MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Doutorado pelo programa de Pós-graduação em Biologia Computacional e Sistemas

Análise de variantes de *splicing* em homem e camundongo por uma abordagem de proteogenômica

RAPHAEL TAVARES DA SILVA

Rio de Janeiro
2016

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

RAPHAEL TAVARES DA SILVA

Análise de variantes de *splicing* em homem e camundongo
por uma abordagem de proteogenômica.

Tese apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em Ciências.

Orientadores: Prof. Dr. Fabio Passetti
Profª. Dra. Nicole de Miranda Scherer

RIO DE JANEIRO
Junho/2016

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

S586 Silva, Raphael Tavares da

Análise de variantes de splicing em homem e camundongo por uma abordagem de proteogenômica / Raphael Tavares da Silva. – Rio de Janeiro, 2016.
xx, 143 f. : il. ; 30 cm.

Tese (Doutorado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2016.
Bibliografia: f. 108-121

1. Proteogenômica. 2. Bioinformática. 3. Biologia computacional. 4. Splicing alternativo. 5. RNA-Seq. 6. Next-generation sequencing. 7. Espectrometria de massas. I. Título.

CDD 572.65

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: RAPHAEL TAVARES DA SILVA

Análise de variantes de *splicing* em homem e camundongo

por uma abordagem de proteogenômica

ORIENTADORES: Prof. Dr. Fabio Passetti
Profª. Dra. Nicole de Miranda Scherer

Aprovada em: 24/06/2016

EXAMINADORES:
Prof. Dr. Jerônimo Conceição Ruiz - **Presidente** (FIOCRUZ/MG)
Prof. Dr. Magno Rodriguez Junqueira (UFRJ)
Prof. Dr. Fabrício Klerynton Marchini (FIOCRUZ/PR)
Profª. Dra. Ana Carolina Paulo Vicente (FIOCRUZ/RJ)
Prof. Dr. Eduardo de Matos Nogueira (UNIRIO)

Rio de Janeiro, 24 de junho de 2016

Dedico este trabalho aos meus pais pelo amor incondicional e por estar sempre ao meu lado

AGRADECIMENTOS

Meus agradecimentos não se resumem aos quatro anos de doutorado, mas sim aos onze anos na academia desde o vestibular até a entrega desta tese. Impossível terminar o doutorado num país como o Brasil e não ser grato por poder estudar e refletir que esta é uma oportunidade (infelizmente) para poucos. Espero algum dia, de alguma forma, passar o que aprendi para outras pessoas que não tiveram a mesma "sorte" que a minha.

Sendo assim, meu primeiro agradecimento são dedicados aos meus pais Armênio e Aida. Obrigado, sobretudo pelo amor, carinho, educação e pelos valores que aprendi com vocês. Pensei mil vezes em desistir da carreira acadêmica, mas vocês estavam sempre ali ao meu lado segurando a onda. Nunca me cobravam nada, apenas que eu me sentisse feliz com as minhas escolhas. Quando eu olhava para trás e via a história da nossa família, de onde viemos e tudo o que vocês me proporcionaram, só me fez pensar se eu dei o meu melhor. Minha consciência diz que sim e espero ter feito por merecer. Obrigado pela vida! Amo vocês!

À minha irmã Aline, por sempre ter os melhores conselhos e por ser a melhor irmã que eu poderia desejar! Obrigado pelo carinho e amizade que compartilhamos durante todos esses anos! Muito do que você me ensinou também está aqui! Ah! E obrigado por ter me dado o sobrinho mais alegre e divertido do mundo: Pedro-pedrudo!

À Thiara, minha namorada e companheira. Dividimos muito até agora e espero poder dividir muito mais daqui pra frente!

Ao meu orientador, Dr. Fabio Passetti. Obrigado pelo aprendizado acadêmico e pelas lições que não estão nos artigos científicos. Obrigado pelo exemplo e pelo compromisso.

À minha coorientadora, Dra. Nicole Scherer. Obrigado por ter me salvado algumas vezes, mas principalmente por ter "disciplinado" minha programação. Ainda não faço programas tão refinados e organizados como você, mas hoje me sinto mais seguro para fazer isso. Pode deixar que eu nunca vou me esquecer de três coisas: "use strict", "use warnings" e seu bom humor!

Aos meus amigos do antigo LBBC! Obrigado à Natasha, Gabriel Wajnberg por me ajudarem ao longo desses anos e por dividirem comigo as vitórias e aflições da vida acadêmica. Um agradecimento especial ao Gabriel (que como cantor é um excelente Bioinformata) pela força no último artigo publicado e pela amizade. Valeu mesmo! Um agradecimento também aos amigos que já passaram pelo LBBC: Edson, Gabriel Renaud, Gabriel Lima-Verde, Gabriel Espíndola e Pedrão!

Aos amigos do LGFB que receberam de braços abertos o grupo do LBBC. Obrigado pelo ambiente descontraído, piadas e momentos de café expresso na veia. Nada mais justo que terminar meu último ano de doutorado na FIOCRUZ e neste laboratório.

Aos amigos que fiz no INCA. Obrigado pela amizade, conversas e risadas no alojamento!

Aos meus eternos amigos do segundo grau: Allan, Sandro, Ana, Cássia, Thomaz, Bia, Bruno e Fernanda! Quase quinze anos depois e o estoque de risadas não acaba! Que sejam eternas as conversas fiadas, o carinho entre a gente e os sítios, é claro! E obrigado aos agregados também porque aguentar essa galera não é pra qualquer um! A alegria de vocês está em cada página da tese! Obrigado pela amizade!

À Fundação Oswaldo Cruz, ao Instituto Oswaldo Cruz e a Pós-graduação em Biologia Computacional e de Sistemas por terem fornecido todo suporte e apoio.

Ao Instituto Nacional de Câncer por ter fornecido toda a infraestrutura e suporte ao longo dos oito anos que desenvolvi meus projetos de iniciação científica, mestrado e doutorado.

Ao Dr. Carlos Gil por ter acreditado e investido no antigo Laboratório de Bioinformática e Biologia Computacional.

À Dra. Adriana Paes-Leme e à Dra. Bianca Pauletti pela colaboração em dois artigos publicados pelo nosso grupo.

Às agências de fomento por tornarem possível a viabilização deste projeto.

