alternativo (WANG et al., 2008). Por exemplo, o gene SLC25A3 possui 17 variantes conhecidas geradas por eventos de *splicing* diferentes. A variante SLC25A3-204 é gerada a partir de um evento de retenção de íntron entre o éxon 2 e o éxon 3, já a variante SLC25A3-211 está associada um evento de uso alternativo de éxon (WANG et al., 2008; ZERBINO et al., 2018).

O mecanismo de *splicing* alternativo pode gerar diversas combinações de transcritos. Entretanto, os dados apresentados pelo projeto GENCODE indicam que o número de traduções distintas observadas é inferior a 62.000 (FRANKISH et al., 2019). Ou seja, somente uma parte dos transcritos codificadores de proteína são efetivamente traduzidos (total ou parcialmente).

FIGURA 2 – PRINCIPAIS EVENTOS DE *SPLICING*



FONTE: O autor (2020).

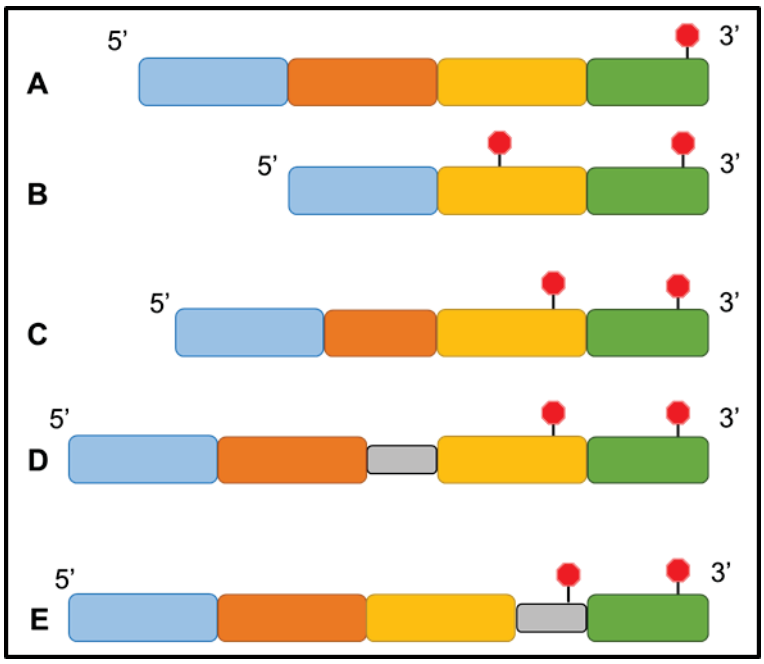
LEGENDA: **A** - *splicing* canônico. **B-F** Eventos de *splicing* alternativo. **B** – Uso de sítio de *splicing* 3' alternativo. **C** - Uso de sítio de *splicing* 5' alternativo. **D** – Éxons mutualmente exclusivos. **E** – Uso alternativo de éxon. **F** - Retenção de íntron. (Adaptado de (CHEN; MANLEY, 2009; WANG; BURGE, 2008).

Nem todas as variantes produzidas por *splicing* alternativo são viáveis para desenvolver uma função biológica (TRESS; ABASCAL; VALENCIA, 2017). Algumas dessas variantes podem não apresentar potencial codificador ou até mesmo desempenhar uma função prejudicial para o organismo (FAUSTINO; COOPER, 2003). Além de mutações, os eventos de *splicing* também podem gerar alterações na sequência do RNA maduro resultando em eventos que gerem impacto na tradução (WANG; LEE, 2018). A alteração de um sítio SS 5', por exemplo, pode mudar a fase de leitura de um transcrito codificador, o que pode resultar no reconhecimento de um códon de terminação não canônico no transcrito, resultando em uma proteína truncada (FAUSTINO; COOPER, 2003). Quando esse códon de terminação está presente há uma distância superior a 50 nucleotídeos à montante da última junção entre os éxons ele é denominado PTC (KUROSAKI; MAQUAT, 2016). Por exemplo, a perda de função dos canais de Cloro músculo-específico (do inglês, *muscle-specific*

chloride channel, ou CIC-1) observada na Distrofia miotônica está associada a inserção de um PTC na sequência do pré-mRNA (FAUSTINO; COOPER, 2003).

A perda de função dos CIC-1 está associada a degradação do RNA devido a ação de mecanismos de controle. Esses mecanismos identificam e levam transcritos que possuem PTC à degradação, impedindo que os mesmos sejam traduzidos (FAUSTINO; COOPER, 2003). A via de NMD atua na degradação amplamente estudada que atua tais transcritos (DA COSTA; MENEZES; ROMÃO, 2017). Na FIGURA 3 estão resumidos diferentes transcritos que possuem PTC e podem ou não ser sensíveis a via canônica de NMD.

FIGURA 3 – ESQUEMA REPRESENTATIVO DE TRANSCRITOS CONTENDO UM CÓDON DE TERMINAÇÃO ALTERNATIVO GERADO POR EVENTOS DE *SPLICING*



FONTE: O autor (2020).

LEGENDA: **A** – códon de terminação no transcrito canônico, não ativa a via canônica de NMD. **B** – PTC inserido no éxon 3 devido a um evento de uso alternativo do éxon 2. **C** - PTC inserido no éxon 3 devido a uso de um sítio 3' SS e encurtamento do éxon 2. **D** – PTC inserido no éxon 3 devido uma retenção de íntron entre os éxon 2 e 3. **E** - Códon de terminação alternativo presente no íntron retido entre o éxon 3 e 4. **B-D** – Eventos de inserção de PTC nos transcritos, presentes a mais de 50 nt à montante da última junção entre os éxons ativam a via de NMD. **E** - Códon de terminação alternativo não ativa a via canônica de NMD devido à ausência de junção entre o éxon à jusante (adaptado de GE; PORSE, 2014).

2.3 MECANISMO DE CONTROLE DA VIA DE NMD

A atuação da via de NMD acontece no citoplasma após a exportação do RNA maduro do núcleo. A ativação canônica dessa via, está associada a presença de um EJC remanescente no RNA maduro. Neste caso, a presença do EJC é resultado da presença de um PTC que impede que o ribossomo remova o complexo (DA COSTA; MENEZES; ROMÃO, 2017).

Em suma, a ativação da via de NMD está associada, principalmente, a duas famílias de proteínas conservadas em eucariotos, as UPF (do inglês, *UP frameshift*) e as SMG (do inglês, *Suppressor with Morphogenetic effect on Genitalia*) (SMITH; BAKER, 2015). Quando um RNA maduro é exportado do núcleo para o citoplasma, as proteínas UPF2 e UPF3 se associam ao EJC. No citoplasma, o RNA maduro contendo um PTC é reconhecido pelo ribossomo, quando o PTC se localiza no sítio A do ribossomo ele indica o término precoce da tradução e recruta um complexo de proteínas denominado SURF (do inglês, Smg-1-Upf1-release factor), que contém as proteínas SMG1, UPF1 e as eRF1 e 3 (do inglês, *eukaryotic release factor*) (HE; JACOBSON, 2015; KUROSAKI; MAQUAT, 2016). A presença de um EJC a mais de 50 nucleotídeos à jusante do PTC resulta na interação de UFP1 com UFP2 e 3 o que promove a fosforilação de UPF1 mediada por SMG1 e consequentemente a ativação da via de NMD (HE; JACOBSON, 2015). UFP1 fosforilada recruta SMG6, 5 e 7 e DCP2 (do inglês, *Decapping mRNA 2*). SMG6 atua na clivagem do RNA na proximidade do PTC gerando dois fragmentos de RNA. O heterodímero SGM5-SGM7 promove a remoção da cauda poli-A da extremidade 3’ do RNA e DCP2 atua na remoção do cap5’ da extremidade 5’ do RNA, o que torna o RNA exposto a ação de exonucleases no citoplasma (HE; JACOBSON, 2015).

Além da interação com EJC, outros sinais podem desencadear a ativação da via de NMD por uma via alternativa. A proteína UPF1 é conhecida por ter a capacidade de se ligar de forma inespecífica a regiões do RNA maduro. Normalmente, proteínas UPF1 remanescentes no transcrito são removidas através do processo de tradução pelo ribossomo. Entretanto, transcritos que apresentam longas regiões 3’-UTR (do inglês, *Untranslated Region*), resultantes ou não da presença de um PTC, podem possuir proteínas UPF1 associadas que possuem potencial para serem ativadas e desencadearem a via de NMD (HE; JACOBSON, 2015; KUROSAKI; MAQUAT, 2016). Outro fator associado a longas regiões 3’ UTR e ativação da via de NMD é a distância

das proteínas eRFs da proteína PABPC1 (do inglês, *poly(A)-binding protein 1*). Uma longa distância impede a interação entre essas proteínas, o que resulta em um término de tradução ineficiente aumentando a interação de UPF1 com as eRFs resultando em eventos que ativam a via de NMD, caracterizando o modelo denominado faux 3' UTR (SMITH; BAKER, 2015).

A via de NMD é de grande importância devido ao potencial de eliminar transcritos que poderiam ser traduzidos em proteínas truncadas, mas também, pode participar da regulação de alguns genes (HE; JACOBSON, 2015). Por exemplo, alterações no gene DMD (Distrofina) que tornem o mRNA sensível a via de NMD já foram associadas com desordens musculares (KERR et al., 2001). Variantes com a presença de PTC também foram encontradas em Condrodisplasia Metafisária do tipo Schmid, como é o caso de uma mutação no gene COL10A1 (cadeia tipo X do colágeno) (CHAN et al., 1998). Mutações no gene FBN1 (Fibrilina-1) que resultem termino precoce da tradução estão relacionadas com fenótipos clínicos da Síndrome de Marfan (SCHRIJVER et al., 2002).

Além disso, a falha no mecanismo de NMD está associada a diversas patologias. Xu e colaboradores (2012), identificaram uma mutação no gene UPF3, que leva a inserção de um PTC no transcrito e consequentemente à redução dos níveis do RNA. Essa mutação foi associada como fator etiológico de retardo mental familiar em uma população chinesa. Ademais, os autores observaram que a redução dos níveis de UPF3 afeta a ativação canônica da via de NMD, mas não a via alternativa (XU et al., 2013). Liu e colaboradores (2014) reportaram que mutações no gene UPF1 que resultem na perda da função da proteína, impactam na ativação da via de NMD e estão associadas a carcinoma adenoescamoso pancreático. Essa mutação no gene UPF1, resulta na manutenção de uma variante do gene TP53 que contém um PTC no íntron 6 que é mantido no transcrito devido a um evento de *splicing* alternativo (LIU et al., 2014).

É válido destacar que, estratégias relacionadas ao mecanismo de NMD estão sendo empregadas em análises de algumas doenças, entre elas o câncer. Huusko e colaboradores (2004) após a inativação da via NMD em linhagens celulares, identificaram o gene EphB2 como candidato a supressor de tumor em câncer de próstata (HUUSKO et al., 2004). Desta forma, a identificação de transcritos que possuem PTC e são suscetíveis a degradação via NMD é de grande importância na

busca de possíveis biomarcadores em diversas doenças, podendo também contribuir para identificação de falhas nesse mecanismo (NOENSIE; DIETZ, 2001).

2.4 TRANSCRITOS ALVOS DA VIA DE NMD EM BANCO DE DADOS DE SEQUÊNCIAS BIOLÓGICAS

O estudo da via de NMD, bem como a identificação de transcritos que podem ser alvo desta via merecem destaque. Tal informação é relevante em diversos estudos, principalmente os que dependem da utilização de bancos de dados como referência (LAU et al., 2019; WANG et al., 2012). Devido a sua alta importância, bancos de dados de sequências biológicas como NCBI, Ensembl e UniProt possuem identificação para transcritos que são alvo da via de NMD (O'LEARY et al., 2015; THE UNIPROT CONSORTIUM, 2019; ZERBINO et al., 2018).

Além disso, diversos programas já foram desenvolvidos para a identificação computacional de transcritos sensíveis a via de NMD a partir de dados de sequências biológicas (DE LA GRANGE et al., 2007; HSU; LIN; CHEN, 2017; RYAN et al., 2008; VITTING-SEERUP et al., 2014). O repositório SpliceCenter (RYAN et al., 2008) por exemplo é uma interface online que agrupa diferentes programas para análise de microarranjos, experimentos com RNA de interferência, PCR em tempo real, entre outros, e possui classificação para sensibilidade a via de NMD (RYAN et al., 2008). O banco de dados fastdb2 também reúne diferentes ferramentas para avaliação de dados provenientes de experimentos de transcriptômica com destaque para avaliação de variantes por *splicing* alternativo e possui em sua interface a funcionalidade de predição de variantes suscetíveis a degradação via NMD (DE LA GRANGE et al., 2007). O SpliceR é um pacote disponível no repositório Bioconductor capaz de avaliar dados de sequenciamento de RNA (RNAseq) e identificar eventos de *splicing* alternativo, bem como identificar variantes sensíveis a degradação via NMD (VITTING-SEERUP et al., 2014). O programa NMDClassifier é um programa em linguagem Perl desenvolvido com o foco de identificação de variantes por *splicing* alternativo sensíveis a degradação por NMD. Além da predição das potenciais variantes alvo da via de NMD ele é capaz de indicar o evento de *splicing* alternativo associado a inserção de um PTC na sequência (HSU; LIN; CHEN, 2017).

Ainda que existam diferentes programas disponíveis que realizem a predição computacional de transcritos que são potenciais alvos da via de NMD, o método

principal e o estado da arte utilizado para esta predição é a identificação de um PTC presente no transcrito (HSU; LIN; CHEN, 2017; VITTING-SEERUP et al., 2014). Como consequência, esses programas possuem ênfase na identificação de mRNAs com potencial codificador suscetíveis a via canônica de ativação de NMD (HE; JACOBSON, 2015; KUROSAKI; MAQUAT, 2016).

2.5 PROTEOGENÔMICA E ALVOS DA VIA DE NMD

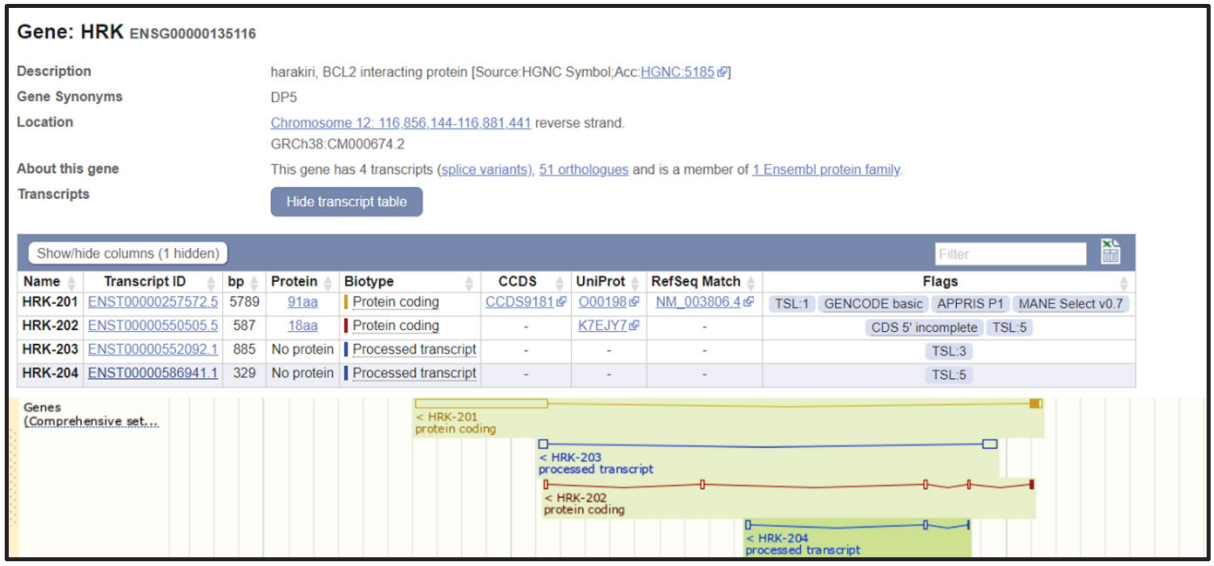
A Proteogenômica é uma área que engloba a proteômica e genômica. Por exemplo, ao analisar espectros de peptídeos de determinada amostra é possível mensurar o nível de expressão proteica de determinado gene (NESVIZHSKII, 2014), para experimentos de proteômica em larga escala, a identificação dos peptídeos necessita da construção e/ou utilização de bancos de dados personalizados (NESVIZHSKII, 2014). Desta forma, é necessário que esses bancos de dados possuam informações a respeito de variantes por *splicing* alternativo associadas ao potencial codificador desta variante. Além disso é de grande importância a identificação de variantes suscetíveis a NMD, uma vez que a presença de um PTC em um mRNA irá leva-lo a degradação impedindo que ele seja traduzido em alguma proteína (LIU et al., 2017; WANG et al., 2012). Sendo assim, tais identificações necessárias a fim de permitir uma avaliação mais verossímil em comparação com o sistema biológico estudado (WANG et al., 2012).

2.5.1 O projeto Ensembl

O projeto Ensembl, por exemplo, possui diferentes bancos de dados com informações para genes de diferentes espécies, bem como para as suas variantes por *splicing* alternativo (FIGURA 4). Nessa base de dados as informações para as diferentes variantes são geradas via anotação automática, ou anotadas manualmente pelo projeto HAVANA (do inglês, Human and Vertebrate Analysis and Annotation) (http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/). Para tal, são considerados diferentes tipos de dados como sequências completas do RNA e também sequências de EST (do inglês: *Expressed Sequence Tag*), que são fragmentos de sequências de determinado transcrito derivados de um sequenciamento randômico. Informações de EST podem verificar a expressão de

determinado transcrito, mas não necessariamente são capazes de capturar a informação da sequência total do transcrito (PARKINSON; BLAXTER, 2009). Na FIGURA 4 está ilustrado as informações para o gene HRK, bem como as respectivas variantes por *splicing* alternativo. Tais informações também estão disponíveis nos arquivos de anotação do genoma e no arquivo de sequência em formato fasta.