"Não esmorecer para não desmerecer"
(Oswaldo Cruz)

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Análise de variantes de *splicing* em homem e camundongo por uma abordagem de proteogenômica

RESUMO

TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Raphael Tavares da Silva

Os avanços obtidos no estudo do transcriptoma pelo uso de sequenciadores de alta vazão e na proteômica por meio da espectrometria de massas, resultaram num grande volume de dados que passou a ser integrado em diversos estudos na Bioinformática. A proteogenômica é a área da pesquisa que reúne essas tecnologias, atuando na interface entre a genômica e a proteômica para interpretar eventos moleculares como, por exemplo, o *splicing* alternativo. Este evento é capaz de gerar RNAs mensageiros diferentes de um mesmo gene, podendo alterar a sequência polipeptídica após a tradução e consequentemente afetar a função das proteínas. Estudos utilizando dados de RNA-Seq, estimaram que até 90% dos genes humanos sejam afetados por este evento, expandindo assim, a capacidade de geração de proteínas com diferentes funções. Esta tese reúne o desenvolvimento e a aplicação de diferentes abordagens voltadas para a identificação de variantes de *splicing* em dados de espectrometria de massas de linhagens celulares e amostras de diferentes tecidos de homem e camundongo. Para tal, foram criados repositórios personalizados de sequências proteicas, constituídos por sequências canônicas e peptídeos digeridos *in silico* derivados de isoformas preditas computacionalmente. O primeiro repositório personalizado foi aplicado em dados de espectrometria de massas de uma linhagem de células T. Foram identificados 54 peptídeos oriundos de variantes de *splicing* que não seriam identificados utilizando repositórios tradicionais. O segundo repositório personalizado foi aplicado em dados de espectrometria de massas de uma linhagem celular de oligodendrócitos humanos. Foram identificadas 39 isoformas que apresentaram um perfil de atuação no citoesqueleto desse tipo celular, além de terem sua função discutida no âmbito do tecido cerebral. Algumas destas variantes de *splicing* tiveram a expressão de seus mRNAs confirmada experimentalmente (*EEF1D*, *KRAS*, *MFF*, *SDR39U1* e *SUGT1*). Ademais, foram apresentadas propostas para atribuir maior confiabilidade às isoformas encontradas a partir do número de espectros e peptídeos únicos. O terceiro e quarto repositórios personalizados foram usados em dados de espectrometria de massas das regiões cerebrais de homem e camundongo. Para a composição desses repositórios foi utilizada a montagem de transcriptoma com genoma de referência e *de novo* para reconstrução de transcritos provenientes de corridas de RNA-Seq. Foram identificadas variantes de *splicing* já conhecidas e exclusivas derivadas da montagem de transcriptoma. Entre os genes das isoformas encontradas na região do corpo caloso de homem e camundongo, sete foram identificados como ortólogos (*CDC42*, *TPM3*, *EEF1D*, *PKM*, *SEPT7*, *SET* e *RUFY3*). Até o momento, as abordagens desenvolvidas indicam que a utilização de

repositórios proteicos personalizados e a montagem de transcriptoma contribuem para a identificação de isoformas anotadas e potencias variantes de *splicing.*

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Analysis of splicing variants of human and mouse by a proteogenomics approach

ABSTRACT

PHD THESIS IN SYSTEMS AND COMPUTATIONAL BIOLOGY

RAPHAEL TAVARES DA SILVA

Technological improvements in data generation in transcriptomics by next generation sequencers, and in proteomics by mass spectrometry resulted in a large volume of data, which is progressively becaming integrated in Bioinformatics studies. Proteogenomics is a research area in which these technologies are combined, acting in the interface of genomics and proteomics to elucidate molecular events, for instance, alternative splicing. This event is capable to alternatively process different mRNAs from the same gene, and its translation to polypeptide sequence may be affected, consequently influencing protein function. Studies using RNA-Seq estimated that at least 90% of human genes are subject to alternative splicing, expanding the number of functionally different proteins. The aim of this thesis is the development and application of different approaches focused on alternative splicing isoforms in mass spectrometry data from cell lines and different samples from human and mouse tissues. For this purpose, we designed customized protein sequence repositories, composed by canonical sequences and *in silico* digested peptides from predicted isoforms. The first customized repository was applied to mass spectrometry data of a T cell line. We were able to identify 54 peptides derived by splicing variants, which could not be detected using canonical protein sequence repositories. The second customized repository was applied to mass spectrometry data of a human oligodendrocyte cell line. The 39 isoforms found are related to cytoskeleton in this cell type, and their functions in the cerebral tissue context have been considered. Five of these splicing variants were experimentally validated (EEF1D, KRAS, MFF, SDR39U1 e SUGT1). We propose an approach to assign reliability to the isoforms identified by their number of unique peptides and spectra. The third customized repository was applied to mass spectrometry data of brain regions from human and mouse. For this purpose, genome-guided and *de novo* transcriptome assembly were used to reconstruct transcripts from RNA-Seq data. We detected annotated alternative splicing isoforms as well as isoforms predicted by the computational transcriptome assembly. Comparing the isoforms identified in corpus callosum of human and mouse, seven were classified as orthologs (CDC42, TPM3, EEF1D, PKM, SEPT7, SET e RUFY3). The developed approaches show that customized protein repositories and transcriptome assembly contribute to identify annotated isoforms and potential alternative splicing variants.

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

LISTAS DE SIGLAS E ABREVIATURAS

ATL – Lobo Temporal Anterior, do Inglês “Anterior Temporal Lobe”

AUGUSTUS – preditor gênico utilizado para avaliação

BAM – do Inglês “Binary Alignment/Map”

BLAST – do Inglês “Basic Local Alignment Search Tool”

BLAT – do Inglês “BLAST-like alignment tool”

UCSC – Universidade da Califórnia, *campus* Santa Cruz

CC – Corpo Caloso, do Inglês “Corpus Callosum”

cDNA – DNA complementar, do Inglês “complementary DNA”

CDS – sequencia codificadora, do Inglês “Coding Sequence”

C-HUPO - Chromosome-based Human Proteome Project

CID – do Inglês “Collision Induced Dissociation”

DNA – Ácido Desoxiribonucléico, do Inglês “deoxyribonucleic acid”

EMBL – do Inglês “European Molecular Biology Laboratory”

EMBOSS – do Inglês “European Molecular Biology Open Software Suite”

ESTs – do Inglês “Expressed Sequence Tag”

FDR - Taxa de falso positivos, do Inglês “False Discovery Rate”

FIOCRUZ – Fundação Oswaldo Cruz

GO – do Inglês “Gene Ontology”

GSEA – do Inglês “Gene Set Enrichment Analysis”

GTF – do Inglês “Gene Transfer Format”

INCA – Instituto Nacional de Câncer

INPI – Instituto Nacional da Propriedade Industrial

IPKB – do Inglês “Ingenuity Pathways Knowledge Base”

mRNA – RNA mensageiro, do Inglês “messenger RNA”

MS - Espectrometria de Massas, do Inglês, “Mass Espectrometry”

ncRNAs – RNAs não codificantes, do Inglês “non-coding RNA”

NGS – Sequenciamento de alta vazão, do Inglês “Next Generation Sequencing”

NM – mRNAs completos, validados experimentalmente e de alta confiabilidade

NP – sequências de aminoácidos integrais ou parciais de proteínas, também validados experimentalmente e de alta confiabilidade

PASA – do Inglês “Program to Assemble Spliced Alignments”

PCR - Reação em cadeia da polimerase, do Inglês “Polymerase Chain Reaction”

PEP - probabilidade de erro posterior, do Inglês “Posterior Error Probability”

Perl – do Inglês “Practical Extration and Report Language”

PSA – do Inglês “Prostate-specific Antigen”

PSM – do Inglês, “Peptide Spetrum Match”

RefSeq – do Inglês, “Reference Sequence”

RNA – Ácido Ribonucléico, do Inglês “ribonucleic acid”