FIGURA 4 – ENTRADA PARA O GENE HRK NA INTERFACE ONLINE DO PROJETO ENSEMBL



FONTE: Ensembl (2020).

LEGENDA: Informações sobre o gene HRK (ENST00000135116) disponíveis na interface online do projeto Ensembl (https://bit.ly/3cAMmhe). Nesta interface é possível obter informações sobre a anotação do gene e suas respectivas variantes por *splicing*, assim como a alinhamento e disposição dos éxons dos transcritos.

Para espécie humana, todas as informações disponíveis são combinadas e em caso de conflito a anotação manual fornecida pelo HAVANA é priorizada. Na base de dados do Ensembl, os transcritos podem ser classificados como codificadores de proteína, não-codificadores de proteína, RNAs que são suscetíveis a vias de degradação, RNAs que possuem íntrons retidos, pseudogenes, entre outros (ZERBINO et al., 2018) (FIGURA 4). Essa informação também pode ser encontrada em outros bancos de dados como RefSeq e SwissProt (O'LEARY et al., 2015; THE UNIPROT CONSORTIUM, 2019).

Mesmo em bancos amplamente utilizados na literatura, a anotação de diversos transcritos alternativos de um mesmo gene pode ser deficiente. Para a versão 96 do projeto Ensembl para *Homo sapiens* (12 de março de 2019), de 208.460 transcritos anotados, somente 31.782 possuem validação para todas as junções de

splicing, ou seja, possuem TSL valor 1 (do inglês, *Transcript Support Level*). O TSL, é um método de avaliar a confiabilidade das variantes por *splicing* alternativo dos genes baseado nas informações disponíveis para sequência do transcrito. No QUADRO 1, está descrito o significado para cada tipo de TSL adotado pelo Ensembl.

Devido à baixa resolução dos bancos de dados de sequências biológicas em relação a variantes por *splicing* alternativo e suas respectivas proteoformas, em 2014, Tavares e colaboradores utilizaram a metodologia de matrizes ternárias para avaliar determinados transcritos. Sendo assim, foi criado o repositório SpliceProt (TAVARES et al., 2014) que utiliza dados de transcriptômica para predizer eventos de *splicing* alternativo em um conjunto de dados. Além disso, realiza a tradução *in silico* dessas variantes com o objetivo final de gerar um banco de dados de peptídeos personalizados.

QUADRO 1 – SIGNIFICADO DE CADA TIPO DE TSL NO BANCO DE DADOS ENSEMBL

tsl 1	Todas as junções de *splicing* do transcrito são corroboradas por pelo menos 1 RNA confiável.
tsl 2	O transcrito é sustentado por um mRNA não confiável ou por múltiplas ESTs.
tsl 3	O transcrito é corroborado por uma única EST.
tsl 4	O transcrito é sustentado por uma EST não confiável.
tsl 5	Não há suporte para o modelo estrutural do transcrito.
tsl NA	O transcrito não foi analisado devido a: ser uma anotação de pseudogene; transcrito proveniente de gene HLA; transcrito proveniente de gene de imunoglobulina; transcrito do receptor de célula T; transcrito com apenas um éxon.

TSL - do inglês: *Transcript Support Level*; mRNA – RNA mensageiro; EST – do inglês: *Expressed Sequence Tag*; HLA – do inglês: *Human Leukocyte Antigen.*

FONTE: Ensembl release 98 - September 2019 ©

O SpliceProt foi contruído como um repositório online disponível gratuitamente composto por variantes por *splicing* alternativo a partir de dados de transcriptômica (TAVARES et al., 2014). Quando foi criado, em 2014, sua metodologia foi baseada em sequências obtidas do projeto UNIGENE (PONTIUS; WAGNER; SCHULER, 2003) que compunha dados de ESTs e dados do repositório RefSeq. As variantes por *splicing* foram avaliadas utilizando a metodologia de matrizes ternárias e a tradução dos transcritos foi realizada pelo software TRANSEQ (RICE; LONGDEN; BLEASBY, 2000; TAVARES et al., 2014). Entretanto, foi observado que variantes que possuíam PTC e eram potencialmente alvo da via de NMD, também eram traduzidas. A tradução

desses transcritos representava um viés na representatividade de sistemas biológicos, uma vez que a via de NMD levaria esses mRNAs à degradação via exonucleases e impede a sua tradução.

Pelo exposto, a fim de se obter um banco de dados com uma maior representação de eventos biológicos é necessário identificar variantes que possuam PTC e sejam suscetíveis a degradação via NMD. Além disso, como citado anteriormente, essas variantes podem estar associadas a diversas doenças. Desta forma a identificação de variantes por *splicing* alternativo sensíveis a via de NMD no conjunto de dados do SpliceProt permite aumentar as funcionalidades disponibilizadas para o usuário permitindo que o repositório seja utilizado em estudos de busca de novos candidatos a biomarcadores e estudos que tenham como foco transcritos que apresentem um PTC.

Desta forma, a origem dos dados de entrada para construção do repositório SpliceProt, foi modificada, uma vez que o banco de dados utilizado por Tavares e colaboradores (2014), o Unigene (PONTIUS; WAGNER; SCHULER, 2003) foi descontinuado. Em sequência, foi realizado a identificação das variantes por *splicing* alternativo com potencial para serem alvo da via de NMD. A seguir (item 3), está descrita a estratégia empregada para realização deste trabalho.

3 MATERIAL E MÉTODOS

Originalmente o foco deste projeto era aprimorar a etapa de tradução *in silico* do repositório SpliceProt (TAVARES et al., 2014), adicionando uma etapa anterior de filtragem para variantes de *splicing* que seriam suscetíveis a degradação pela via de NMD por possuírem PTC. Entretanto, foi necessário a reestruturação do banco de dados a fim de receber dados atualizados disponíveis na literatura.

3.1 SPLICEPROT 2014

A primeira versão do repositório SpliceProt foi desenvolvida por Tavares e colaboradores (2014). Inicialmente os dados de transcritos humanos foram obtidos a partir do repositório Unigene (versão 228) assim como do repositório SRA (do inglês: *Sequence Read Archive*). Dados do projeto dbEST (BOGUSKI; LOWE; TOLSTOSHEV, 1993) foram usados para compor os dados de entrada e aumentar o

potencial em identificar variantes de splicing, uma vez que esses dados poderiam conter informações de ESTs referentes a diferentes variantes por *splicing* alternativo. Dados de Transcriptômica e Proteômica do projeto RefSeq (O'LEARY et al., 2015) foram utilizados como parâmetro de confiabilidade para fins de comparação. As sequências utilizadas como dados de entrada foram alinhadas com o genoma humano (versão hg19) utilizando o programa SIBsim4 (FLOREA et al., 1998). Nesse método, transcritos com apenas 1 éxon foram excluídos da análise.

A identificação de variantes por *splicing* alternativo foi realizada através do método das matrizes ternárias (TAVARES et al., 2014). As matrizes foram construídas a partir da definição de coordenadas confiáveis do transcrito. Para transcritos que estavam representados no RefSeq, todas as coordenadas foram consideradas como confiáveis. Entretanto, para os demais transcritos, associados a dados de EST, apenas as coordenadas internas dos éxon foram utilizadas para construção das matrizes.

Para cada gene presente na base relacional, foi criado uma matriz ternária de tamanho NxM, em que N representa o número de total de transcritos e M o número de total de íntrons e éxons para o respectivo gene. Para tal, éxons são definidos por uma região gênica presente em pelo menos um transcrito, enquanto íntrons são definidos por regiões genômicas ausentes em todos os transcritos da respectiva região gênica. A representação matricial dos transcritos é dada por 3 caracteres, sendo "1" a presença de determinada região do éxon, "0" a ausência da região do éxon e "|" o limite entre os éxons (FIGURA 4)(TAVARES et al., 2014).

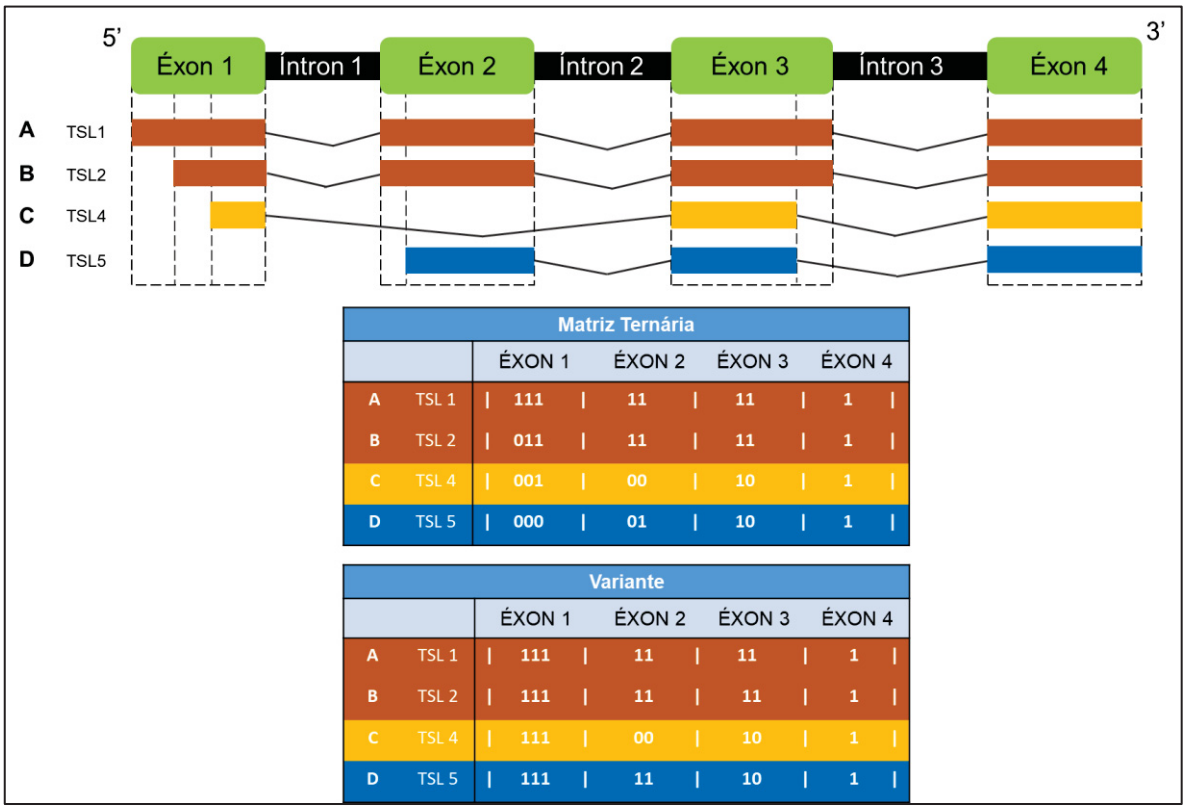
A partir das matrizes ternárias de cada transcrito foi construído uma matriz denominada 'variante', que agrupa os eventos de *splicing* associados ao transcrito. Um transcrito é definido como uma variante de *splicing* quando são encontrados um ou mais caracteres "0" flanqueados pelo caractere "1" (TAVARES et al., 2014).

Adicionalmente, Tavares e colaboradores (2014) também criaram um programa em Perl que identificava os padrões de splicing alternativos preditos pela metodologia matricial ternária.

Os programas PASA (HAAS et al., 2003), AUGUSTUS (STANKE et al., 2006) and TRANSEQ (RICE; LONGDEN; BLEASBY, 2000) foram avaliados quanto a eficiência para realizar a tradução *in silico* das variantes identificadas a partir de metodologia empregada. Devido a sua curadoria manual, os dados do repositório Refseq que possuíam identificadores NM (mRNA validado experimentalmente) e NP

(sequência de aminoácidos validada experimentalmente) foram utilizados para avaliação dos programas citados (O'LEARY et al., 2015). O programa TRANSEQ apresentou a melhor performance de tradução, sendo capaz de traduzir todos os transcritos NM que foram utilizados nessas análises, com mais de 96% de sequências apresentando um valor de identidade maior que 95% e cobertura de 90-100% com as respectivas sequências NP utilizada nessas análises (TAVARES et al., 2014).

FIGURA 5 – ESQUEMA REPRESENTATIVO DA CONSTRUÇÃO DE UMA MATRIZ TERNÁRIA



Matriz Ternária

		ÉXON 1	ÉXON 2	ÉXON 3	ÉXON 4
A	TSL 1	111	11	11	1
B	TSL 2	011	11	11	1
C	TSL 4	001	00	10	1
D	TSL 5	000	01	10	1

Variante

		ÉXON 1	ÉXON 2	ÉXON 3	ÉXON 4
A	TSL 1	111	11	11	1
B	TSL 2	111	11	11	1
C	TSL 4	111	00	10	1
D	TSL 5	111	11	10	1

FONTE: O autor (2020).

LEGENDA: Todos os transcritos são alinhados e as coordenadas identificadas. A matriz ternária é uma representação fiel do transcrito, onde o caractere "|" representa os limites dos éxons, o caractere "0", representa a ausência de determinada região e "1" a presença da região do éxon. A matriz variante representa os eventos de *splicing* alternativo preditos para o gene baseado na lógica de seleção de coordenadas confiáveis baseadas no TSL e/ou tamanho da sequência. Na figura, A e B representam a mesma variante, sem indicação de *splicing* alternativo. O transcrito C apresenta dois eventos de *splicing* alternativo, uso alternativo do éxon 2 e uso alternativo de SS 5'. O transcrito D apresenta somente o evento de *splicing* alternativo de uso alternativo de SS 5'.

Tavares e colaboradores (2014) avaliaram o potencial do repositório SpliceProt em identificar variantes por *splicing* alternativo associados à peptídeos em experimentos de espectrometria de massas. Para isso foi realizada uma digestão *in silico,* pela enzima tripsina, das sequências de proteínas dos repositórios SpliceProt,

RefSeq (O'LEARY et al., 2015), ENSEMBL Gene (ZERBINO et al., 2018) e UniProtKB/Swiss-Prot (THE UNIPROT CONSORTIUM, 2019). Nesse contexto, Foram identificados 181.174 peptídeos não redundantes gerados exclusivamente pelo repositório SpliceProt (TAVARES et al., 2014).

Além disso, Tavares e colaboradores (2014) compararam os peptídeos identificados a partir dos dados do repositório SpliceProt e do UniProtKB/Swiss-Prot utilizando dados de proteômica para uma linhagem de linfócito T disponível publicamente (SHEYNKMAN et al., 2013). Nessa análise, foram identificados 173 peptídeos associados a variantes por *splicing* alternativo, sendo 54 encontrados apenas no repositório SpliceProt, sugerindo que o repositório SpliceProt pode contribuir para a anotação de variantes por *splicing* alternativo em experimentos de proteômica (TAVARES et al., 2014).

Sabendo que eventos de *splicing* alternativo podem gerar variantes que não possuem potencial de serem traduzidas, como aquelas sensíveis a via de NMD, foi observado que era relevante identificar a presença de PTC nos transcritos que faziam parte dos dados de entrada para a criação do repositório SpliceProt (FAUSTINO; COOPER, 2003). Entretanto, o repositório Unigene (PONTIUS; WAGNER; SCHULER, 2003) foi descontinuado em julho de 2019, o que levou a necessidade de reconstruir a estrutura dos dados de entrada do repositório SpliceProt.

Todas as etapas deste projeto foram desenvolvidas utilizando a infraestrutura do núcleo de bioinformática associado ao Laboratório de Regulação da Expressão Gênica do Instituto Carlos Chagas FIOCRUZ-PR. A restruturação do banco de dados SpliceProt foi realizada em conjunto com a doutoranda Letícia Graziela Costa Santos.

Todos os programas utilizados na execução deste projeto foram desenvolvidos em linguagem PERL (versão 5.26.1) e no Sistema de Gerenciamento de Bancos de Dados PostgreSQL (versão 10.1). Os gráficos para comparação entre os arquivos de anotação foram criados utilizando o IGV (versão 2.4.14). O diagrama representando a estrutura relacional do banco de dados foi criado utilizando a plataforma online Ludichart disponível em: https://www.lucidchart.com/pages/.

3.2 RESUMO DO MÉTODO

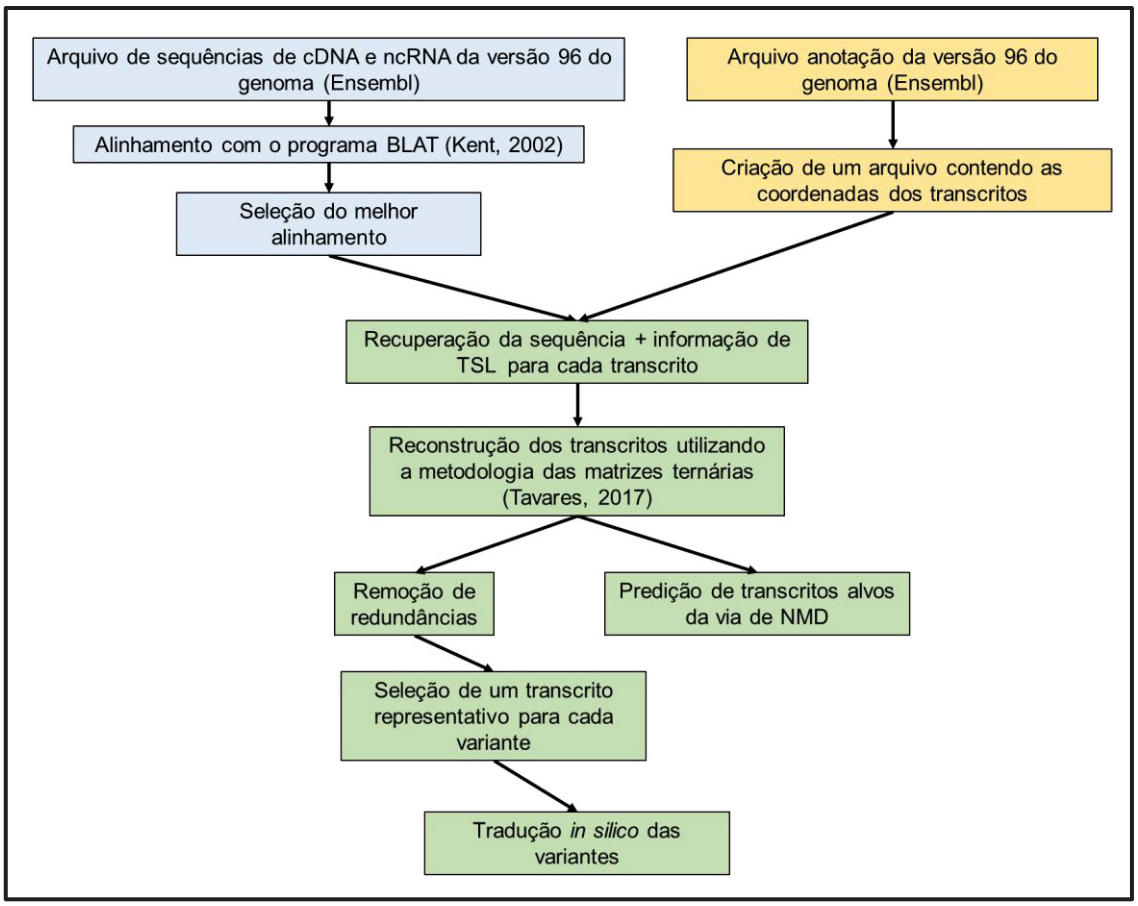
A lógica de construção do banco de dados foi alterada para receber dados de entradas provenientes do banco de dados Ensembl. Foram utilizados como entrada dados de sequência dos transcritos humanos do genoma GRCh38/hg38 do Ensembl

(versão 96) (ZERBINO et al., 2018). Utilizando o programa BLAT (KENT, 2002), a sequência desses transcritos foram alinhadas com o genoma humano para fornecer dados para a construção do banco de dados *Homo_sapiens_hg38*. Também foram utilizados como entrada, dados de anotação para fornecer as coordenadas dos transcritos já anotadas pelo projeto Ensembl e assim, construir o banco de dados *Homo_sapiens_ENSEMBL*. A partir dos dados de anotação do projeto Ensembl, o parâmetro TSL foi utilizado um critério de confiabilidade para definir as coordenadas dos éxons dos transcritos que eram confiáveis para as duas bases de dados. Sendo assim, as informações de coordenadas, TSL e sequência foram utilizadas para reconstruir os transcritos e predizer seus respectivos padrões de *splicing* baseados no método de matrizes ternárias (item 3.1) (TAVARES et al., 2014).

Após a elaboração das matrizes e variantes paras as duas bases de dados, foram identificadas as variantes presentes em cada gene e um programa em Perl foi construído para eliminar as redundâncias presentes. Tal redundância foi removida selecionando um transcrito que melhor representava tal variante. Para isso, os critérios para inclusão foram: anotação como codificador de proteína, menor valor de TSL e maior sequência de nucleotídeos.

As tabelas relacionais sem redundância de variantes para cada gene dos bancos de dados *Homo_sapiens_hg38* e *Homo_sapiens_ENSEMBL* foram utilizadas como dados de entrada para etapa de tradução *in silico* utilizando o método já empregado por Tavares e colaboradores (2014). Na FIGURA 7 está representado o resumo do método empregado neste trabalho para a duas bases de dados.

FIGURA 6 – ESQUEMA REPRESENTATIVO DO MÉTODO EMPREGADO NESTE TRABALHO



FONTE: O autor (2020).

LEGENDA: Em azul, as etapas exclusivas para construção das tabelas relacionais da base de dados *Homo_sapiens_hg38*. Em amarelo, as etapas exclusivas para construção das tabelas relacionais da base de dados *Homo_sapiens_ENSEMBL*. Em verde, etapas compartilhadas para construção das duas bases de dados, que foram executadas individualmente para cada.

3.3 PREPARAÇÃO DOS DADOS DE ENTRADA PARA CONSTRUÇÃO DOS BANCOS DE DADOS

3.3.1 Dados de alinhamento para base relacional *Homo_sapiens_hg38*

A fim de compor o repositório SpliceProt com diferentes sequências para aplicação em futuros projetos do nosso grupo, foram utilizadas sequências provenientes do arquivo de cDNA (ftp://ftp.ensembl.org/pub/release-96/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz) e também de RNAs não codificadores (ftp://ftp.ensembl.org/pub/release-

96/fasta/homo_sapiens/ncrna/Homo_sapiens.GRCh38.ncrna.fa.gz). As sequências utilizadas como dados de entrada foram aquelas disponíveis para a versão 96 do genoma humano Homo_sapiens_GRCh38

Todas as sequências foram concatenadas em um só arquivo no formato fasta. Esse arquivo foi separado em arquivos menores com cerca de 200 sequências cada e o alinhamento dessas sequências foi realizado pelo programa BLAT (KENT, 2002). Para o alinhamento, as sequências dos cromossomos humanos foram utilizadas como referência (Ensembl, Homo_sapiens_GRCh38 Versão 96 - ftp://ftp.ensembl.org/pub/release-96/fasta/homo_sapiens/dna/). A execução do BLAT (KENT, 2002) foi realizada em paralelo para cada arquivo por meio de um programa na linguagem Perl, utilizando os parâmetros -t=dna (tipo da sequência do arquivo de referência), -q=rna (tipo do arquivo a ser alinhado), -out=psl ou -out=blast9. Dois arquivos de alinhamento foram gerados, um em formato psl e outro em formato blast9, uma vez que o arquivo em formato blast9 era o único que fornecia informação sobre a porcentagem de identidade do alinhamento.

Os melhores alinhamentos do arquivo em formato psl foram selecionados utilizando o programa do pacote BLAT, PSLReps (KENT, 2002), utilizando como argumento -singlehit, a fim de capturar apenas o melhor alinhamento para cada transcrito. Entretanto, foi necessário remover as redundâncias que ainda se mantinham no arquivo psl utilizando um programa em linguagem Perl.

3.3.2 Dados provenientes de anotação para base relacional *Homo_sapiens_ENSEMBL*

As informações necessárias para povoar as tabelas do banco SpliceProt para a base relacional *Homo_sapiens_ENSEMBL* foram retiradas do arquivo de anotação do genoma humano disponível no projeto Ensembl (Homo_sapiens_GRCh38 Versão 96 - ftp://ftp.ensembl.org/pub/release-96/gtf/homo_sapiens/Homo_sapiens.GRCh38.96.gtf.gz). Para isso, foi criado um programa em Perl para retirar as informações para cada éxon e como padrão foi adotado uma identidade de 100% para todos transcritos. As coordenadas dos transcritos (início e fim do transcrito e tamanho de cada éxon) foram calculadas considerando as coordenadas cromossômicas de cada éxon e a orientação do transcrito no DNA.

3.4 ADAPTAÇÃO DA ESTRUTURA DO BANCO DE DADOS SPLICEPROT

3.4.1 Criação da estrutura relacional para os bancos de dados

Foi criada a mesma estrutura relacional para ambas as bases de dados, *Homo_sapiens_hg38* e *Homo_sapiens_ENSEMBL*. Como indicado na FIGURA 6, o relacionamento entre as tabelas foi o mesmo. Entretanto, os dados presentes em cada base de dados foram provenientes de diferentes origens (item 1.1). Para a base de dados *Homo_sapiens_hg38* foram utilizados dados obtidos no alinhamento, enquanto para a base de dados *Homo_sapiens_ENSEMBL* foram utilizados dados provenientes da anotação do projeto Ensembl.

Esses dados foram utilizados para povoar a tabela CLUSTERS que continham as respectivas informações organizadas nos atributos: id_clusters (chave primária sequencial da TABELA), chr (número dos cromossomos humanos), seq_acc (identificador do projeto Ensembl para a sequência do transcrito), start_chr (coordenada de início do mapeamento do éxon), end_chr (coordenada de fim do mapeamento do éxon), start_est (coordenada de início do transcrito), end_est (coordenada de fim do transcrito), strand (orientação do mapeamento do éxon no genoma), identity (identidade do mapeamento) e cluster_id (identificador do projeto Ensembl para o gene sem o prefixo “ENSG” e os caracteres “0” que antecedem o primeiro número inteiro”).

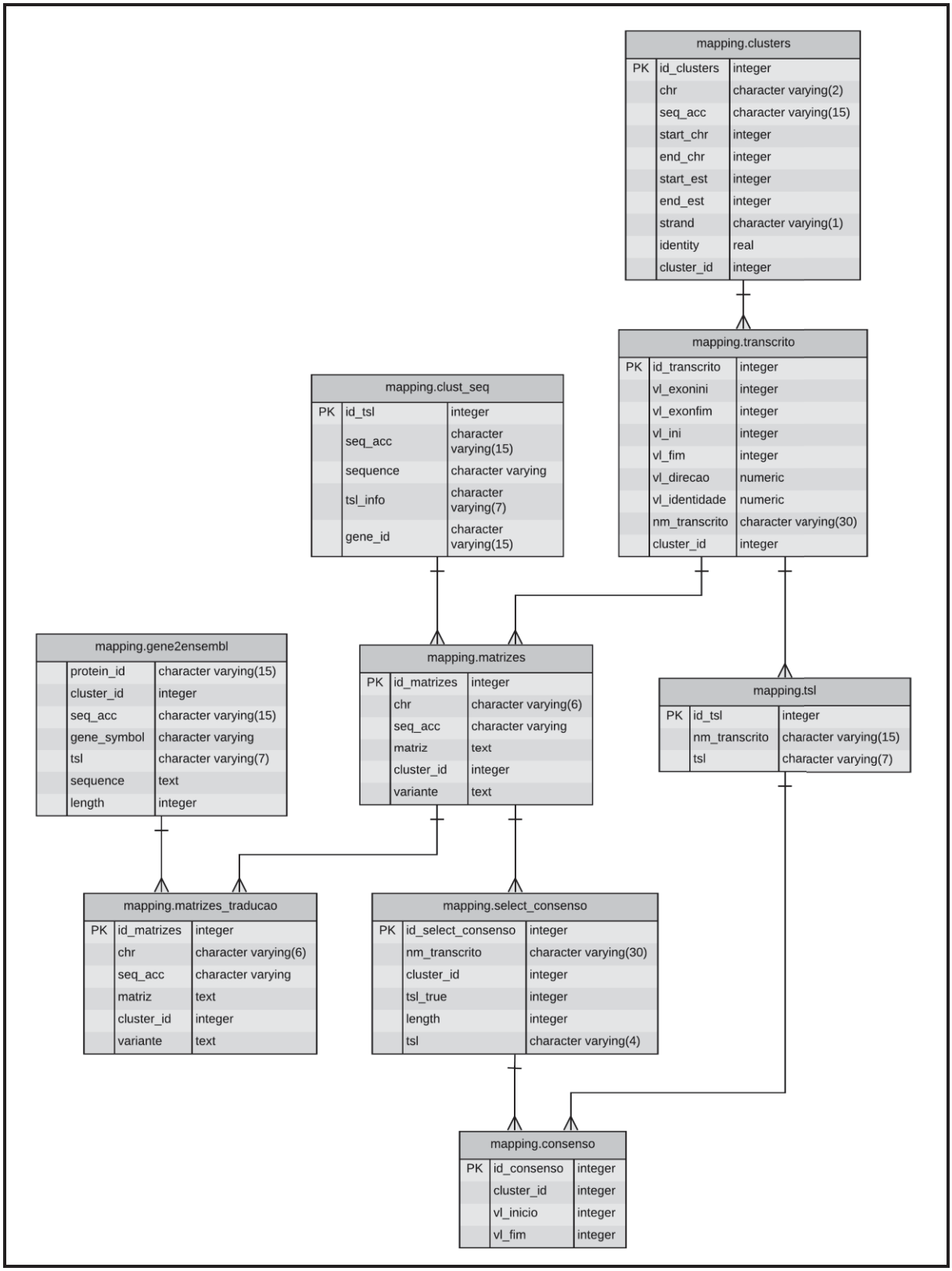
A partir dos dados das tabelas CLUSTERS foram criadas as tabelas TRANSCRITO contendo as informações ordenadas por gene e seus respectivos transcritos, destacando as informações para cada éxon. Para tal, foram utilizados os atributos: id_transcrito (chave primária sequencial da tabela), vl_exonini (coordenada de início do mapeamento do éxon), vl_exonfim (coordenada de fim do mapeamento do éxon), vl_ini (coordenada de início do éxon), vl_fim (coordenada de fim do éxon), vl_direcao (orientação do mapeamento do éxon no genoma), vl_identidade (identidade do mapeamento), nm_transcrito (identificador do projeto Ensembl para a sequência do transcrito), cluster_id (identificador do projeto Ensembl para o gene sem o prefixo “ENSG” e os caracteres “0” que antecedem o primeiro número inteiro”).

Para a nova estrutura do banco de dados, o parâmetro TSL utilizado pelo Ensembl, foi adotado como padrão de confiabilidade para as coordenadas dos transcritos. Para tal, foi desenvolvido um programa em Perl a fim de obter a informação

de TSL para cada transcrito (QUADRO 1) a partir da anotação do genoma (versão 96). A tabela TSL foi, então, preenchida com os atributos: id_tsl (chave primária sequencial da tabela), nm_transcrito (identificador do projeto Ensembl para a sequência do transcrito) e tsl (valor de TSL para o transcrito). Para transcritos que não possuíam valor de TSL anotado e estavam presentes na tabela TRANSCRITO foi considerado o valor “NULL”, sendo considerados equivalentes àqueles que possuíam o valor “NA”. Desta forma, tanto “NULL” quanto “NA” foram considerados os valores mais altos de TSL.

Neste método, o parâmetro TSL foi utilizado para definir qual transcrito poderia ser utilizado como referência de coordenadas externas do primeiro e último éxons (extremidades 5’ e 3’) (FIGURA 7). Quanto menor o valor de TSL maior é a confiabilidade das informações sobre a sequência e junções de *splicing* para o transcrito (QUADRO 1) (http://Jan2020.archive.ensembl.org/info/genome/genebuild/transcript_quality_tags.html). Na ausência de transcritos com TSL1 para o gene, foi escolhido o transcrito com menor valor de TSL (FIGURA 7C). Em casos de genes com mais de um transcrito que possuem o mesmo valor de TSL, foi escolhido aquele com a maior sequência (FIGURA 7B e 7D).

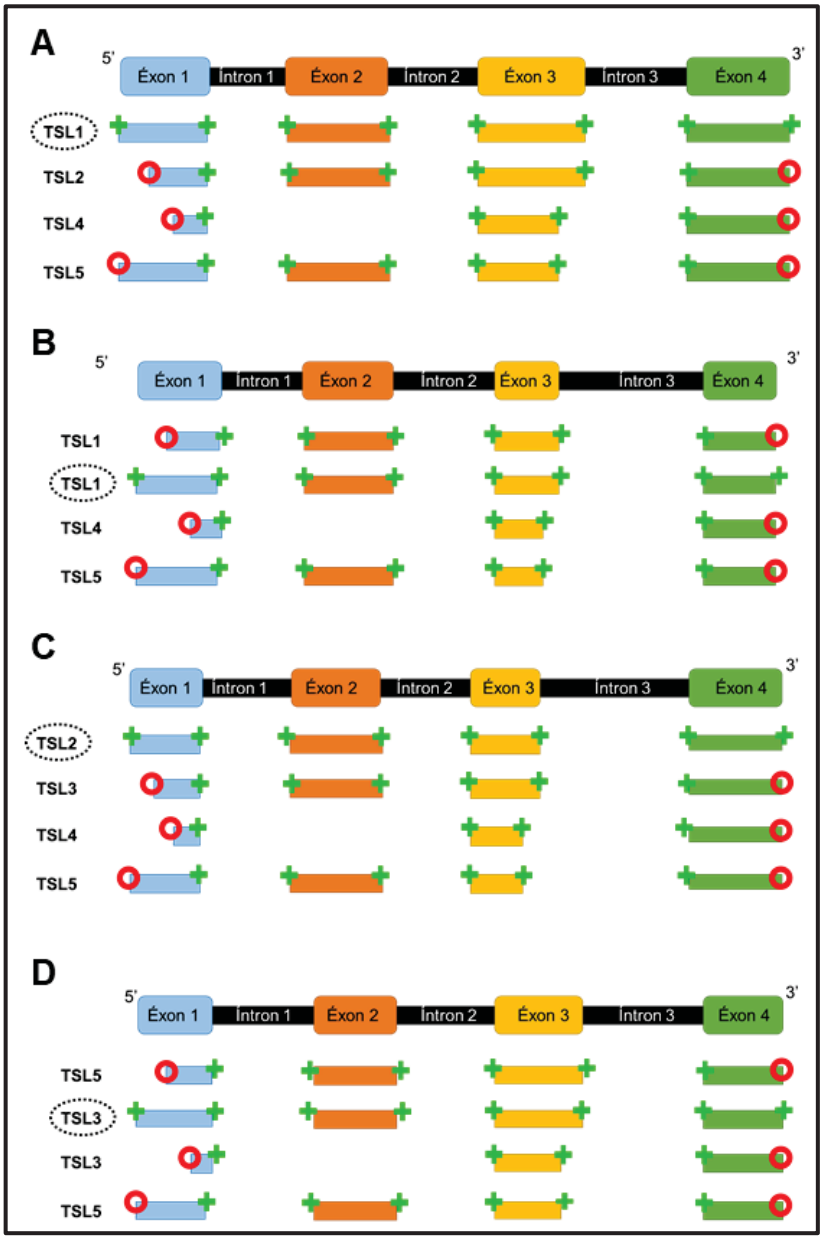
FIGURA 7 – ESTRUTURA RELACIONAL DOS BANCOS DE DADOS



FONTE: O autor (2020).