RNA-Seq – Sequenciamento RNA de alta vazão, do Inglês “RNA sequencing”

rRNA – RNA ribossomal, do Inglês “ribosomal RNA”

RT-qPCR – do Inglês “Real time quantitative PCR”

SAM – do Inglês “Sequence Alignment/Map”

SGBD – Sistema Gerenciador de Banco de Dados

SNP – Polimorfismo de nucleotídeo único, do Inglês “Single Nucleotide Polymorphism”

SRA – do Inglês “Sequence Read Archive”

TRANSEQ – preditor gênico utilizado para avaliação

tRNA - RNA transportador, do Inglês “transporter RNA”

UCSC – Universidade da Califórnia *campus* Santa Cruz, do Inglês “University of California, Santa Cruz”

VEGF – fator de crescimento endotelial vascular, do Inglês “Vascular Endothelial Growth Gactor”

1. INTRODUÇÃO

1.1 - Visão geral

O grande volume de dados produzidos pelas diferentes plataformas na genômica, transcriptômica e proteômica, criou desafios quanto ao seu armazenamento, processamento, visualização e interpretação. A conexão entre as informações obtidas por cada área conduz a uma necessária e frequente transferência de conhecimento entre diferentes campos da pesquisa em Bioinformática e Biologia Computacional.

O transcriptoma pode ser definido como um conjunto de todas as moléculas de RNA expressas em uma determinada célula, tecido ou em um organismo como todo (Blencowe *et al*., 2009). O transcriptoma de um organismo é composto por RNAs de diferentes origens e funções, podendo ser divididos em duas classes principais: mensageiros (mRNAs) e não codificadores de proteínas (ncRNAs) (revisto por Wery, 2011). Os ncRNAs vêm sendo caracterizados como importantes reguladores da atividade celular e ainda precisam de estudos mais aprofundados para o seu melhor entendimento apesar de algumas classes como, por exemplo, tRNA e rRNA já serem extensivamente estudados. Já os mRNAs compõem uma fração do transcriptoma que é responsável pela produção protéica de uma célula. Muito se sabe sobre os mecanismos de maturação de um mRNA, que em geral envolve a adição de uma guanina modificada na extremidade 5', a inclusão de uma cauda de adeninas na extremidade 3' e a retirada de regiões (íntrons) dos pré-mRNAs, através de um evento molecular conhecido como *splicing*.

O *splicing* alternativo é um dos mecanismos moleculares capazes de aumentar significativamente o repertório protéico de um organismo (revisto por Hallegger *et al*., 2010). Este mecanismo ocorre geralmente durante a transcrição, pelo processamento alternativo de íntrons no pré-mRNA. Assim, diferentes transcritos (isoformas) são produzidos a partir de um mesmo gene, podendo ser traduzidos em proteínas diferentes. Desta forma, especula-se que o *splicing* alternativo seja um dos mecanismos que mais influencie na mudança do conjunto das cadeias polipeptídicas (revisto por Black, 2000).

Com o advento das novas tecnologias de sequenciamento conhecidas como *Next Generation Sequencing* (NGS) (revisto por Kircher & Kelso, 2010), o volume de informação disponível para a investigação e aplicação de estudos voltados para o genoma e transcriptoma humano aumentou drasticamente durante a última década.

No que tange ao mecanismo de *splicing* alternativo, a possibilidade da descoberta de isoformas inéditas faz com que esta tecnologia venha a ser utilizada como uma importante estratégia para a identificação de novos biomarcadores, auxiliando no acompanhamento de patologias e também para o desenvolvimento de novas ferramentas de diagnóstico (Tavares *et al*., 2015; Wajnberg e Passetti, 2016).

A Espectrometria de Massas (MS) (do Inglês, *Mass Espectrometry*) tem contribuído ao longo dos anos no campo da proteômica (Yates *et al*., 2009). O grande volume de peptídeos identificados possibilita a confirmação e quantificação de proteínas já conhecidas em diferentes condições. Além disso, é possível identificar peptídeos que caracterizem novas isoformas proteicas provenientes de diferentes eventos moleculares como promotores alternativos, poliadenilação alternativa e *splicing* alternativo (Figura 1.1).

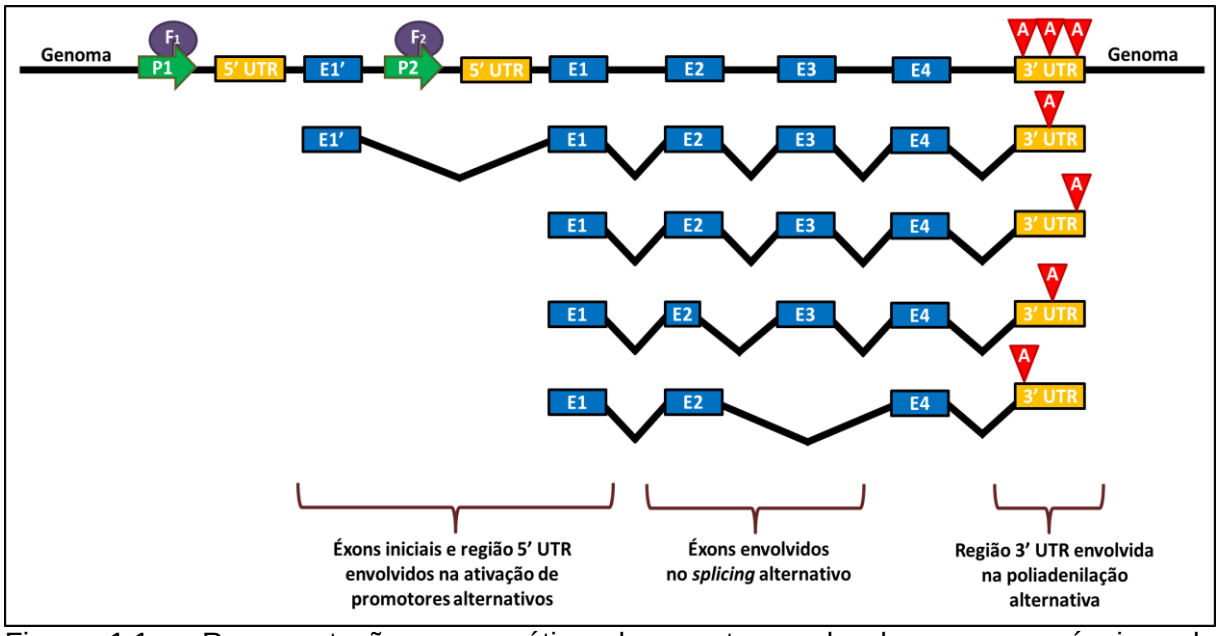

Figura 1.1 - Representação esquemática de eventos moleculares responsáveis pela diversidade proteica (retângulos em azul: éxons codificadores; retângulos em amarelo: regiões não codificadoras; seta verde: regiões promotoras; círculos roxos: fatores de transcrição; triângulos invertidos em vermelho: sítios de poliadenilação).