LEGENDA: A estrutura indicada caracteriza a relação entre as tabelas para os bancos *Homo_sapiens_hg38* e *Homo_sapiens_ENSEMBL*. A descrição de cada campo presente nas tabelas relacionais está apresentada ao longo do texto.

FIGURA 8 – ESQUEMA REPRESENTATIVO DE UM ALINHAMENTO DAS SEQUÊNCIAS DE TRANSCRITOS NO GENE DE REFERÊNCIA



FONTE: O autor (2020).

LEGENDA: São indicados diferentes variantes por *splicing* alternativo de um mesmo gene com diferentes valores de TSL. A cruz verde indica coordenadas consideradas ao criar a TABELA relacional CONSENSO e o círculos vermelhos coordenadas que não serão consideradas. **A** – Gene com um transcrito com TSL 1. **B** – Gene com mais de um transcrito com TSL 1. **C, D e E** – Genes com ausência de transcritos com TSL 1. Em cada conjunto de transcritos o TSL que está circundado com linha pontilhada pertence ao transcrito que será considerado como referência para as coordenadas externas seguindo a lógica adotada neste trabalho, ou seja, está presente na tabela SELECT_CONSENSO. TSL – do inglês, *Transcript Suport Level*

Um programa na linguagem de programação Perl foi desenvolvido para selecionar as coordenadas confiáveis das variantes por *splicing* alternativo de um gene a partir dos transcritos e dos parâmetros de confiabilidade (TSL e tamanho da

sequência de nucleotídeos) disponíveis na tabela CLUSTERS. Esses transcritos foram adicionados na TABELA SELECT_CONSENSO seguindo os seguintes atributos: *id_select_consenso* (chave primária sequencial da tabela), *nm_transcrito* (identificador do projeto Ensembl para a sequência do transcrito), *cluster_id* (identificador do projeto Ensembl para o gene sem o prefixo “ENSG” e os caracteres “0” que antecedem o primeiro número inteiro maior que “0”), *tsl_true* (valor de TSL para o transcrito), length (tamanho da sequência do transcrito) e *tsl* (contendo o valor “true” que foi utilizado para identificar o transcrito como referência nas próximas etapas do método).

Os dados das tabelas TRANSCRITO, CLUSTERS e SELECT_CONSENSO foram utilizados para gerar a tabela CONSENSO que agrupa todas as coordenadas que serão recuperadas para definir o limite dos éxons utilizados no método das matrizes ternárias. Além das coordenadas de todos os éxons dos transcritos da tabela SELECT_CONSENSO, são consideradas confiáveis aquelas que são internas a sequência total dos transcritos (FIGURA 7). A tabela CONSENSO é organizada nos atributos: id_consenso (chave primária sequencial da tabela), cluster_id (identificador do projeto Ensembl para o gene sem o prefixo “ENSG” e os caracteres “0” que antecedem o primeiro número inteiro maior que “0”), vl_inicio (coordenada de início do mapeamento do éxon), vl_fim (coordenada de início do mapeamento do éxon).

3.4.2 Reconstrução das coordenadas dos transcritos e predição das variantes de *splicing* através do método de Matrizes Ternárias

O método de elaboração das matrizes ternárias, desenvolvido por Tavares e colaboradores (2014), foi empregada de forma independente para os bancos de dados *Homo_sapiens_hg38* e *Homo_sapiens_ENSEMBL*. Seguindo a lógica criada anteriormente, a função para construção das matrizes e predição das variantes por *splicing* alternativo utiliza informações das tabelas relacionais TRANSCRITO e CONSENSO. Ao final da execução foram criadas as tabelas MATRIZES contento os atributos: id_matrizes (chave primária sequencial da tabela), chr (identificador do cromossomo onde está localizado o transcrito), seq_acc (identificador do transcrito), matriz (representação do transcrito por uma matriz ternária), cluster_id (identificador do gene) e variante (representação da variante por *splicing* alternativo por uma matriz ternária).

3.4.3 Identificação de variantes suscetíveis a degradação pela via de NMD

O programa NMDClassifier (HSU; LIN; CHEN, 2017) foi utilizado para criar um subconjunto de variantes de *splicing* que não seriam degradadas pela via de NMD e paralelamente, um subconjunto das variantes identificadas que são sensíveis a essa via. Tal programa utiliza a regra da distância do PTC da última junção entre os éxons, adotando uma distância de 50 nucleotídeos à montante da região (HSU; LIN; CHEN, 2017), ou busca no arquivo de anotação a identificação do transcrito como alvo ou não da via de NMD.

Um programa em Perl foi elaborado para utilizar as informações da tabela MATRIZES, CLUSTERS e CONSENSO e construir um arquivo em formato gtf, necessário a utilização do programa NMDClassifier. Presente na tabela MATRIZES, a representação ternária (atributo: *matriz*) de cada gene foi utilizada a fim de reconstruir as coordenadas dos éxons de cada respectivo transcrito. A presença do carácter 1 indicava a presença de determinada região (coordenada a ser recuperada) e 0 a ausência da região, enquanto o carácter "|" indicava o limite do éxon. A informação de coordenadas foi recuperada a partir da tabela CONSENSO, e a informação do cromossomo e fita de orientação foi obtida da tabela CLUSTERS. Por fim, o arquivo GTF foi preenchido com os campos: chr (o cromossomo em que se localiza a sequência do transcrito, origem (informação de onde foram retirados os dados) tipo (indica se a linha contém informações de um transcrito ou de um éxon), início (posição inicial do transcrito ou éxon no cromossomo), final (posição final do transcrito ou éxon no cromossomo), pontuação e fase (receberam o valor ".", pois não se aplicam em nossas análises) fita (orientação do transcrito ou éxon) e informação (informações adicionais separadas por ';' contendo: identificador do gene e identificador do transcrito).

A identificação do transcrito foi formatada a fim de remover o termo "ENST" mantendo apenas os números, como o objetivo de impedir que o NMDClassifier realizasse a busca utilizando a anotação do Ensembl, ou seja, para que ele realizasse a predição baseado na presença de PTC.

Para execução do programa NMDClassifier, além do arquivo gtf respectivo à cada base de dados, foi fornecido um arquivo de anotação do Ensembl (versão 96

para GRCh38/hg38) e um arquivo de sequência dos cromossomos humanos (versão 96 para GRCh38/hg38).

3.4.4 Comparação entre as abordagens quanto a identificação para transcritos sensíveis a via de NMD

Um programa em Perl foi desenvolvido para realizar a comparação, nas duas bases de dados, entre os alvos preditos como sensíveis a via de NMD e a anotação do projeto Ensembl (versão 96). Para esta comparação foram considerados as classificações “alvos da via de NMD”, “codificadores de proteína”, “retenção de íntron”, “transcritos processados” e “outros eventos”, que incluía outro tipo de anotação, como por exemplo, transcritos antisenso, pseudogene processado, lincRNAS (do inglês: *Long intronic noncoding RNA*) entre outros. Também foi considerado o valor de TSL dos transcritos a fim de associar possíveis concordâncias e discordâncias entre as bases de dados ao grau de confiabilidade do transcrito.

Além disso, para as duas bases de dados um programa em Perl foi criado a fim de comparar as coordenadas genômicas das variantes preditas como alvo de NMD. As coordenadas para os esses transcritos foram obtidas dos arquivos GTF criados para as bases *Homo_sapiens_hg38* e *Homo_sapiens_ENSEMBL*. Além disso, para a comparação, também foi utilizado o arquivo GTF de anotação do genoma (versão 96 para GRCh38/hg38) disponível na base de dados do projeto Ensembl (ZERBINO et al., 2018). O programa IGV (versão 2.4.14) (THORVALDSDÓTTIR; ROBINSON; MESIROV, 2013) foi utilizado para uma melhor visualização das diferenças observadas entre as coordenadas genômicas.

3.4.5 Seleção das variantes por *splicing* alternativo não redundantes para cada gene

Para realizar a tradução das variantes preditas pelo método das matrizes ternárias, foi necessário selecionar as variantes distintas para cada gene. Para um mesmo gene, quando mais de um transcrito foi representado pela mesma variante por *splicing* alternativo, foi identificado o transcrito com maior confiabilidade para representar a variante e ser traduzido. Para isso foi desenvolvido um programa em linguagem Perl que recuperava as informações da anotação do transcrito quanto a classificação em “transcript_biotype” e “transcript_support_level” (informação de TSL). Desta forma, em caso de variantes redundantes, foi selecionado àquele que possuía

a classificação protein_coding (tradução: codificador de proteína) para “transcript_biotype”. Em casos em que mais de um transcrito ou nenhum foi considerado protein_coding foi selecionado o que possuía menor TSL. Por fim, caso ainda restasse mais de um transcrito como possível representante de determinada variante, foi selecionado àquele com a maior sequência de nucleotídeos.

Desta forma, a partir das tabelas MATRIZES foram criadas as tabelas relacionais MATRIZES_TRADUCAO. Essas tabelas foram povoadas para os mesmos campos que a tabela de origem e continham um transcrito representativo para cada variante não redundante.

3.4.6 Tradução das variantes por *splicing* alternativo não redundantes

Para realizar a tradução *in silico* das variantes preditas, foram utilizadas informações do arquivo no formato fasta de proteínas disponível no projeto Ensembl (Homo_sapiens_GRCh38 Versão 96 - ftp://ftp.ensembl.org/pub/release-96/fasta/homo_sapiens/pep/Homo_sapiens.GRCh38.pep.all.fa.gz). A TABELA GENE2ENSEMBL foi construída com os campos: protein_id (identificador da proteína), cluster_id (identificador do gene), seq_acc (identificador do transcrito), gene_symbol (nome do gene), tsl (valor de TSL para o transcrito), sequence (sequência de aminoácidos), length (tamanho da sequência da proteína).

A tabela MATRIZES_TRADUCAO foi utilizada em conjunto com a TABELA GENE2ENSEMBL para realizar a tradução *in silico*. Sendo assim, as variantes foram traduzidas utilizando o programa TRANSEQ do pacote EMBOSS (versão 6.3.1) (Rice, 2000). A tradução foi realizada para as 3 frames de leitura possíveis para o transcrito. A seleção da fase de leitura foi realizada pelo programa clustalw2 (versão 2.1) comparando as proteínas traduzidas in silico nas 3 fases de leitura com a proteína de referência retirada dos arquivos .fasta de proteínas do projeto Ensembl (Zerbino, et al 2018). Os parâmetros identidade e cobertura foram considerados para selecionar a sequência traduzida mais confiável. Na ausência de uma proteína de referência, a maior sequência de proteína foi considerada.

4 APRESENTAÇÃO DOS RESULTADOS

Este trabalho foi realizado em conjunto com etapas parciais do doutorado da estudante Letícia Graziela Costa Santos. Nosso resultado foi a construção de uma base de dados com a predição de variantes por *splicing* alternativo baseado em dados do Ensembl. O autor realizou a criação de um subconjunto de dados com a ausência de variantes que possuíssem um PTC, que posteriormente foi traduzido *in silico* dando origem ao repositório SpliceProt. A estrutura relacional das tabelas foi construída com base na linguagem SQL e os arquivos de entrada para preencher as tabelas foram formatados utilizando programas em linguagem Perl desenvolvidos pelo autor.

Para fins de comparação de metodologia, foram executadas duas abordagens, uma com dados proveniente de sequências no formato fasta de cDNA e ncRNA e outra proveniente de dados de anotação do projeto Ensembl (FRANKISH et al., 2019), em formato gtf. Desta forma, foram construídos dois bancos de dados para os diferentes dados de entrada, *Homo_sapiens_hg38* e *Homo_sapiens_ENSEMBL*.

4.1 NÚMERO DE TRANSCRITOS PARA OS DOIS BANCOS DE DADOS

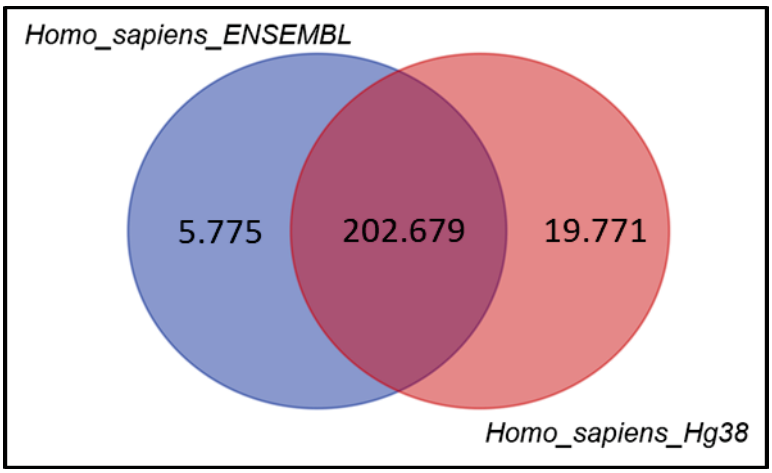
Para a base de dados *Homo_sapiens_hg38*, foram agrupadas as sequências de cDNA e de ncRNA disponíveis para a versão 96 do genoma humano GRCh38/hg38 do Ensembl. Ao todo, um total de 228.297 sequências de transcritos foram alinhadas ao genoma humano utilizando o programa BLAT (KENT, 2002). Ao total, foram produzidos 5.800.541 alinhamentos em formato psl e 63.264.912 alinhamentos no formato blast9. Após a seleção dos melhores alinhamentos não repetidos, restaram 222.462 alinhamentos. O programa format_clusterBlat.pl foi utilizado para formatar as informações dos alinhamentos do arquivo psl no arquivo input_cluters_Blat.csv, entretanto a informação de identidade do alinhamento foi obtida do arquivo com formato blast9.

Para a base de dados *Homo_sapiens_ENSEMBL*, o arquivo de anotação do Ensembl foi usado para fornecer as informações necessárias através do programa format_clusterENSMBL.pl, no total foram obtidas informações para 208.460 transcritos.

A FIGURA 8 mostra um diagrama de Venn comparando os transcritos presentes nas duas bases de dados. Enquanto 202.685 transcritos estavam presentes nas duas

bases de dados, 5.775 transcritos foram exclusivos do arquivo proveniente da anotação do Ensembl e 19.777 transcritos só estavam presentes no arquivo gerado a partir do alinhamento das sequências em formato fasta.

FIGURA 9 – DIAGRAMA DE VENN REPRESENTANDO A QUANTIDADE DE TRANSCRITOS PRESENTES EM CADA BANCO DE DADOS



FONTE: O autor (2020).

LEGENDA: À direita, em vermelho, a quantidade de transcritos presentes somente da base de dados *Homo_sapiens_hg38*. À esquerda, em azul, a quantidade de transcritos presentes somente na base de dados *Homo_sapiens_ENSEMBL*.

4.2 IDENTIFICAÇÃO DE NÍVEL DE CONFIABILIDADE E SEQUÊNCIA DOS TRANSCRITOS.

O padrão TSL foi utilizado para selecionar as coordenadas confiáveis dos transcritos e assim estabelecer as junções de *splicing* confiáveis. Entretanto, considerando o projeto Ensembl, a informação de TSL está presente apenas no arquivo de anotação do genoma, ainda assim para alguns transcritos essa informação estava ausente.

A sequência de cada transcrito foi obtida do arquivo fasta criado com os arquivos de sequência de cDNA e ncRNA obtidos do repositório Ensembl.

Ao todo, foram recuperadas informações de sequência e TSL para 207.368 transcritos. Dos quais, 41.836 possuíam TSL1, 39.419 transcritos com TLS 2, 33.557 transcritos com TSL 3, 17.173 transcritos com TSL 4, 37.614 transcritos com TSL 5, 28.026 transcritos com TSL “NA” e 9.743 que receberam o valor “NULL” (TABELA 1).

TABELA 1 – NÚMERO DE TRANSCRITOS QUE POSSUÍAM SEQUÊNCIA, SEPARADOS POR TSL

Transcript Support Level	**Número de transcritos**
TSL1	41.836
TSL2	39.419
TLS3	33.557
TSL4	17.173
TSL5	37.614
TSLNA	28.026
TSLNULL	9.743
Total	*207.368*

Fonte: O autor (2020)

Neste trabalho, transcritos que não possuíam informação de sequência foram excluídos do banco de dados pois a sequência do transcrito é necessária a execução dos métodos de construção do repositório SpliceProt. Esses transcritos foram armazenados em um arquivo separado e serão tratados em abordagens que ainda serão desenvolvidas.

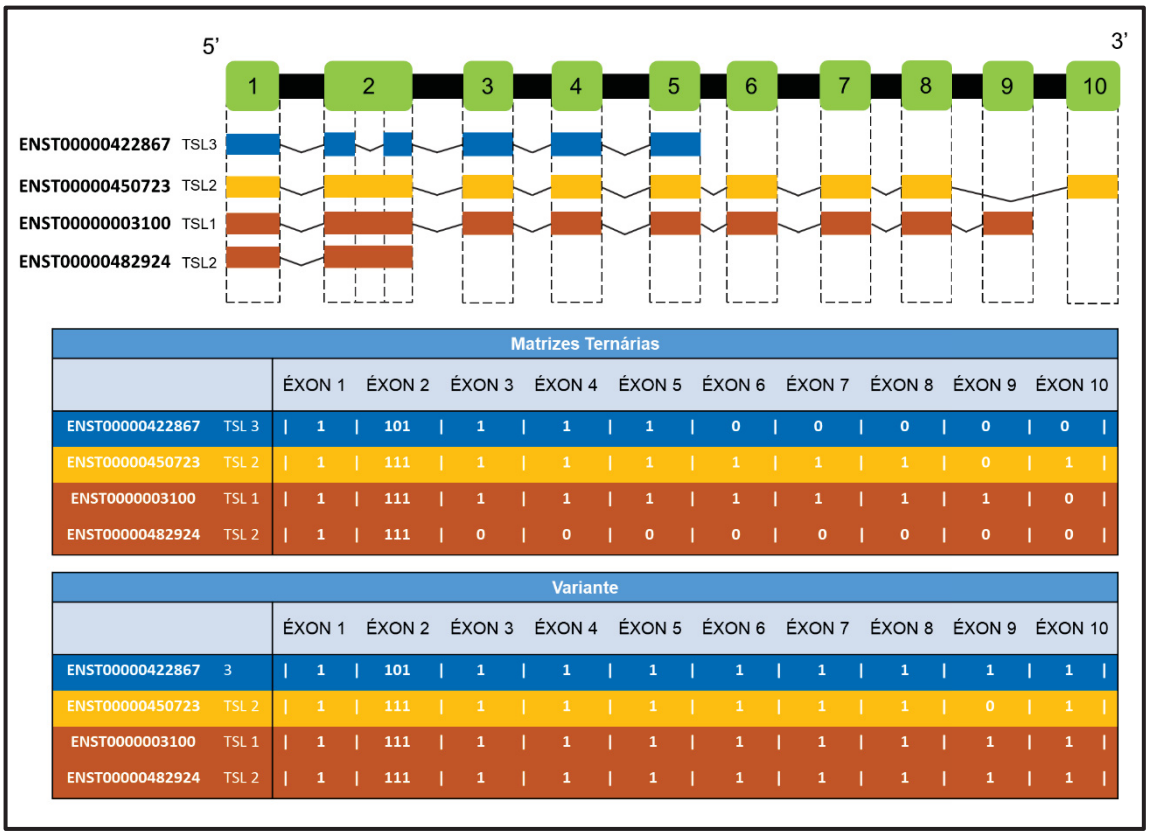
4.3 MATRIZES TERNÁRIAS

O método desenvolvido por Tavares e colaboradores (2014), resumido nos itens 3.1 e 3.3.2, foi aplicado para ambas as bases de dados. Sendo assim, foi construída a matriz representativa de cada transcrito, bem como a matriz variante representando os eventos de *splicing* alternativo associados a cada transcrito (FIGURA 9).

Foram construídas 222.462 matrizes para um total de 61.122 genes na base de dados *Homo_sapiens_hg38*. Para a base de dados *Homo_sapiens_ENSEMBL* foram construídas 208.460 matrizes para um total de 58.825 genes. Ao final, foram observados genes que apresentavam matrizes com discrepância, que receberam somente o caractere “0”, o que indicava ausência de informação na região e “|” que indicava os limites dos éxons. Sendo assim, esses transcritos foram removidos das

bases de dados e ao final, a base de dados *Homo_sapiens_hg38* reuniu 222.450 transcritos com suas respectivas matrizes e a base de dados *Homo_sapiens_ENSEMBL* reuniu 208.454 transcritos (TABELA 2).

FIGURA 10 – ESQUEMA REPRESENTATIVO DA CONSTRUÇÃO DE UMA MATRIZ TERNÁRIA PARA O GENE ENSG00000001630 (CYP51A1) PARA BASE DE DADOS *HOMO_SAPIENS_HG38*.



Matrizes Ternárias

	ÉXON 1	ÉXON 2	ÉXON 3	ÉXON 4	ÉXON 5	ÉXON 6	ÉXON 7	ÉXON 8	ÉXON 9	ÉXON 10
ENST00000422867 TSL 3	1	101	1	1	1	0	0	0	0	0
ENST00000450723 TSL 2	1	111	1	1	1	1	1	1	0	1
ENST0000003100 TSL 1	1	111	1	1	1	1	1	1	1	0
ENST00000482924 TSL 2	1	111	0	0	0	0	0	0	0	0

Variante

	ÉXON 1	ÉXON 2	ÉXON 3	ÉXON 4	ÉXON 5	ÉXON 6	ÉXON 7	ÉXON 8	ÉXON 9	ÉXON 10
ENST00000422867 3	1	101	1	1	1	1	1	1	1	1
ENST00000450723 TSL 2	1	111	1	1	1	1	1	1	0	1
ENST0000003100 TSL 1	1	111	1	1	1	1	1	1	1	1
ENST00000482924 TSL 2	1	111	1	1	1	1	1	1	1	1

FONTE: O autor (2020).

LEGENDA: Os transcritos são alinhados e utilizando como parâmetros o TSL e o tamanho da sequência, são determinadas as coordenadas confiáveis. Ao todo, o gene indicado possui 4 transcritos e 3 variantes. Ou seja, ENST00000003100 e ENST00000482924 são a mesma variante geradas por *splicing* alternativo. O caractere "|" representa os limites dos éxons, o caractere "0", representa a ausência de determinada região e "1" a presença da região do éxon. A TABELA variante representa os eventos de *splicing* alternativo preditos para o gene baseado na metodologia descrita neste trabalho. O transcrito ENST00000522867 apresenta uso alternativo do éxon 2. O transcrito ENST00000540723 apresenta uso alternativo do éxon 9.

TABELA 2 – NÚMERO DE TRANSCRITOS E GENES NAS BASES DE DADOS CRIADAS

Base de dados	***Homo_sapiens_hg38***		***Homo_sapiens_ENSEMBL***	
	Transcritos	***Genes***	***Transcritos***	***Genes***
Número total de transcritos	222.462	61.122	208.460	58.825
Transcritos com matrizes discrepantes	12	11	6	6
Número final de transcritos	222.450	61.122	208.454	58.825

Fonte: O autor (2020)

4.4 IDENTIFICAÇÃO DAS VARIANTES ALVO DA VIA DE NMD.

Para detecção de transcritos alvos da via de NMD foi utilizado o software NMDClassifier (HSU; LIN; CHEN, 2017), que realiza a predição baseado na presença de um PTC na fase de leitura da proteína.

Na base de dados *Homo_sapiens_hg38*, foram preditas 34.497 variantes como alvo da via de NMD. Foram identificados 9.602 genes que possuíam pelo menos uma variante predita como alvo da via de NMD (Tabela 3). O gene ENSG00000281794 (MUC20-OT1) foi identificado como gene com o maior número de variantes preditas como alvo de NMD, totalizando 98 variantes.

Para o banco *Homo_sapiens_ENSEMBL* foram preditas 31.642 variantes alvo da via de NMD. Foram identificados 8.833 genes com pelos menos uma variante alvo de NMD (Tabela 3). O gene ENSG00000127990 (SGCE) foi o gene identificado com o maior número de variantes preditas como alvo de NMD, totalizando 79 variantes.

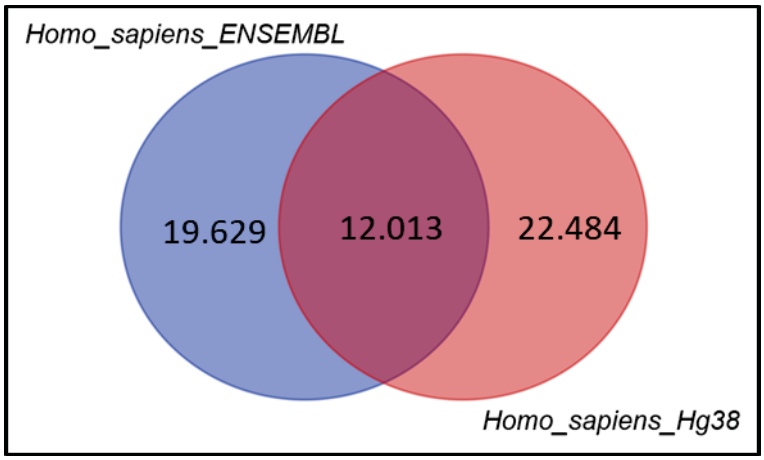
TABELA 3 – NÚMERO DE TRANSCRITOS COM RELAÇÃO A VIA DE NMD NAS BASES DE DADOS CRIADAS

Base de dados	*Homo_sapiens_hg38*	*Homo_sapiens_ENSEMBL*
Número de transcritos no banco de dados	222.450	208.454
Número variantes preditas como alvo da via de NMD	34.497	31.642
Número de genes com pelo menos uma variante predita como alvo de NMD	9.602	8.833
Número máximo de transcritos sensíveis a NMD para um mesmo gene	98	79

Fonte: O autor (2020)

Dentre o total de variantes preditas como alvo de NMD considerando as duas bases de dados criadas, 19.629 foram identificadas apenas na base de dados *Homo_sapiens_ENSEMBL* e 22.484 apenas na base de dados *Homo_sapiens_hg38*, enquanto 12.013 nas duas bases de dados (FIGURA 10).

FIGURA 11 – DIAGRAMA DE VENN REPRESENTANDO A QUANTIDADE DE VARIANTES IDENTIFICADAS COMO SENSÍVEIS A NMD EM CADA BANCO DE DADOS

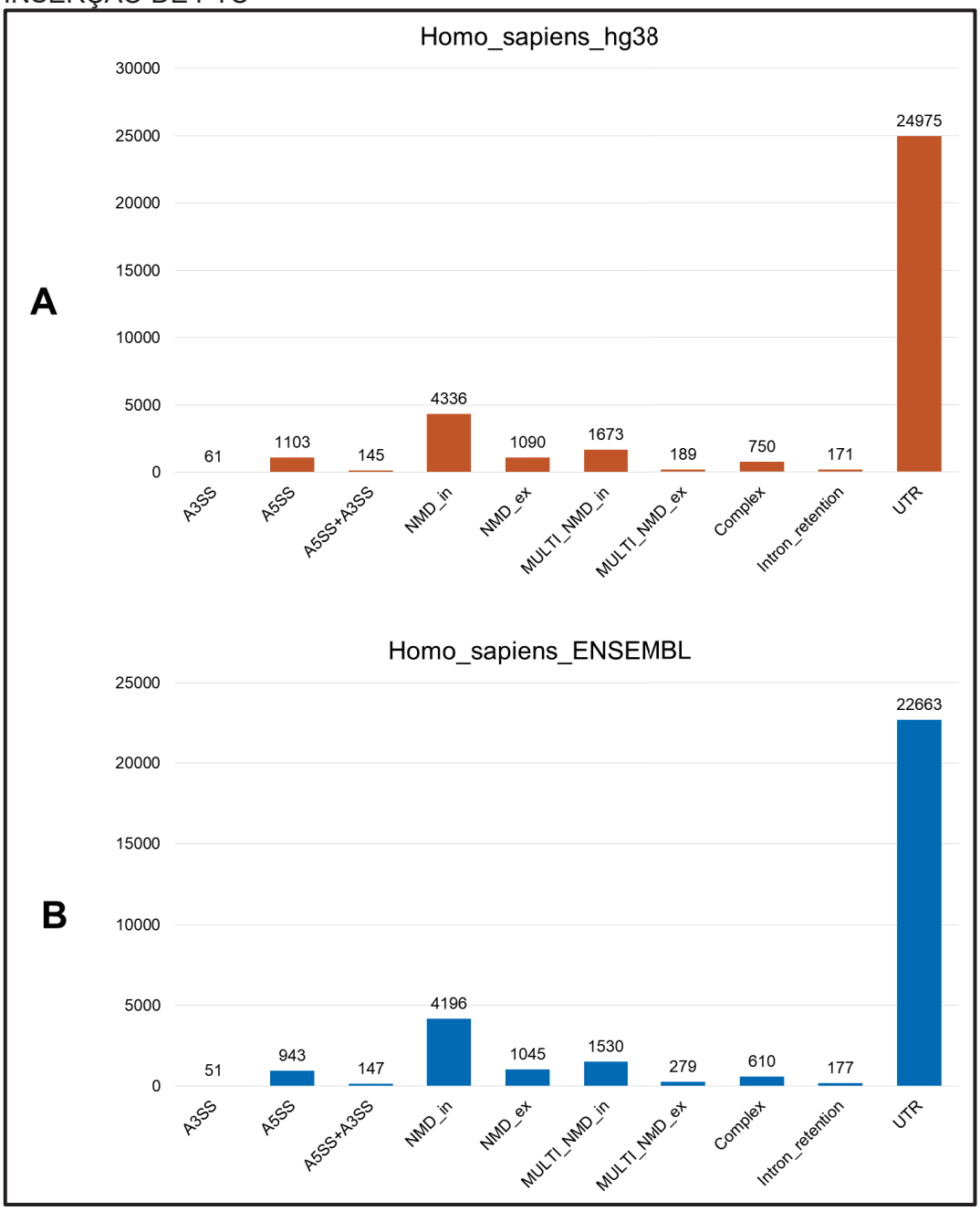


FONTE: O autor (2020).

LEGENDA: À direita em vermelho a quantidade de variantes preditas como alvo da via de NMD para a base de dados *Homo_sapiens_hg38*. À esquerda em azul, a quantidade de variantes preditas como alvo da via de NMD na base de dados *Homo_sapiens_ENSEMBL*.

Além de identificação das variantes preditas como alvo da via de NMD, o programa NMDClassifier faz a predição do evento de *splicing* que poderia ter resultado na inserção de um PTC na determinada variante. Na base de dados *Homo_sapiens_hg38* os maiores números de eventos associados a inserção de um PTC nos transcritos foram alterações nas regiões UTRs (24.975) e eventos de inclusão de 1 éxon (4.336) (GRÁFICO 1A). O mesmo padrão foi observado para base de dados *Homo_sapiens_ENSEMBL* com 22.663 eventos de alterações nas regiões UTRs e 4.196 eventos de inserção de 1 éxon (GRÁFICO 1B).

GRÁFICO 1 – NÚMERO DE EVENTOS DE *SPLICING* ALTERNATIVO ASSOCIADOS A INSERÇÃO DE PTC



FONTE: O autor (2020).

LEGENDA: **A -** Resultados observados para base de dados *Homo_sapiens_hg38.* **B -** Resultados observados para base de dados *Homo_sapiens_ENSEMBL.* A3SS – alterações no 3'SS; A5SS – alterações no 5'SS; A3SS+A5SS – alterações em ambos os SS, 5' e 3'; NMD_in – inclusão de um éxon; NMD_ex – exclusão de um éxon; multi_NMD_in – inclusão de múltiplos éxons; multi_NMD_ex – exclusão de múltiplos éxons; Complex – múltiplos eventos de alteração estrutural; Intron_retention – evento de retenção de íntron; UTR – alterações as regiões UTRs (HSU; LIN; CHEN, 2017).

4.4.1 Diferenças de anotação para variantes preditas como alvo da via de NMD em comparação com o projeto Ensembl

O programa NMDClassifier identificou variantes como alvo da via de NMD que não estavam de acordo com a anotação disponível no projeto Ensembl, tanto para a base de dados *Homo_sapiens_hg38* quanto para a base de dados *Homo_sapiens_ENSEMBL*.

4.4.1.1 Base de dados *Homo_sapiens_hg38*

Foram encontradas diferença quanto à identificação de transcritos alvo de NMD preditos pelo programa NMDClassifier na base de dados *Homo_sapiens_hg38* com relação a anotação do transcrito no projeto Ensembl. Das variantes identificadas como alvo de NMD, 4.634 variantes estavam em concordância com a anotação do projeto Ensembl, 15.697 estavam anotadas como codificadoras de proteínas, 3.940 foram anotadas como retenção de íntron, 4.491 foram anotadas como transcritos processados e 5.735 possuíam outro tipo de anotação (Tabela 4).

TABELA 4 – COMPARAÇÃO ENTRE O NÚMERO DE TRANSCRITOS NA BASE DE DADOS *HOMO_SAPIENS_HG38* E DADOS ANOTADOS DO PROJETO ENSEMBL EM RELAÇÃO A IDENTIFICAÇÃO COMO ALVO DA VIA DE NMD

Homo_sapiens_hg38			
NMDClassifier	***Anotação Ensembl***	**Número de variantes**	**TSL1**
Alvo de NMD	Alvo de NMD	4.634	1.835
Alvo de NMD	Codificador de proteína	15.698	5.804
Alvo de NMD	Retenção de íntron	4.491	550
Alvo de NMD	Transcrito processado	3.887	327
Alvo de NMD	“Outros eventos”	2.369	991
Não sensível a NMD	Alvo de NMD	10.721	4.304
Alvo de NMD	Sem anotação	3.418	-

Fonte: O autor (2020)

4.4.1.2 Base de dados *Homo_sapiens_ENSEMBL*

O banco de dados *Homo_sapiens_ENSEMBL* foi criado a partir dos dados de anotação disponíveis no projeto Ensembl, ainda assim, foram observadas diferenças na predição realizada pelo programa NMDClassifier em relação a notação do projeto Ensembl. 4.744 variantes compartilhavam a mesma classificação para NMD com a

anotação do projeto Ensembl, 15.694 estavam anotadas como codificadoras de proteínas, 4.578 como retenção de íntron, 4.024 como transcritos processados e 2.602 possuíam outro tipo de anotação (TABELA 5).

TABELA 5 – COMPARAÇÃO ENTRE O NÚMERO DE TRANSCRITOS NA BASE DE DADOS *HOMO_SAPIENS_ENSEMBL* E DADOS ANOTADOS DO PROJETO ENSEMBL EM RELAÇÃO A IDENTIFICAÇÃO COMO ALVO DA VIA DE NMD.