A proteogenômica é uma área de pesquisa que tem como proposta integrar dados de diferentes tecnologias (como o NGS e MS) e elucidar o fluxo de informação gerado desde a expressão gênica até a tradução dos RNAs mensageiros em proteínas. Desde então, diferentes abordagens têm sido criadas, proporcionando a identificação de potenciais isoformas dentro do contexto celular (revisto por Sheynkman, 2016).

1.2 - O *splicing* e a sua associação com doenças

Um gene humano apresenta basicamente duas regiões que estão dispostas ao longo da fita do DNA e que participam diretamente do evento de *splicing*, sendo estas classificadas como éxons e íntrons. Éxons são sequências que possuem em média 145 nucleotídeos e que constitutivamente não são removidas durante o evento de *splicing*. Em contrapartida, íntrons são sequências aproximadamente 10 vezes maiores em relação aos éxons, intercalados entre eles e que são removidas durante o evento de *splicing* (Lander *et al*., 2001).

Trabalhos realizados no início dos anos 1980 investigaram a composição nucleotídica ao longo da sequência gênica com o objetivo de determinar os limites entre éxons e íntrons, e assim, auxiliar no entendimento do mecanismo de *splicing* (Breathnach e Chambon, 1981; Mount, 1982; Shapiro e Senapathy, 1987). Estes trabalhos apresentaram duas sequências consenso: a primeira localizada na junção éxon-íntron, caracterizado pela sequência [A|C]AG/GT[A|G]AGT, e a segunda, localizada na junção íntron-éxon, caracterizado pela sequência [C|T]N[C|T]AG/G, onde "/" representa o sítio de *splice*. Com o passar dos anos e o aumento da disponibilidade de dados de genomas completos, a presença destes dois perfis nucleotídicos foi encontrada em outros organismos, assim, reafirmando sua importância biológica (Mount *et al*., 1992; Burset *et al*., 2000).

A partir destes estudos, verificou-se que aproximadamente 99% dos íntrons do genoma humano possuem o par de dinucleotídeos GU-AG em seus extremos 5' e 3', respectivamente (Modrek e Lee, 2002). Por fim, é possível observar ainda uma frequência do nucleotídeo guanina na última posição de cada éxon, além de uma região rica em pirimidina próxima ao extremo 3' dos íntrons. Desta maneira, a literatura destaca três regiões fundamentais onde o maquinário enzimático responsável pelo *splicing* as identifica para a remoção dos íntrons: os sítios de *splice* 5' e 3' (localizados nas junções entre éxon e íntrons) e o *branch site*, localizado aproximadamente a 100-150 nucleotídeos do sítio de *splice* 3' (figura 1.2).

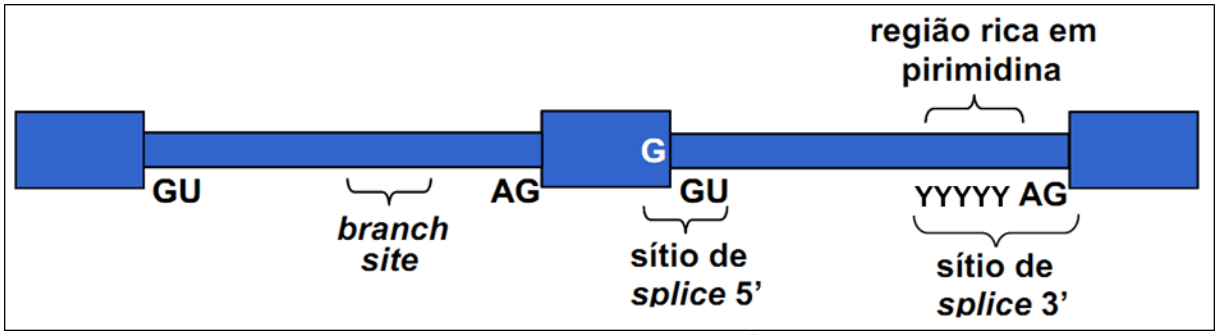

Figura 1.2 - Esquema identificando os principais nucleotídeos atuantes durante o evento de *splicing* (figura adaptada de Faustino e Cooper, 2003).

Como exposto, a região que delimita um íntron foi extensamente investigada onde, por exemplo, foi avaliada a influência de padrões de sequência nessas regiões. Hiller e colaboradores estudaram a distribuição do padrão GYNGYN envolvidas em eventos de *splicing* alternativo em diferentes espécies. Em outro estudo, foi identificado que o padrão de repetição de sequências nos extremos 3' de éxons e íntrons de genes humanos codificadores de proteínas são maiores quando encontrados dentro do mesmo gene (Tavares *et al*., 2012).

Apesar de ser um processo fundamentalmente co-transcricional, o *splicing* pode ocorrer tanto durante como após a transcrição. Quando realizado juntamente com processo de transcrição, o *splicing* permite uma integração funcional da transcrição e o maquinário de processamento de *splicing*, o que e poderia levar a sua mútua modulação. Quando ocorre após a transcrição, o *splicing* poderia facilitar a associação do maquinário enzimático do RNA e sua exportação para assim criar novas etapas na expressão gênica (Han *et al*., 2011).

O *splicing* alternativo é um mecanismo pelo qual múltiplos transcritos são gerados a partir da escolha de diferentes sítios de *splice* de um mesmo mRNA precursor (revisto por Kim *et al*., 2008) (Figura 1.3).

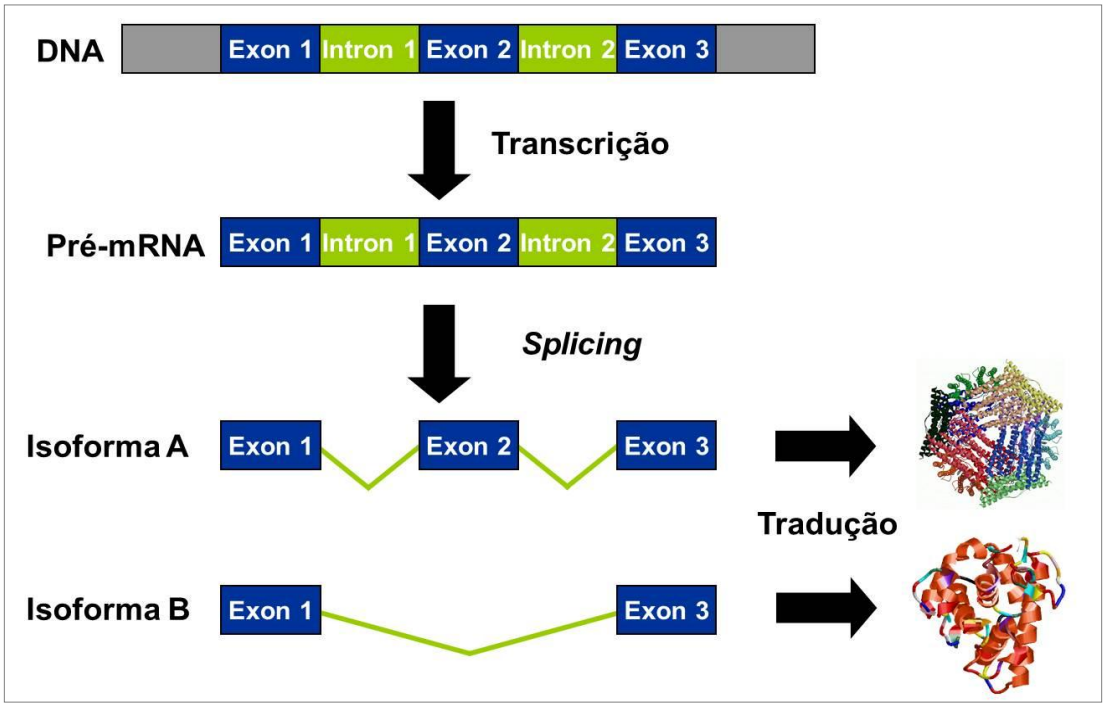

Figura 1.3 - Representação esquemática de um gene hipotético produzindo duas isoformas de mRNA por *splicing* alternativo do éxon 2 (E2).