Homo_sapiens_ENSEMBL			
NMDClassifier	***Anotação Ensembl***	**Número de variantes**	**TSL1**
Alvo de NMD	Alvo de NMD	4744	819
Alvo de NMD	Codificador de proteína	15.694	5.932
Alvo de NMD	Retenção de íntron	4.578	541
Alvo de NMD	Transcrito processado	4.024	1.546
Alvo de NMD	“Outros eventos”	2.602	1.105
Não sensível a NMD	Alvo de NMD	10.798	1.961

Fonte: O autor (2020)

4.4.2 Comparação dos alvos da via de nmd na base de dados *Homo_sapiens_hg38* e *Homo_sapiens_ENSEMBL*

A predição realizada pelo programa NMDClassifier é baseada na presença de um PTC na sequência do transcrito. Visto isso, para os transcritos que foram preditos com alvo de NMD, foram avaliadas as coordenadas genômicas referentes à sequência de nucleotídeos. Nas TABELAS 6 e 7 estão indicadas as quantidades de transcritos preditos como alvo da via de NMD em cada base de dados e que possuíam coordenadas diferentes em pelo menos um éxon do transcrito quando comparadas as bases de dados com a anotação do projeto Ensembl.

TABELA 6 – NÚMERO DE VARIANTES NA BASE DE DADOS *HOMO_SAPIENS_HG38* E COM COORDENADAS GENÔMICAS DIFERENTES DA ANOTAÇÃO DO PROJETO ENSEMBL E A CLASSIFICAÇÃO COMO ALVO DA VIA DE NMD.

HOMO_SAPIENS_HG38		
Anotação Ensembl	**NMDClassifier**	**Número de variantes com pelo menos 1 éxon com coordenadas diferentes**
Alvo de NMD	Alvo de NMD	4.563
Codificador de proteína	Alvo de NMD	15.609
Retenção de íntron	Alvo de NMD	4.466
Transcrito processado	Alvo de NMD	3.838
“Outros eventos”	Alvo de NMD	2.359
Alvo de NMD	Não sensível a NMD	10.599

Fonte: O autor (2020)

Para a base de dados *Homo_sapiens_hg38*, 19.771 transcritos não estavam presentes no arquivo de anotação do projeto Ensembl (Zerbino, 2018). Sendo assim, eles não foram considerados para a comparação de coordenadas genômicas.

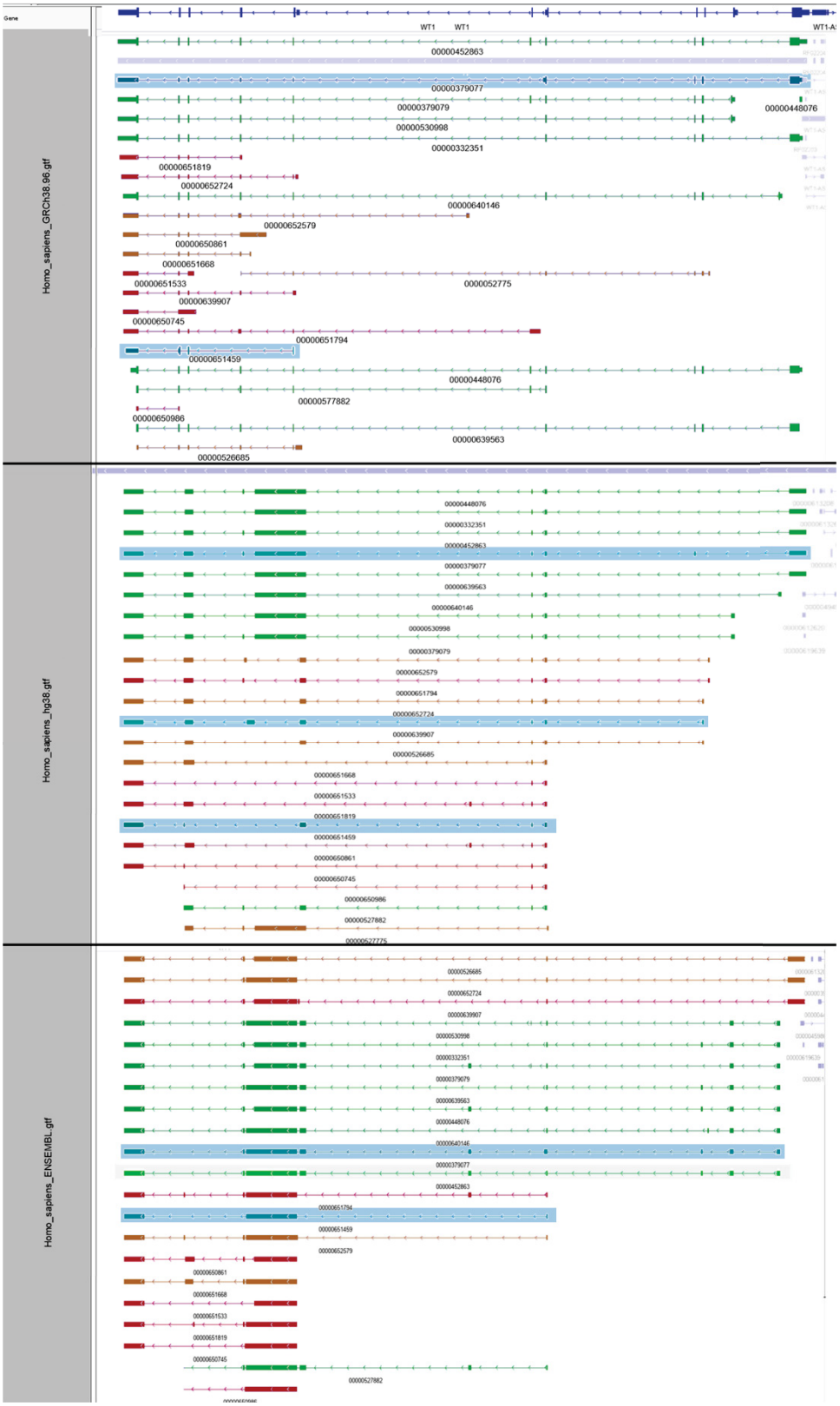
TABELA 7 – NÚMERO DE VARIANTES NA BASE DE DADOS *HOMO_SAPIENS_HG38* E COM COORDENADAS GENÔMICAS DIFERENTES DA ANOTAÇÃO DO PROJETO ENSEMBL E A CLASSIFICAÇÃO COMO ALVO DA VIA DE NMD.

HOMO_SAPIENS_ENSEMBL		
Anotação Ensembl	**NMDClassifier**	**Número de variantes com pelo menos 1 éxon com coordenadas diferentes**
Alvo de NMD	Alvo de NMD	4.670
Codificador de proteína	Alvo de NMD	15.596
Retenção de íntron	Alvo de NMD	4.553
Transcrito processado	Alvo de NMD	3.972
“Outros eventos”	Alvo de NMD	2.596
Alvo de NMD	Não sensível a NMD	10.678

Fonte: O autor (2020)

Para ilustrar as diferenças observadas, foi utilizado o programa IGV para visualizar o alinhamento das variantes preditas como alvo de NMD, e comparar com as coordenadas dos transcritos nas duas bases de dados com as coordenadas anotadas pelo projeto Ensembl. Na figura 11, está indicado a visualização do gene WT1 (do inglês, *Wilms' Tumor-Associated gene 1*), com nenhuma variante identificada com alvo da via de NMD pelo NMDClassifier, diferente do indicado pelo projeto Ensembl. Na FIGURA 12, está representado o gene ROBO3 (do inglês, *Roundabout Guidance Receptor 3*) que possui variantes identificadas como alvo da via de NMD nas duas bases de dados, mas não na anotação do projeto Ensembl.

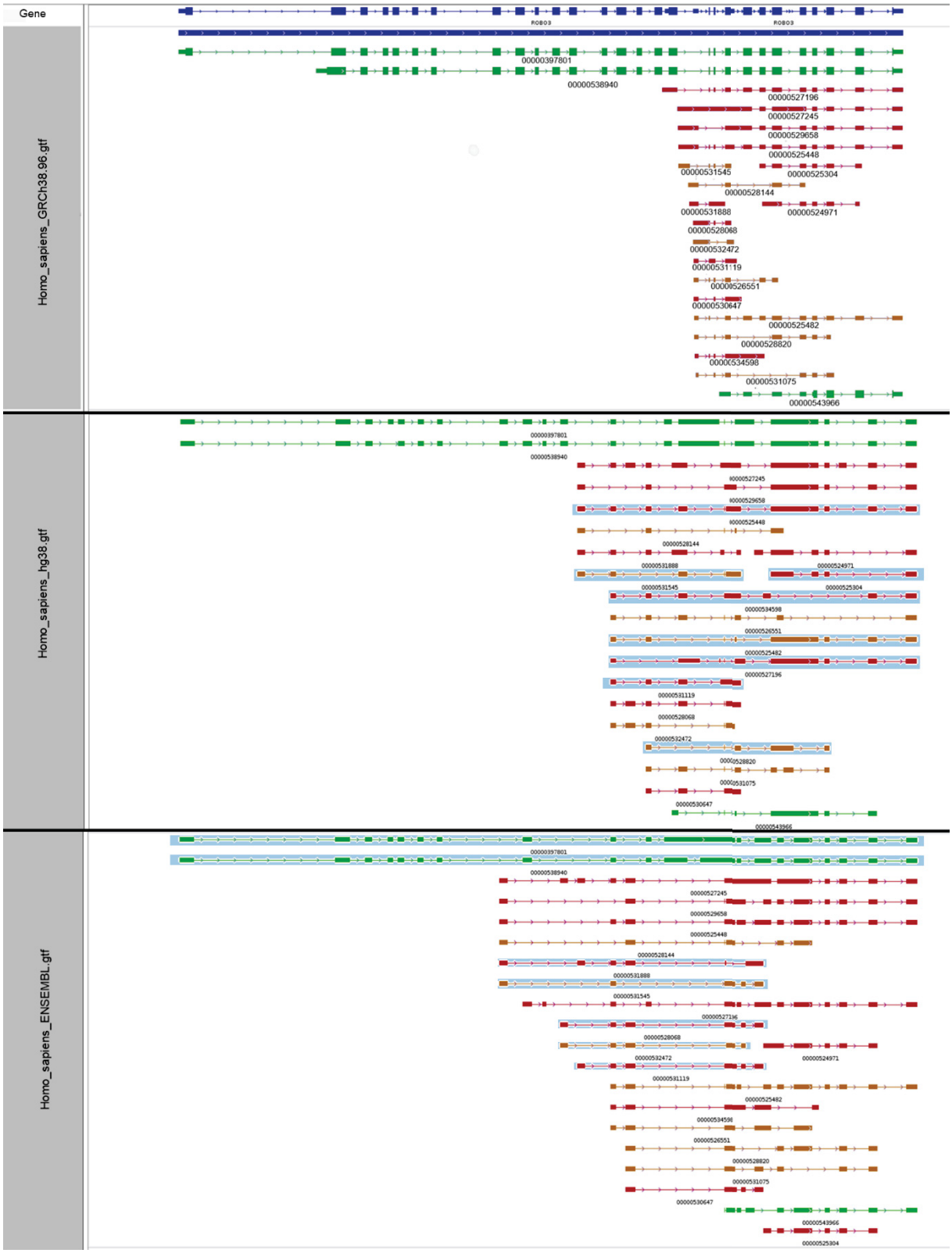
FIGURA 12 - VISUALIZAÇÃO DO ALINHAMENTO PARA O GENE WT1 NAS BASES DE DADOS E SEGUNDO A ANOTAÇÃO DO PROJETO ENSEMBL.



FONTE: O autor (2020).

LEGENDA: Em verde, transcritos anotados como codificadores de proteína, em marrom transcritos anotados como transcritos processados, em vermelho transcritos anotados como retenção de íntron, em azul claro transcritos anotados como alvo da via de NMD pelo projeto Ensembl.

FIGURA 13 – VISUALIZAÇÃO DO ALINHAMENTO PARA O GENE ROBO3 NAS BASES DE DADOS E SEGUNDO A ANOTAÇÃO DO PROJETO ENSEMBL.



FONTE: O autor (2020).

LEGENDA: Em verde, transcritos anotados como codificadores de proteína, em marrom transcritos anotados como transcritos processados, em vermelho transcritos anotados como retenção de íntron. Os transcritos destacados por um sombreado azul foram preditos como alvo da via de NMD pelo programa NMDClassifier na determinada base de dados.

4.5 TRADUÇÃO DAS VARIANTES

4.5.1 Remoção das variantes redundantes.

Na tabela relacional das matrizes ternárias para ambas as bases de dados foram identificados padrões de variantes idênticos para diferentes transcritos (FIGURA 9). Sendo assim, nesses casos, para cada gene foi selecionado o transcrito com coordenadas mais confiáveis, seguindo o critério descrito no item 3.3.5. Para a base de dados *Homo_sapiens_hg38* foram removidas 49.258 variantes redundantes, ou seja, aproximadamente 21,95% dos transcritos presentes na base de dados. Portanto, foram selecionadas 173.192 variantes por *splicing* alternativo para compor a TABELA MATRIZES_TRADUCAO_hg38 (Tabela 8).

Para a base de dados *Homo_sapiens_ENSEMBL*, foram removidas 43.248 variantes redundantes, ou seja, cerca de 25,8% dos transcritos. Desta forma, a TABELA relacional MATRIZES_TRADUCAO_ENSEMBL foi povoada com informações de 165.206 variantes (Tabela 8).

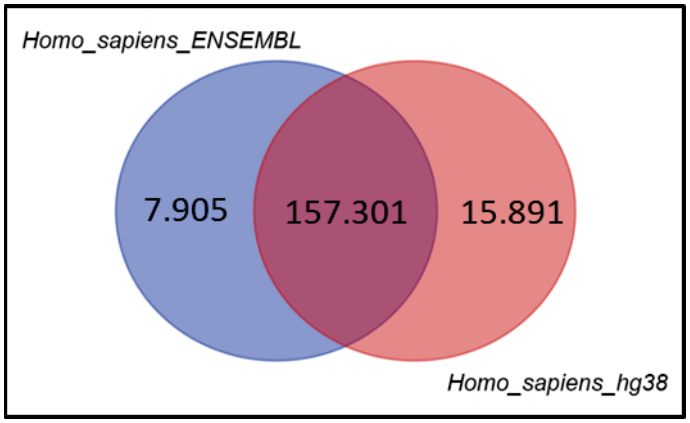
TABELA 8 – NÚMERO DE TRANSCRITOS PARA BASES DE DADOS CRIADAS

Base de dados	***Homo_sapiens_hg38***	***Homo_sapiens_ENSEMBL***
Número variantes preditas	222.450	208.454
Número variantes redundantes	49.258	43.248
Número de transcritos para tradução	173.192	165.206

Fonte: O autor (2020)

Na FIGURA 13 está representado um diagrama de Venn com quantidade de variantes presentes em cada base de dados após a remoção das redundâncias.

FIGURA 14 – DIAGRAMA DE VEEN REPRESENTANDO A QUANTIDADE DE VARIANTES À SEREM TRADUZIDAS PRESENTES EM CADA BANCO DE DADOS.



FONTE: O autor (2020).

LEGENDA: À direita em vermelho a quantidade de transcritos com variantes não redundantes presentes somente da base de dados *Homo_sapiens_hg38*. À esquerda em azul, a quantidade de transcritos com variantes não redundantes na base de dados *Homo_sapiens_ENSEMBL*.

4.5.2 Execução do método de tradução

A respectiva TABELA relacional MATRIZES_TRADUCAO para cada base de dados foi utilizada para aplicar o método de Tavares e colaboradores (2014) para realizar a tradução das variantes não redundantes. Para a base de dados *Homo_sapiens_hg38*, foram traduzidas 113.665 proteínas não redundantes enquanto, para a base de dados *Homo_sapiens_ENSEMBL* foram traduzidas 107.464 proteínas não redundantes.

Na TABELA 9 está indicado a quantidade de transcritos anotados pelo projeto Ensembl como retenção de íntron, transcritos processados e outros eventos, além disso está indicado a quantidade de transcritos traduzidos que foram classificados como alvos da via de NMD pelo programa NMDClassifier.

TABELA 9 – NÚMERO DE TRANSCRITOS TRADUZIDOS PARA BASES DE DADOS CRIADAS.

Base de dados	***Homo_sapiens_hg38***	***Homo_sapiens_ENSEMBL***
Número traduções	113.665	107.464
Anotação Ensembl retenção de íntron	6.965	7.081
Anotação Ensembl transcrito processado	10.513	10.812
Anotação Ensembl outros eventos	18.367	19.270
NMDClassifier – alvo de NMD	17.688	17.929

Fonte: O autor (2020)

Por meio da abordagem descrita neste trabalho, foi possível identificar 35.845 variantes para a base de dados *Homo_sapiens_hg38* e 37.163 variantes para a base de dados *Homo_sapiens_ENSEMBL* que foram traduzidas hipoteticamente e não possuíam a anotação como “codificadora de proteína” no projeto Ensembl (FRANKISH et al., 2019).

5 DISCUSSÃO

5.1 REPOSITÓRIO SPLICEPROT (2014)

Como consequência do splicing alternativo, é observado um aumento na diversidade do transcriptoma (TRESS; ABASCAL; VALENCIA, 2017). Entretanto, nem todos os bancos de dados de sequências biológicas possuem informações verossímeis para grande parte dessas variantes (WANG et al., 2012). Devido a isso, abordagens de bioinformática são uma forma de aumentar a resolução da identificação dessas variantes, o que pode contribuir para a identificação de peptídeos em experimentos de Proteogenômica (NESVIZHSKII, 2014). Desta forma, se faz necessário também a identificação do potencial codificador de tais variantes. Visto isso, Tavares e colaboradores (2014) desenvolveram uma abordagem para desenvolver um banco de dados com a tradução de peptídeos a partir da predição de variantes por *splicing* alternativo (TAVARES et al., 2014).

Sendo assim, Tavares e colaboradores (2014) desenvolveram o repositório SpliceProt utilizando dados do repositório Unigene (PONTIUS; WAGNER; SCHULER, 2003) e reconstruindo os transcritos através do método de matrizes ternárias (TAVARES et al., 2014). Com a abordagem utilizada, foi possível identificar novas variantes de *splicing* com potencial codificador. Em suas análises, os autores verificaram que o método utilizado para construção do SpliceProt era capaz de detectar um maior número de variantes de *splicing* quando comparado às outras bases de dados como RefSeq (O’LEARY et al., 2015) e o projeto Ensembl (FRANKISH et al., 2019). Além disso, após realizada a tradução in silico, identificou 54 peptídeos associados à essas variantes que não estavam presentes na base de dados UniprotKB/TrEMBL (THE UNIPROT CONSORTIUM, 2019).

Após o repositório Unigene (PONTIUS; WAGNER; SCHULER, 2003) ser descontinuado em 2019, foi necessário atualizar o método de construção do repositório SpliceProt. Além disso, foi observada a presença de variantes por *splicing* com PTC na base de dados criada por Tavares e colaboradores, que consequentemente seriam degradadas pela via de NMD e não traduzidas. Sendo assim, este estudante em conjunto com a doutoranda Letícia Graziela Costa Santos atualizaram a metodologia de construção do banco de dados e inseriu no fluxo de execução do método uma etapa para identificar variantes com a presença de PTC.

5.2 CONSTRUÇÃO DA NOVA VERSÃO DO REPOSITÓRIO SPLICEPROT - 2020

Os dados necessários para a criação do SpliceProt foram obtidos do projeto Ensembl (Homo_sapiens_GRCh38 Versão 96 ftp://ftp.ensembl.org/pub/release-96). Além de possuir todas as informações necessárias à execução da metodologia do SpliceProt, o projeto Ensembl possui diversas informações para variantes por *splicing* alternativo e é constantemente atualizado e amplamente utilizado em diversos trabalhos (HUNT et al., 2018; ZERBINO et al., 2018).

Ao verificar os dados disponibilizados na base de dados do projeto Ensembl (Zerbino,2018) foi observado que diferentes arquivos possuem informações não correspondentes. Foi observado uma variedade de transcritos (cerca de 9,5%) no arquivo de sequências do arquivo disponibilizado em formato fasta que foi utilizado para a criação da base de dados *Homo_sapiens_hg38*. Mesmo depois do alinhamento com o programa BLAT (KENT, 2002) e a filtragem das redundâncias, foi observado cerca de 6,5% de diferença na quantidade de transcritos entre as duas bases de dados criadas. Sendo assim, pudemos concluir que há transcritos que possuem sequência disponibilizada no projeto Ensembl (FRANKISH et al., 2019) ainda não possuem anotação quando comparado ao arquivo no formato gtf disponibilizado pelo referido projeto.

Por outro lado, foi observado que uma parte dos transcritos estavam presentes exclusivamente no arquivo de anotação do genoma (*Homo_sapiens_ENSEMBL*), ou seja, não possuíam informação de sequência de nucleotídeos disponível. Ainda que tenha sido realizada a etapa de filtragem (item 3.2.1), somente uma parte desses transcritos haviam sido removidas da base de

dados *Homo_sapiens_hg38* devido incongruências ou redundâncias nos dados de alinhamento.

Devido a isso e também a necessidade de avaliação da nova abordagem utilizada para construção do SpliceProt, foram criadas as bases de dados *Homo_sapiens_hg38* e *Homo_sapiens_ENSEMBL*, nas quais foi aplicado o método para construção do SpliceProt.

5.2.1 Avaliação de um novo método para análise da confiabilidade das coordenadas dos transcritos

Para a reconstrução dos transcritos e de suas variantes por *splicing* alternativo, foi necessário determinar as coordenadas confiáveis de cada transcrito. Anteriormente, o método de construção do SpliceProt (TAVARES et al., 2014) utilizava duas bases de dados o RefSeq (O'LEARY et al., 2015) e o dbEST (BOGUSKI; LOWE; TOLSTOSHEV, 1993), ou seja, utilizava informações das sequências de transcritos curados manualmente e também de sequências geradas em projetos que utilizaram a abordagem de ESTs. Entretanto, o Ensembl define o grau de confiabilidade dos seus transcritos e a precisão da identificação das junções de *splicing* através de outros argumentos.

O TSL e o tamanho da sequência do transcrito foram escolhidos como primeiro e segundo parâmetros de confiabilidade para o novo método, respectivamente. Transcritos com valor de TSL igual a 1 possuíam certa equivalência com dados do repositório RefSeq (O'LEARY et al., 2015), uma vez que todas as junções entre os éxon eram validadas pela presença de um RNA confiável. Os demais valores de TSL incluíam transcritos com apenas evidências por EST ou com ausência de evidência estrutural (Quadro1).

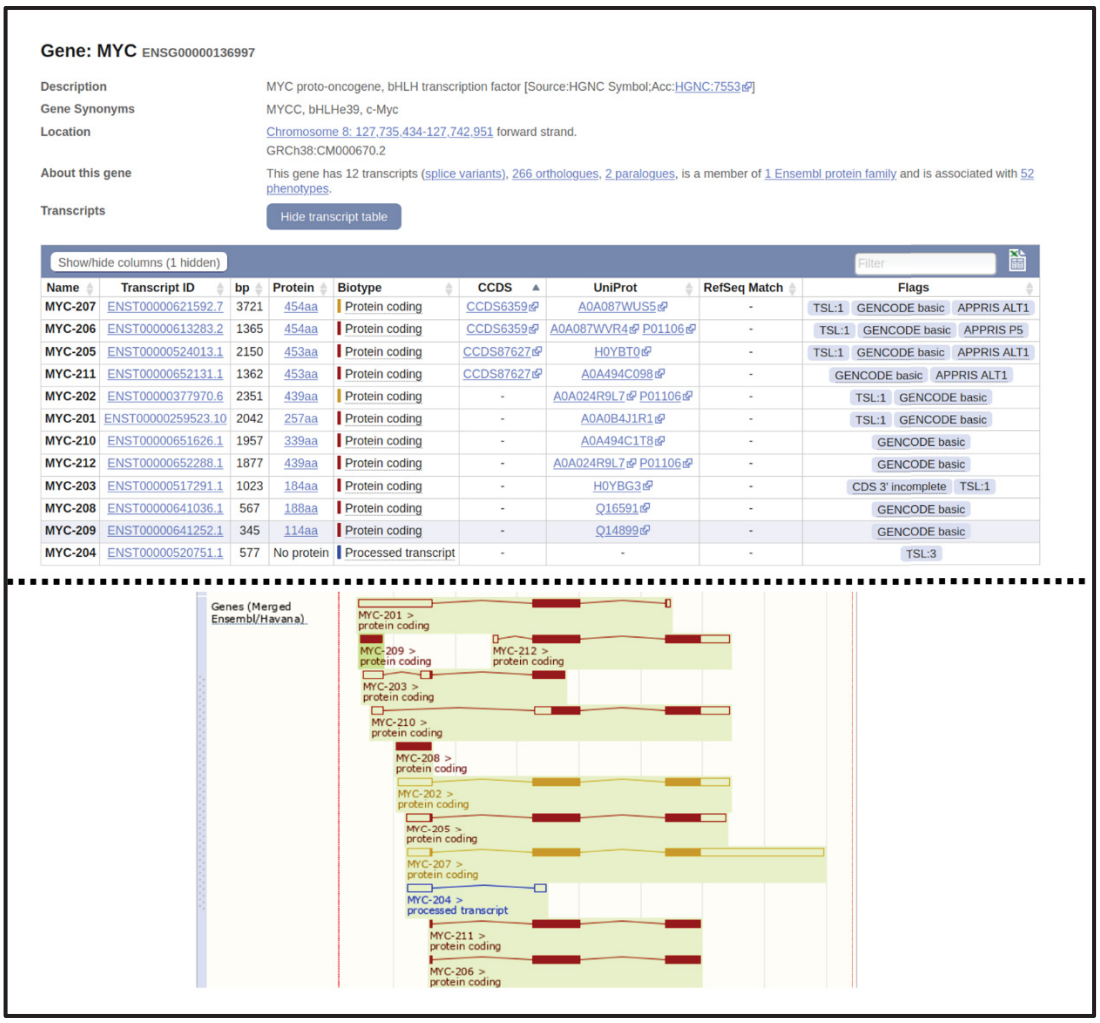
Somente para 207.368 transcritos (o que representa 93,2% dos transcritos presentes na base de dados *Homo_sapiens_hg38* e 99,47% dos transcritos da base de dados *Homo_sapiens_ENSEMBL*) foram obtidas informações para ambos os parâmetros de confiabilidade (TSL e sequência). Esses transcritos incluíam apenas transcritos anotados, pois somente esses possuíam valor de TSL, ainda que o campo de identificação desse argumento estivesse ausente na anotação de alguns transcritos (TSL NULL). Como não havia informação a respeito da confiabilidade das

sequências para o arquivo fasta (mRNA total ou EST) esses transcritos não foram considerados para a avaliação de coordenadas confiáveis.

5.2.2 Reconstrução dos transcritos e predição das variantes de *splicing* alternativo pelo método das matrizes ternárias.

Para as duas bases de dados, foi criado uma matriz com a representação ternária de cada transcrito (coluna matrizes, item 3.3.2). Utilizando a informação das coordenadas confiáveis, foi construído uma matriz com a representação ternária das variantes por *splicing* alternativo de cada gene (coluna variantes, item 3.3.2) (FIGURA 4). Ao final de execução do método, foi observado a reconstrução de matrizes com o valor 0 para todos os éxons do transcrito de determinados genes. Devido à baixa quantidade de ocorrências, esses transcritos foram avaliados e removidos das bases de dados manualmente. Foi observado que para esses casos, o transcrito foi reconstruído com ausência de algum éxon, pois possuíam apenas 1 éxon cujas coordenadas foram consideradas não confiáveis. (FIGURA 7). Na figura 13, está representado o gene ENSG00000136997 exemplificando um dos casos observados. O transcrito ENST00000641252.1 é definido como sendo TSL NULL e possui apenas 1 éxon. Em nosso método, as coordenadas desse transcrito foram consideradas não confiáveis. Ainda que o transcrito ENST00000641036.1 também possua somente 1 éxon, a coordenada da extremidade 3’ é confirmada pelo éxon 2 do transcrito ENST00000517291.1.

FIGURA 15 – REPRESENTAÇÃO NA BASE DE DADOS ONLINE DO PROJETO ENSEMBL DOS ÉXONS DO GENE ENSG00000136997.



FONTE: O autor (2020).

LEGENDA: O gene possui 12 transcritos, 6 deles possuem TSL1, 1 possui TSL3 e 5 transcritos foram considerados como TSL NULL. Em nosso método somente para o transcrito ENST00000621592.7 todas as coordenadas foram definidas como confiáveis.

5.2.3 Identificação de variantes contento PTC nas duas bases de dados.

Ainda que haja uma via alternativa de ativação de NMD que possa ser ativada independente da presença de PTC, a via canônica é ativada quando há presença de um códon de terminação à 50-55 nt à montante da última junção entre os éxons (DA COSTA; MENEZES; ROMÃO, 2017). As abordagens in silico para identificação de transcritos suscetíveis à via de NMD são direcionadas a identificação de transcritos

que possuam PTC. Ainda são necessários estudos para desenvolver ferramentas in silico para identificar transcritos que ativem a via alternativa de NMD (DE LA GRANGE et al., 2007; HSU; LIN; CHEN, 2017; RYAN et al., 2008).

O programa NMDClassifier (HSU; LIN; CHEN, 2017) foi utilizado para predizer as variantes que possuíam um PTC e eram potenciais alvo da via de NMD. A abordagem utilizada por esse programa está de acordo com os demais programas que possuem o mesmo objetivo. Entretanto, o programa NMDClassifier se destaca pelo foco em avaliar variantes por *splicing* alternativo. Além disso, Hsu e colaboradores (2017) avaliaram em seu trabalho que a busca por PTC por uma abordagem computacional utilizando uma distância de 50 ou 55 nucleotídeos à montante da última junção entre os éxons apresentaram uma acurácia em torno de 97,8%. Entretanto, outros métodos estatísticos apresentaram uma maior correlação associada a distância de 50 nucleotídeos (HSU; LIN; CHEN, 2017).

Devido à utilização de dados de entrada distintos para as duas bases de dados, foi observado uma diferença na quantidade e na variedade dos transcritos preditos como alvo da via de NMD. Entretanto, essa diferença possuía a mesma proporção da diferença entre a quantidade de transcritos presentes em cada base de dados (aproximadamente 8%).

A identificação da presença de um PTC é realizada pela busca na sequência do transcrito. Além disso, o NMDClassifier identifica o transcrito não sensível à via de NMD mais próximo do que foi predito como alvo desta via (*best-partner*). A partir da definição de um *best-partner* o programa realiza a predição do possível evento de *splicing* alternativo, que poderia ter originado a inserção do PTC. A definição do evento de *splicing* é realizada utilizando a premissa de que uma menor quantidade de eventos e as alterações mais simples possuem uma maior probabilidade de ocorrer. Vale destacar que eventos de inserção de PTC originados por mutação não são identificados pelo programa NMDClassifier (HSU; LIN; CHEN, 2017).

Para as duas bases de dados a alteração da região UTR dos transcritos foi a principal causa da inserção de PTC. Entretanto, esse evento reúne diversos tipos de alterações que incluem alterações nas regiões UTR, alterações que não causam a modificação na fase de leitura das proteínas, mas alteram a sequência codificadora e eventos que não puderam ser identificados em outras categorias (HSU; LIN; CHEN, 2017). Sendo assim, quando avaliado especificamente eventos de *splicing* alternativo, para as duas bases de dados, a inserção de um éxon (que pode incluir os eventos

éxons mutualmente exclusivos, uso alternativo de éxon) foi o principal responsável pela inserção de PTC nos transcritos. É interessante observar que dentre os eventos de *splicing* alternativo o uso alternativo de éxon possui a maior frequência no genoma (WANG et al., 2017).

5.2.4 Comparação entre a predição realizada pelo programa NMDClassifier e a anotação do projeto Ensembl (versão 96).

Para a base de dados *Homo_sapiens_hg38* apenas cerca de 13% dos transcritos preditos como alvo da via de NMD pelo NMDClassifier também estavam anotados da mesma forma pelo projeto Ensembl. Entretanto, cerca de 4,8% dos transcritos anotados pelo Ensembl como alvo da via de NMD nesta base de dados, não foram identificados pelo NMDClassifier. Quando considerado a base de dados *Homo_sapiens_ENSEMBL*, cerca de 2,3% dos transcritos apresentaram concordância como alvo da via de NMD com a anotação do Ensembl. Ao mesmo tempo, 5,2% dos transcritos que foram classificados como alvo da via de NMD pela anotação do Ensembl não foram identificados pelo NMDClassifier.

Utilizando o método descrito neste trabalho, para o gene ROBO3 o programa NMDClassifier identificou 8 variantes sensíveis a NMD na base de dados *Homo_sapiens_hg38* e 7 variantes na base de dados *Homo_sapiens_ENSEMBL* (TABELA 10). Entretanto, não foi observada anotação de nenhuma variante como alvo da via de NMD nos dados de anotação do projeto ENSEMBL (https://bit.ly/2ROQEcA). A TABELA 10 mostra exemplos de discordâncias da anotação do Ensembl e a predição realizada pelo programa NMDClassifier para as duas bases de dados, sublinhado estão os transcritos que foram preditos como alvo da via de NMD nas duas bases de dados. O gene ROBO3 está relacionado com a formação dos axônios na medula espinhal, e a expressão de variantes distintas está associada a regulação deste processo (COLAK et al., 2013). Na TABELA 10, a variante destacada em negrito foi classificada como presença de um PTC devido a retenção de um íntron. Além disso, Colak e colaboradores (2013) relataram que a via de NMD participa da regulação deste processo atuando sobre uma variante ROBO3.2 que possui um PTC devido a retenção de um íntron na variante (COLAK et al., 2013). O gene MX1 e LRPPRC também foram identificados nas duas bases de dados como possuindo variantes alvos da via de NMD, sendo discordante do anotado no projeto

Ensembl (MX1 - https://bit.ly/2ScdNoe; LRPPRC - https://bit.ly/2SdBHQn) (TABELA 10). Rausell e colaboradores (2014) avaliaram que a inserção de um PTC em uma variante do gene MX1 está associado a redução dos níveis da proteína e o aumento da patogenicidade do vírus Influenza A (RAUSELL et al., 2014). Gaweda-Walerych e colaboradores (2016) identificaram que eventos de *splicing* alternativo que estão associados a inserção de PTC na sequência do gene LRPPRC podem contribuir com a progressão da doença de Parkinson, devido a possível degradação da proteína (GAWEDA-WALERYCH et al., 2016).

Ainda que o parâmetro TSL das variantes que apresentaram resultados discordantes tenha sido observado, não foi observado nenhuma relação entre a diferença entre anotação do projeto Ensembl e a predição pelo programa NMDClassifier relacionada a este parâmetro.

TABELA 10 – EXEMPLOS DE VARIANTES PREDITAS COMO ALVO DA VIA DE NMD PELO SOFTWARE NMDCLASSIFIER CUJOS GENES NÃO POSSUEM VARIANTES ALVO DESTA VIA ANOTADOS NO PROJETO ENSEMBL.