A primeira sugestão do processamento alternativo de mRNA foi dada por Gilbert (1978) após a descoberta das regiões intragênicas (íntrons) em genes de eucariotos e também em Adenovirus. Gilbert propôs que diferentes combinações de éxons poderiam ser processados conjuntamente para produzir diferentes isoformas de mRNA de um mesmo gene. A proposta de Gilbert levaria futuramente a uma reformulação do dogma central da biologia molecular onde um gene geraria apenas um transcrito podendo este ser traduzido em uma proteína.

Estudos realizados no início da década de 1980 identificaram diversos genes onde foi possível caracterizar eventos de *splicing* alternativo (Early *et al*., 1980). Em 1994, Sharp afirmou que apenas uma pequena fração dos genes em eucariotos superiores poderiam sofrer *splicing* alternativo. Hoje, esta estimativa é de que aproximadamente 90% dos genes humanos tenham transcritos sob a influência de eventos de *splicing* alternativo (Wang *et al*., 2008). Sendo assim, o genoma humano expande sua capacidade de produção de diferentes isoformas de uma mesma proteína, tendo como origem um único gene.

A literatura relacionada a este mecanismo molecular identifica basicamente quatro eventos moleculares, classificados como: retenção de íntrons, uso alternativo de sítios de *splice* 5', uso alternativo de sítios de *splice* 3' e uso alternativo de éxons. É possível também identificar isoformas onde a presença de determinados éxons

está diretamente relacionada a ausência de outros e vice-versa, sendo este evento classificado como éxon mutualmente exclusivos (Figura 1.4).

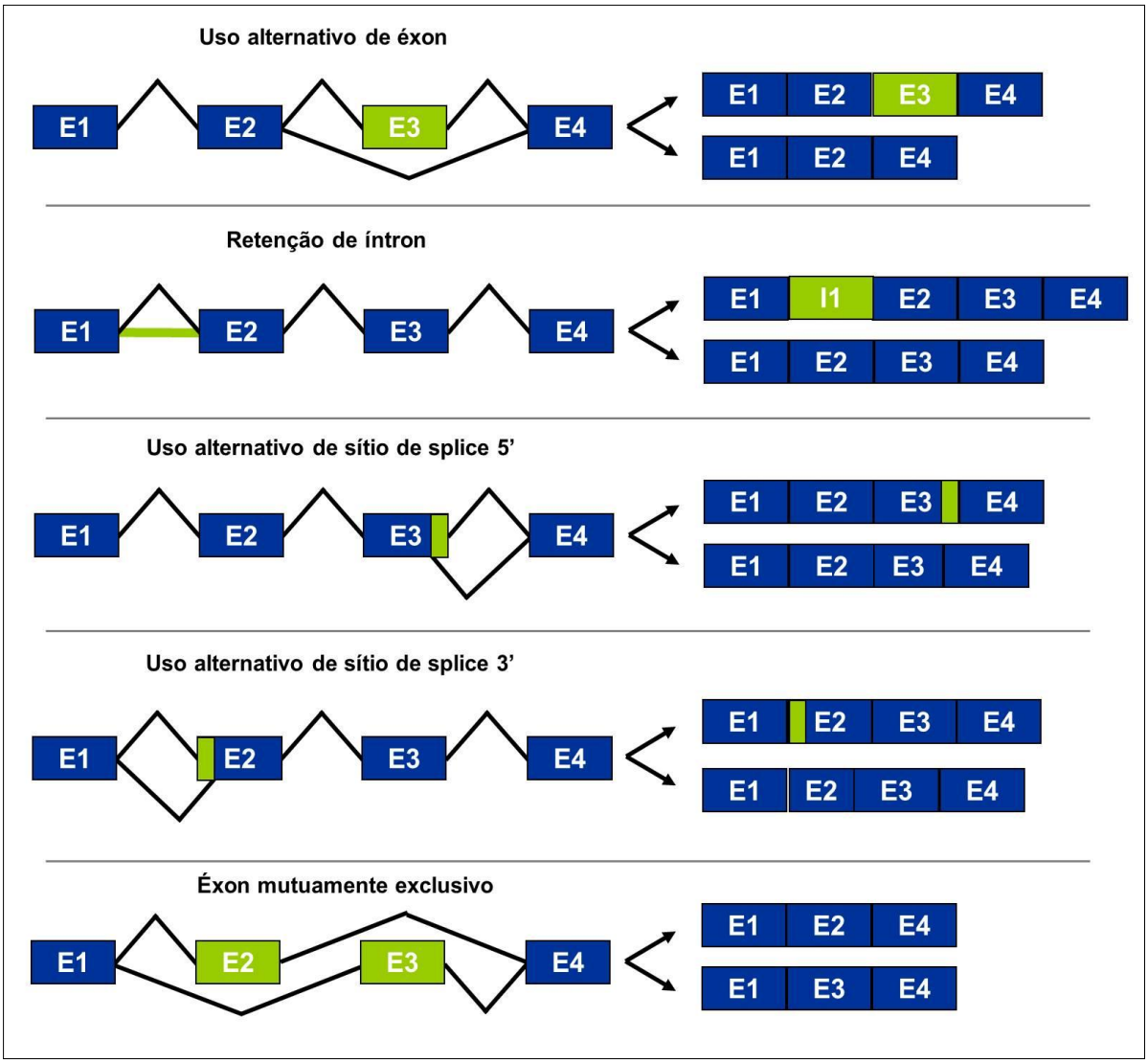

Figura 1.4 - Classificação dos eventos do *splicing* alternativo. Em verde estão representadas as regiões afetadas por cada tipo de evento, sendo E: éxon e I: íntron (Adaptado de Blewcome, 2006).

A ocorrência de um evento de *splicing* alternativo em um mRNA pode depender do estágio celular, ser específico para um determinado tecido, depender das condições biológicas em que a célula se encontra ou estar relacionado ao desenvolvimento de uma patologia, como o câncer (revisto por Orengo e Cooper, 2007; Ward e Cooper. 2010). Como consequência, as proteínas produzidas podem exibir um comportamento diferente e até mesmo antagonista ao seu original devido ao fato de possuírem propriedades estruturais distintas. Logo, o fenótipo resultante dependerá dos níveis de cada isoforma expressa (revisto por Florea *et al*., 2005).