Gene	Homo_sapiens_ hg38	Ensembl	tsl	Homo_sapiens_ ENSEMBL	Ensembl	tsl
ROBO3	ENST00000525304	int_ret	5	ENST00000397801	ptn_cod	1
	ENST00000525448	int_ret	1	ENST00000528068	int_ret	4
	ENST00000526551	process	2	ENST00000531119	int_ret	4
	ENST00000527196	int_ret	5	ENST00000531545	proces	2
	ENST00000528820	process	3	ENST00000531888	int_ret	1
	ENST00000531119	int_ret	4	ENST00000532472	process	4
	ENST00000531545	process	2	ENST00000538940	ptn_cod	5
	ENST00000534598	int_ret	2			
MX1	ENST00000413378	ptn_cod	5	ENST00000288383	ptn_cod	2
	ENST00000417963	ptn_cod	5	ENST00000398598	ptn_cod	1
	ENST00000424365	ptn_cod	5	ENST00000398600	ptn_cod	2
	ENST00000486275	int_ret	5	ENST00000413778	ptn_cod	5
	ENST00000619682	ptn_cod	5	ENST00000455164	ptn_cod	1
				ENST00000478268	int_ret	2
				ENST00000484465	int_ret	3
				ENST00000490220	process	5
				ENST00000619682	ptn_cod	5
LRPPRC	ENST00000419884	ptn_cod	5	ENST00000409659	ptn_cod	5
	ENST00000447246	ptn_cod	1	ENST00000409946	ptn_cod	1
	ENST00000467058	int_ret	3			

Int_ret – retenção de íntron; process – transcrito processado; ptn_cod – codificador de proteína.

Fonte: O autor (2020)

Por outro lado, ao avaliar o gene WT1, por exemplo, o programa NMDClassifier não identificou nenhuma variante alvo da via de NMD, diferente do observado na anotação do projeto Ensembl (https://bit.ly/393o7Ga). O gene WT1 é um gene supressor de tumor associado ao desenvolvimento do tumor de Wilms (nefroblastoma) e já foi identificado a presença de uma variante deste gene que é alvo da via de NMD (JIN KIM; ALINOOR RAHMAN, 2018).

Não foi sido observado um padrão claro de diferenças entre a anotação no projeto Ensembl para as variantes preditas como alvo da via de NMD em nossas bases de dados. Entretanto as diferenças quando à anotação de retenção de íntron do projeto Ensembl merecem destaque. Segundo a documentação do projeto HAVANA, ainda que casos de retenção de íntron resultem em transcritos que sejam alvos da via de NMD eles são preferencialmente anotados como retenção de íntron (HAVANA PROJECT, 2016). Ainda assim, o programa NMDClassifier não foi capaz de identificar alguns alvos validados da via de NMD.

5.2.5 Comparação da localização genômica entre os alvos da via de NMD entre as bases de dados e os dados de anotação do projeto Ensembl (versão 96).

A inserção de um PTC poder estar relacionada a mudança da fase de leitura da tradução associada a adição/remoção de apenas um nucleotídeo. Sendo assim, e devido as diferenças observadas entre a predição realizada pelo NMDClassifier e os dados de anotação do projeto Ensembl a localização genômica das variantes foram comparadas entre as bases de dados e a respectiva anotação do projeto Ensembl.

Adicionalmente, a base de dados *Homo_sapiens_hg38* foi construída a partir de um método de alinhamento utilizando o programa Blat (Kent, 2002), o que resultou em diferenças quanto a localização genômica de determinados transcritos quando comparado com os dados de anotação. Além disso, os métodos utilizados para reconstrução ternária dos transcritos nas duas bases de dados consideraram apenas as coordenadas externas extremas (5’ do primeiro éxon e 3’ do último éxon) do maior transcrito com TLS1. Essa abordagem resultou na ausência de coordenadas validadas na TABELA relacional CONSENSO das duas bases de dados criadas. Para a predição e identificação de um PTC é de grande importância que a reconstrução das coordenadas e sequência dos transcritos sejam verossímeis (DE LA GRANGE et al., 2007; HSU; LIN; CHEN, 2017). Desta forma, tal etapa ainda precisa ser

aprimorada antes da predição e seleção de variantes suscetíveis a degradação pela via de NMD pela abordagem desenvolvida por Hsu e colaboradores (2017).

5.2.6 Tradução das variantes por *splicing* alternativo preditas pela metodologia de matrizes ternárias.

O método de construção de matrizes ternárias permitiu identificar diversas redundâncias quanto a identificação de variantes por *splicing* alternativo nos dados provenientes do projeto Ensembl. Como ilustrado na figura 9, os transcritos ENST00000482924 e ENST000003100 mesmo que possuam conjuntos de éxons diferentes, segundo a metodologia empregada por esse trabalho, a tradução hipotética dos trancritos resulta na mesma sequência de proteína.

A partir da reconstrução dos transcritos pelo método de Tavares e colaboradores (2014), foi possível observar a tradução de diversas variantes que não estavam anotadas como codificadoras de proteína no projeto Ensembl. A identificação do potencial codificador de variantes por *splicing* alternativo é de grande importância para a construção de bancos de dados de sequência de proteína (BLACK, 2000; WANG et al., 2012). A identificação de novos peptídeos em experimentos de proteômica *shotgun* está diretamente relacionado a utilização de bancos de dados de proteínas verossímeis e que sejam capazes de identificar diferentes proteoformas (NESVIZHSKII, 2014; WANG et al., 2012).

A abordagem desenvolvida por Tavares e colaboradores (2014) em sua primeira versão permitiu identificar, in silico, novos peptídeos que não estavam anotados nas bases de dados UniprotKB/TrEMBL (THE UNIPROT CONSORTIUM, 2019). Neste trabalho, essa abordagem foi atualizada e aprimorada para utilizar dados do projeto Ensembl e identificar variantes que sejam alvo em potencial da via de NMD. Ainda que precise ser aprimorada, a abordagem realizada neste trabalho foi capaz de identificar variantes por *splicing* alternativo redundantes na base de dados do projeto Ensembl (FRANKISH et al., 2019). Além disso, para as duas bases de dados criadas foram identificadas mais de 35.000 variantes por *splicing* alternativo que apresentavam uma proteína hipotética e não estavam anotadas como “codificadoras de proteínas” no projeto Ensembl. Isso demonstra que método aprimorado neste trabalho é de relevância para a construção de um banco de dados de proteínas que

pode contribuir para a identificação de novos peptídeos em abordagens de proteômica *shotgun*.

6 CONSIDERAÇÕES FINAIS

A abordagem utilizada por este trabalho se mostrou eficiente para reconstrução de variantes por *splicing* alternativo. Foi possível adaptar o fluxo de processamento criado por Tavares e colaboradores (2014) e atualizar o repositório SpliceProt para receber dados do projeto Ensembl.

Utilizando a metodologia das matrizes ternárias foi possível reconstruir as coordenadas dos transcritos e realizar a predição de variantes por *splicing* alternativo a partir de dados de sequência e anotação do projeto Ensembl.

Não foi encontrado um parâmetro de classificação para definir um conjunto de dados ótimos entre dados provenientes de alinhamento (arquivo em formato fasta) ou de anotação (arquivo em formato gtf). Devido a diferença encontrada entre as duas bases de dados, principalmente quanto as coordenadas das sequências, dada a definição do parâmetro TSL (Quadro 1) (FRANKISH et al., 2019) sugerimos que seja construída uma única base de dados contendo transcritos com transcritos com valor de TSL igual a 1 provenientes de dados de anotação e dos demais provenientes de dados de alinhamento. Para uma melhor definição, tais dados necessitariam de serem avaliados manualmente e validados experimentalmente, etapas a serem consideradas após a otimização do método aqui descrito.

Os parâmetros TSL e tamanho da sequência de nucleotídeos foram eficientes para a seleção das coordenadas confiáveis dos transcritos. Entretanto, foi observado que na próxima versão do SpliceProt todos os transcritos com valor de TSL igual a 1 devem ser considerados para recuperar coordenadas confiáveis, não somente àquele com a maior sequência. Ainda assim, o método de matrizes ternárias foi capaz de identificar as variantes por *splicing* alternativo das bases de dados criadas e identificar as redundâncias presentes nos dados.

Além disso, foi possível aprimorar o método de desenvolvimento do SpliceProt adicionando em seu fluxo de trabalho o programa NMDClassifier (HSU; LIN; CHEN, 2017) que permite identificar variantes que seriam degradadas por possuir um PTC. De fato, após realizar a tradução das variantes foi identificado a presença de que no conjunto de dados haviam variantes preditas como alvo da via de NMD. Sendo assim,

foi criado um novo conjunto de dados com proteínas hipotéticas traduzidas *in silico* retirando aquelas cujos transcritos são alvos da via de NMD.

Entretanto, após observar os dados apresentados nesta dissertação, sugerimos a utilização em conjunto da anotação do projeto Ensembl com a predição realizada pelo programa NMDClassifier. Ainda mais, em casos de transcritos que possuam valor de TSL igual a 1 e sejam codificadores de proteína, a anotação do Ensembl deve prevalecer sobre o método de predição. Por outro lado, quando os transcritos apresentarem outros valores de TSL e não forem anotados como codificadores de proteína, a predição realizada pelo programa NMDClassifier (HSU; LIN; CHEN, 2017) a partir dos dados de base de dados construída nesse trabalho deve ser considerada. Isso se deve ao fato que o método de predição está relacionado diretamente à informação de sequência do transcrito, e com o método realizado neste trabalho, foi possível predizer a sequência de variantes por *splicing* alternativo com informações apenas de EST.

As etapas finais do aprimoramento do repositório SpliceProt serão realizadas durante o doutorado da doutoranda Letícia Santos em conjunto com o doutorado deste estudante.

REFERÊNCIAS

BLACK, D. L. Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biologyCell, 2000.
BOGUSKI, M. S.; LOWE, T. M. J.; TOLSTOSHEV, C. M. dbEST — database for “expressed sequence tags”Nature Genetics, 1993.
CHAN, D. et al. A nonsense mutation in the carboxyl-terminal domain of type X collagen causes haploinsufficiency in schmid metaphyseal chondrodysplasia. The Journal of clinical investigation, v. 101, n. 16, p. 1490–1499, 1998.
CHANG, Y.-F.; IMAM, J. S.; WILKINSON, M. F. The Nonsense-Mediated Decay RNA Surveillance Pathway. Annual Review of Biochemistry, 2007.
CHEN, M.; MANLEY, J. L. Mechanisms of alternative splicing regulation: Insights from molecular and genomics approachesNature Reviews Molecular Cell Biology, 2009.
COLAK, D. et al. Regulation of Axon Guidance by Compartmentalized Nonsense-Mediated mRNA Decay. 2013.
DA COSTA, P. J.; MENEZES, J.; ROMÃO, L. The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. International Journal of Biochemistry and Cell Biology, v. 91, p. 168–175, 2017.
DE LA GRANGE, P. et al. A new advance in alternative splicing databases: from catalogue to detailed analysis of regulation of expression and function of human alternative splicing variants. 2007.
FAUSTINO, N. A.; COOPER, T. A. Pre-mRNA splicing and human disease. Genes & Development, 2003.
FLOREA, L. et al. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Research, 1998.
FRANKISH, A. et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research, v. 47, p. 767, 2019.
GAWEDA-WALERYCH, K. et al. Parkinson’s disease–related gene variants influence pre-mRNA splicing processes. Neurobiology of Aging, v. 47, p. 127–138, 1 nov. 2016.
GE, Y.; PORSE, B. T. The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression. BioEssays, v. 36, n. 3, p. 236–243, 2014.
HAAS, B. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research, v. 31, n. 19, p. 5654–5666, 2003.
HAVANA PROJECT. havana Annotation Guidelines. v. 24, n. March, p. 1–30, 2016.
HE, F.; JACOBSON, A. Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. Annual Review of Genetics, v. 49, n. 1, p. 339–366, 2015.
HSU, M. K.; LIN, H. Y.; CHEN, F. C. NMD Classifier: A reliable and systematic classification tool for nonsense-mediated decay events. PLoS ONE, v. 12, n. 4, p. e0174798, 1 mar. 2017.
HUNT, S. E. et al. Ensembl variation resources. Database : the journal of biological databases and curation, v. 2018, p. 1–12, 2018.
HUUSKO, P. et al. Nonsense-mediated decay microarray analysis identifies mutations of EPHB2 in human prostate cancer. Nature Genetics, 2004.
IDEUE, T. et al. Introns play an essential role in splicing-dependent formation of the exon junction complex. 2007.

JIN KIM, Y.; ALINOOR RAHMAN, M. NMD in disease and potential therapies. 2018.
KASTNER, B. et al. Structural Insights into Nuclear pre-mRNA Splicing in Higher EukaryotesCold Spring Harbor perspectives in biology, 2019. Disponível em: <http://cshperspectives.cshlp.org/>. Acesso em: 13 dez. 2019
KENT, W. J. BLAT - The BLAST-like alignment tool. Genome Research, 2002.
KERR, T. P. et al. Long mutant dystrophins and variable phenotypes: Evasion of nonsense-mediated decay? Human Genetics, v. 109, n. 4, p. 402–407, 2001.
KUROSAKI, T.; MAQUAT, L. E. Nonsense-mediated mRNA decay in humans at a glance. Journal of Cell Science, v. 129, n. 3, p. 461–467, 2016.
LAU, E. et al. Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. 2019.
LEE, C.; WANG, Q. Bioinformatics analysis of alternative splicing. Briefings in Bioinformatics, 2005.
LIU, C. et al. The UPF1 RNA surveillance gene is commonly mutated in pancreatic adenosquamous carcinoma. Nature Medicine, 2014.
LIU, Y. et al. Impact of Alternative Splicing on the Human Proteome. Cell Reports, v. 20, n. 5, p. 1229–1241, 1 ago. 2017.
MILLER, J. N.; PEARCE, D. A. Nonsense-Mediated Decay in Genetic Disease : Friend or Foe ? Mutation Research - Reviews in Mutation Research, v. 762, p. 52–64, 2014.
NESVIZHSKII, A. I. Proteogenomics: Concepts, applications and computational strategiesNature MethodsNature Publishing Group, , 1 nov. 2014.
NOENSIE, E. N.; DIETZ, H. C. A strategy for disease gene identification through nonsense-mediated mRNA decay inhibition. Nature biotechnology, 2001.
O'LEARY, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, v. 44, p. 733–745, 2015.
PARKINSON, J.; BLAXTER, M. Expressed sequence tags: An overview. Methods in Molecular Biology, v. 533, p. 1–12, 2009.
PONTIUS, J. U.; WAGNER, L.; SCHULER, G. D. UniGene: a unified view of the transcriptome. The NCBI Handbook, 2003.
RAUSELL, A. et al. Analysis of Stop-Gain and Frameshift Variants in Human Innate Immunity Genes. PLoS Computational Biology, 2014.
RICE, P.; LONGDEN, L.; BLEASBY, A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics, 2000.
RYAN, M. C. et al. SpliceCenter: A suite of web-based bioinformatic applications for evaluating the impact of alternative splicing on RT-PCR, RNAi, microarray, and peptide-based studies. 2008.
SCHRIJVER, I. et al. Premature termination mutations in FBN1: distinct effects on differential allelic expression and on protein and clinical phenotypes. American journal of human genetics, v. 71, n. 2, p. 223–37, 2002.
SHEYNKMAN, G. M. et al. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. Molecular and Cellular Proteomics, v. 12, n. 8, p. 2341–2353, 2013.
SMITH, J. E.; BAKER, K. E. Nonsense-mediated RNA decay-a switch and dial for regulating gene expression. BioEssays : news and reviews in molecular, cellular and developmental biology, v. 37, n. 6, p. 612–23, 2015.
STANKE, M. et al. AUGUSTUS: A b initio prediction of alternative transcripts. Nucleic Acids Research, v. 34, n. WEB. SERV. ISS., 2006.

TAVARES, R. et al. SpliceProt: A protein sequence repository of predicted human splice variants. Proteomics, p. 181–185, 2014.
THE UNIPROT CONSORTIUM. UniProt: a worldwide hub of protein knowledge The UniProt Consortium. Nucleic Acids Research, 2019.
THORVALDSDÓTTIR, H.; ROBINSON, J. T.; MESIROV, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Briefings in Bioinformatics, 2013.
TRESS, M. L.; ABASCAL, F.; VALENCIA, A. Alternative splicing may not be the key to proteome complexity HHS Public Access. Trends Biochem Sci, v. 42, n. 2, p. 98–110, 2017.
VAN DEN HOOGENHOF, M. M. G.; PINTO, Y. M.; CREEMERS, E. E. RNA Splicing regulation and dysregulation in the heart. Circulation Research, v. 118, n. 3, p. 454–468, 2016.
VITTING-SEERUP, K. et al. SpliceR: An R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. BMC Bioinformatics, 2014.
WANG, B. D.; LEE, N. H. Aberrant RNA splicing in cancer and drug resistance. Cancers, v. 10, n. 11, 2018.
WANG, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature, v. 456, n. 7221, p. 470–476, 2008.
WANG, J. et al. Computational methods and correlation of Exon-skipping events with splicing, transcription, and epigenetic factors. In: Methods in Molecular Biology. [s.l: s.n.]. v. 1513p. 163–170.
WANG, X. et al. Protein identification using customized protein sequence databases derived from RNA-seq data. Journal of Proteome Research, v. 11, n. 2, p. 1009–1017, 2012.
WANG, Y. et al. Mechanism of alternative splicing and its regulation. Biomedical Reports, v. 3, n. 2, p. 152–158, 2015.
WANG, Z.; BURGE, C. B. Splicing regulation: From a parts list of regulatory elements to an integrated splicing codeRNA, 2008.
XU, X. et al. Exome sequencing identifies UPF3B as the causative gene for a Chinese non-syndrome mental retardation pedigree. Clinical Genetics, 2013.
ZERBINO, D. R. et al. Ensembl 2018. Nucleic Acids Research, v. 46, n. D1, p. D754–D761, 4 jan. 2018.
ZHENG, Q. F. et al. Epigenetic alterations contribute to promoter activity of imprinting gene IGF2. Biochimica et Biophysica Acta - Gene Regulatory Mechanisms, v. 1861, n. 2, p. 117–124, 2018.