Estudos relacionando o *splicing* alternativo e determinadas patologias tem sido relatados ao longo dos anos na literatura (revisto por Wang e Cooper, 2007). Porém, grande destaque tem sido dado a sua associação com câncer, principalmente devido a relação de determinadas isoformas com o controle do ciclo celular e seu potencial para a aplicação de novos tratamentos.

A família de proteínas Bcl-2 que está envolvida na morte programada das células, a apoptose. Dentro desta família, existem duas isoformas em destaque, uma pró-apoptótica (Bcl-xS) e uma anti-apoptótica (Bcl-xL) que podem ser produzidas a partir de eventos de *splicing* alternativo e que possuem função antagônica (Schwerk e Schulze-Osthoff, 2005). Fatores de *splicing* como o SRSF1 também participam da regulação da apoptose podendo promover o câncer de mama (Anczuków *et al*., 2012).

A terapêutica também se apresenta como uma área promissora para do desenvolvimento de tratamentos envolvendo transcritos oriundos de *splicing* alternativo. O fator de crescimento endotelial vascular A (VEGF-A) apresenta isoformas que podem promover ou inibir o desenvolvimento de uma nova microcirculação. O controle sobre as formas alternativas VEGF-A165 (pró-angiogênica) e VEGF-A165 b (anti-antiangiogênica) ganham destaque no tratamento contra o câncer, já que a neovascularização é fundamental para o crescimento tumoral e seu possível papel na matástase. (Harper e Bates, 2008). Em outro estudo realizado por Elias e Dias (2008), foi demonstrado que a variante VEGF-121 tem sua expressão elevada a partir do momento em que células tumorais da mama e da próstata são submetidas a um pH reduzido devido a condições de hipóxia.

As isoformas também podem se destacar como uma importante estratégia não apenas para o diagnóstico, mas também no monitoramento do câncer. O receptor CD44 está relacionado a interações celulares (Goodison *et al*., 1999) e as variantes expressas pelo seu gene estão relacionadas com vários tipos de tumor (McFarlane *et al*., 2015; Tjhay *et al*., 2015; Hu *et al*., 2016). Outro exemplo é a detecção de algumas variantes do gene *KLK3*, responsável pela produção do antígeno PSA (do Inglês *prostate-specific antigen*). O monitoramento dos níveis séricos de PSA juntamente com a detecção de determinadas isoformas pode acarretar a uma diferenciação entre um câncer de próstata e uma hiperplasia prostática (Brinkman, 2004).

1.3 - Sequenciamento de alta vazão e o seu uso para a análise de transcriptomas

Os sequenciadores de alta vazão (do Inglês *Next Generation Sequencing;* NGS) surgiram há aproximadamente dez anos possibilitando o sequenciamento de genomas e transcriptomas em quantidade sem precedentes (Nene *et al*., 2007; Mortazavi *et al*; 2010; Scally *et al*., 2012) e a criação de grandes projetos como o Projeto 1000 genomas (Sudmant *et al*., 2015), ENCODE (Dunham *et al*., 2012) e TCGA (The Cancer Genome Atlas Network, 2013). Com o passar dos anos, as plataformas de sequenciamento sofreram inovações que acarretaram na redução de seu custo-benefício e consequentemente no acesso desta tecnologia à pequenos laboratórios e instituições de pequeno porte (revisto por van Dijk, 2014).

O processo de sequenciamento varia de acordo com a plataforma utilizada e consequentemente, o volume de dados e o tratamento aplicados a eles também (Metzker, 2010). Entretanto, o produto gerado pelo sequenciamento, conhecido como *read*, passa por tratamentos como remoção de adaptadores, filtro por qualidade e seleção a partir do tamanho da sequência (Williams *et al*., 2016). O tipo de *read* gerado também influencia na estratégia do estudo. Enquanto os *Single-end* auxiliam numa melhor contagem dos transcritos e sua expressão, os *paired-end reads* ajudam a resolver rearranjos cromossômicos como inserções, deleções e inversões (Ekblom e Wolf, 2014).

O sequenciamento de alta vazão do transcriptoma ou RNA-Seq (do Inglês, *RNA sequencing*) também obteve avanços tanto para os RNAs não-codificadores (Zhao *et al*., 2016; Zhang *et al*., 2016) como também para os RNAs mensageiros (mRNA-Seq; Herzel e Neugebauer, 2015; Tembe *et al*., 2014). No que tange a descoberta de potenciais transcritos processados alternativamente, os trabalhos abrangem diferentes áreas como câncer (Hong *et al*., 2016), doença de Alzheimer (Love *et al*., 2015) e sequenciamento de uma única célula (Welch *et al*., 2016). A quantificação dos transcritos expressos e principalmente sua expressão diferencial em diferentes condições também são abordagens estudadas (Aschoff *et al*., 2013).

Como a maioria dos sequenciadores ainda produzem *reads* com tamanhos limitados, há o desafio de recriar os transcritos completos provenientes dos dados de RNA-Seq. Desta forma, a montagem de transcriptoma é uma área que tem se desenvolvido nos últimos anos e contribuído para a identificação de variantes de *splicing*.

A montagem de transcriptoma tem como objetivo recriar os transcritos a partir dos dados de RNA-Seq. Quando o genoma de um organismo é conhecido, este pode ser utilizado como referência, auxiliando na construção dos transcritos. Entretanto, quando o genoma de um determinado organismo não se encontra disponível, é utilizada a montagem *de novo* como estratégia. Demais grupos vêm utilizando as duas abordagens mencionadas como forma garantir um controle melhor quanto à construção dos transcritos (com genoma de refrência), bem como a identificação de novas isoformas de *splicing* alternativo (*de novo*) (Martin e Wang, 2011).

Uma das formas de montar um transcriptoma é usar uma referência, também conhecida como *ab initio*. Esta abordagem tem como foco a construção dos transcritos através de grafos após o mapeamento dos *reads* no genoma. Os programas Cufflinks (Trapnell *et al*., 2010) e Scripture (Guttman *et al*., 2010) são os principais representantes desta abordagem, variando basicamente no modo de construção dos grafos. O primeiro tenta encontrar o menor conjunto de transcritos para inferir as junções dos íntrons no *reads*, ou seja, busca o menor caminho para a construção do grafo. Já o segundo, constrói grafos baseados na sequência genômica e procura pelos *reads* que complementam estes grafos. Apesar da necessidade de um volume de dados menor para realizar esta abordagem, a montagem com genoma de referência torna-se dependente da qualidade do genoma do organismo em estudo (Martin e Wang, 2011).

A montagem de transcriptoma *de novo* caracteriza-se pela construção dos transcritos sem utilizar um genoma de referência. Basicamente, a proposta é retirar as redundâncias dos *reads* e realizar a sobreposição destes para a montagem dos transcritos utilizando grafos de Brujin. Para esta abordagem, uma série de programas têm sido desenvolvidos, sendo eles: Trans-ABySS (Robertson *et al*., 2010), EBARDenovo (Chu *et al*., 2013), KisSplice (Sacomoto *et al*., 2012) , IDBA-Tran (Peng *et al*., 2013). Porém, segundo comparações feitas, os programas Trinity (Grabherr *et al*., 2011) e Oases (Schulz *et al*., 2012) tem se mostrado com desempenho mais satisfatório (Clarke *et al*., 2012; Lu *et al*., 2013). Esta abordagem tem como vantagem a sua utilização para aqueles organismos que não possuem seu genoma sequenciado e amplamente anotado. Entretanto, comparados os volumes de dados entre a montagem com genoma de referência e *de novo*, a

segunda necessita cobertura de 30 vezes para a construção hipotética dos transcritos (Martin e Wang, 2011).

A combinação entre as duas abordagens têm sido utilizada recentemente visando o aprimoramento da montagem de transcritos. Esta estratégia mista pode ser aplicada de duas formas: (a) alinhamento dos *reads* seguido da montagem com o genoma de referência e a montagem *de novo* para os *reads* não mapeados ou (b) montagem *de novo* com os *reads* para a formação dos *contigs*, seguido da utilização dos *reads* não-montados com um genoma de referência para estender estes *contigs* e ampliar a informação transcritômica (Figura 1.5).

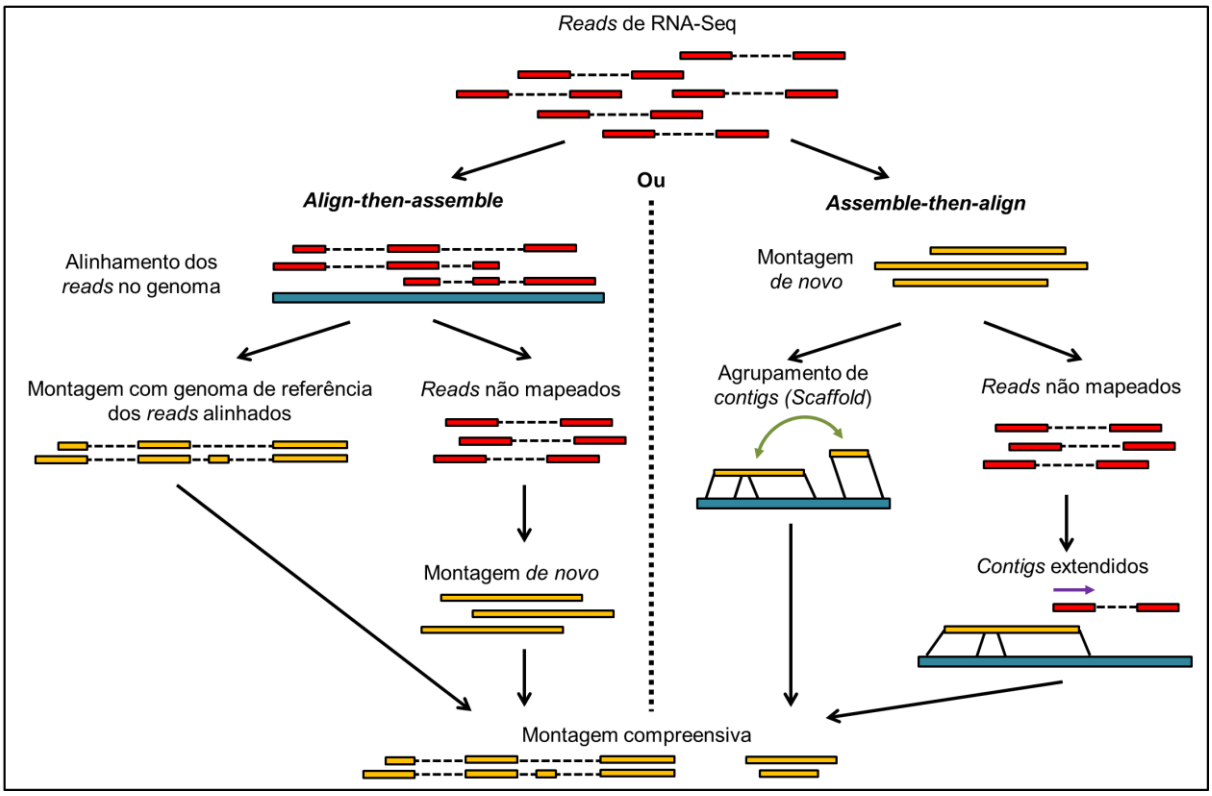

Figura 1.5 - Combinação de estratégias para a montagem de transcriptoma *de novo* e com genoma de referência (Figura adaptada de Martin e Wang, 2011)

1.4 - Proteômica

A Espectrometria de massas (MS; do Inglês, *mass spectrometry*) é uma técnica que tem sua origem no trabalho pioneiro de Thomson (1897), quando a relação de massa e carga do elétron foi determinada. O primeiro espectrômetro de massas foi desenvolvido por Dempster (1918) e seu aprimoramento realizado por Aston (1919), onde a relação de massa para carga (*m/z*) foi calculada de forma mais precisa. O trabalho de Munson e Field (1966) também se destaca pela a descoberta da ionização química, que possibilitou a análise de misturas e consequentemente a caracterização de peptídeos e açúcares (Meurer, 2003). Ao longo dos anos, a espectrometria de massas tem sido aprimorada em termos de equipamento e aplicações (revisto por Zhang e Rockwood, 2015).

A MS possui grande aplicação na proteômica (revisto por Domon e Aebersold, 2006) para a determinação do perfil proteico de um organismo através de uma amostra proveniente de um tecido (Wilhelm *et al*., 2014) ou de uma única célula (Chen *et al*., 2016). Antes de ser inserida no espectrômetro de massas, a amostra é preparada através de técnicas como géis de eletroforese bidimensionais, bandas isoladas de SDS-PAGE e a cromatografia líquida. O preparo como um todo depende da amostra em si e do objetivo do estudo (Gundry *et al*., 2009).

As estratégias para a identificação de proteínas através da MS também podem ser divididas em três: (a) *top-down*, onde a proteínas são analisadas diretamente no espectrômetro de massas sem passar por uma quebra enzimática; (b) *bottom-up*, onde as proteínas são digeridas por uma enzima (geralmente tripsina), e seus peptídeos são inseridos no espectrômetro de massas. Esta estratégia também é conhecida como *shotgun* e é a mais utilizada nos estudos de proteômica; (c) *middle-down*, onde as duas estratégias anteriores são utilizadas. É feita a digestão das proteínas de forma parcial, gerando fragmentos (um pouco maiores que os obtidos por *bottom-up*) que são inseridos no espectrômetro de massas.

Como mencionado, dependendo do preparo e da estratégia utilizada, as proteínas e seus peptídeos podem ser analisados no espectrômetro de massas de formas diferentes. Uma das abordagens consiste na digestão das proteínas por uma enzima, gerando fragmentos peptídicos que terão a sua massa calculada. Esses fragmentos têm o seu espectro comparado contra os espectros teóricos gerados a partir de um repositório de proteínas conhecidas. Quando os espectros teórico e

experimental coincidem, é utilizado o termo “Peptide spetrum match” (PSM), indicando que aquele peptídeo foi registrado e consequentemente sua proteína de origem foi identificada (Thiede *et al*., 2005). Já a abordagem em Tandem, MS/MS ou MS^2 possibilita a análise de misturas contendo várias proteínas. Após a digestão enzimática das proteínas, os peptídeos são fragmentados em aminoácidos através da colisão com um gás inerte pela técnica Collision Induced Dissociation (CID). Com a ionização, esses aminoácidos são analisados e detectados, sendo possível determinar a sequência peptídica através da massa de cada um deles. Esses espectros experimentais serão comparados contra os espectros teóricos de um repositório proteico (na busca por PSMs) para a identificação da proteína. (Yates, 1998).

A quantificação de proteínas através de experimentos de MS tem se desenvolvido ao longo dos anos (revisto por Bantscheff, 2012). Entre as técnicas utilizadas podemos citar a *stable isotope labelling* que é fundamentada na utilização de isótopos estáveis na amostra em estudo. Como esses isótopos possuem uma massa específica, o espectrômetro de massas consegue diferenciar os peptídeos que possuem esses isótopos daqueles que não marcados. Assim, a quantificação é feita a partir da intensidade dos sinais (Gygi *et al*., 1999; Hanke *et al*., 2008; Schmidt *et al*., 2004). Já a técnica *label-free* é feita baseada na quantidade absoluta ou na intensidade do sinal dos peptídeos, sem a utilização de isótopos (Liu *et al*., 2004; Ishihama *et al*., 2005).

A identificação das proteínas em experimentos de MS depende, entre outros aspectos, do repositório proteico utilizado para a comparação entre os espectros teóricos e experimentais. As bases de dados do RefSeq (Pruitt *et al*., 2014), UniProt/SwissProt (The UniProt Consortium, 2015) e ENSEMBL (Yates *et al*., 2016) são tradicionalmente utilizadas por serem frequentemente atualizadas e por conterem sequências com validação experimental. Entretanto, a pequena quantidade de sequências curadas limita o número de proteínas identificadas e dificulta identificação daquelas geradas por eventos como *splicing* alternativo, promotor alternativo e poliadenilação alternativa.

Outro desafio encontrado na intepretação de resultados de MS na proteômica é determinar quais isoformas de um determinado gene ou de uma família gênica estavam realmente presentes na amostra. Isto ocorre porque muitos dos peptídeos encontrados nos experimentos de MS são compartilhados por outras proteínas. Quando um peptídeo pertence somente a uma única proteína ele é classificado

como único ou proteotípico, pois através dele é possível confirmar a presença de uma determinada isoforma (Figura 1.6). Este problema está diretamente relacionado com a escolha do repositório proteico usado para as análises, pois dependendo de sua composição, a busca por potenciais novas isoformas fica limitada à um conjunto de proteínas já conhecido. Assim, a necessidade de repositórios mais especializados ou personalizados se fez presente e será debatida no próximo tópico.

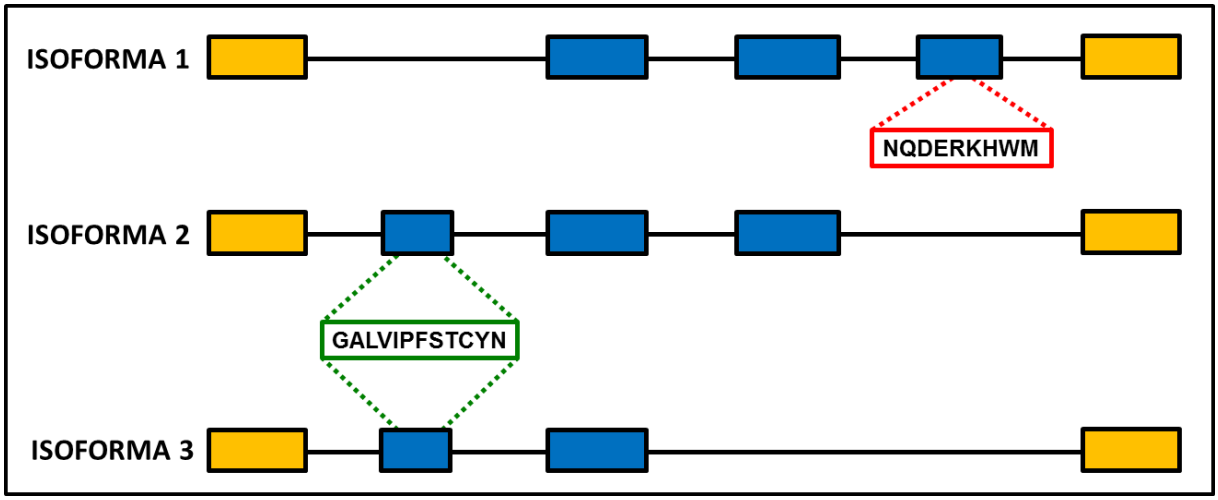

Figura 1.6 - Representação hipotética de peptídeos compartilhados e exclusivos entre isoformas (retângulos azuis: éxons codificadores; retângulos amarelos: regiões não codificadoras; retângulo com borda verde: peptídeo compartilhado entre as isoformas 2 e 3 localizado no éxon 2; retângulo com borda vermelha: peptídeo exclusivo as isoforma 1 localizado no éxon 4).

A validação estatística dos peptídeos e consequentemente das proteínas encontradas também está relacionada com os repositórios encontrados. A taxa de falso positivos (do Inglês, *false discovery rate*; FDR) e a probabilidade de erro posterior (do Inglês, *posterior error probability*; PEP) são os critérios utilizados para validar estatisticamente as proteínas encontradas em experimentos de MS. Em resumo, a FDR procura atribuir uma taxa de erro para um determinado conjunto de proteínas encontradas. Já a PEP atribui uma probabilidade de um determinado peptídeo ter sido identificado de forma errônea. De uma forma geral, a FDR é utilizada em experimentos onde o objetivo é analisar o perfil proteico de uma determinada amostra. Ao passo que a vantagem da PEP, é utilizada para a validação de peptídeos-chave, em especial aqueles que caracterizam uma determinada isoforma (Käll, 2007).

1.5 - A Proteogenômica

Segundo Nesvizhskii (2014), o termo proteogenômica foi primeiramente utilizado por Jaffe e colaboradores (2004), onde em seu trabalho, dados de proteômica foram utilizados para aprimorar a anotação do genoma. Recentemente, ela pode ser definida como é a área de pesquisa que atua na interface entre a genômica e a proteômica (Nesvizhskii, 2014). Isso porque, o desenvolvimento de diferentes tecnologias “omicas”, motivou diversos grupos a utilizarem seus dados, buscando identificar e entender todos os processos e elementos gerados desde a expressão gênica até a codificação das proteínas. Entre as tecnologias mais utilizadas, estão o RNA-Seq e a MS devido ao grande volume de dados gerados. Outra característica presente nos estudos em proteogenômica é a utilização de repositórios proteicos personalizados de acordo com a amostra a ser analisada.

Um dos primeiros estudos voltados esta estratégia foi realizado por Ning e Nesvizhskii (2010), onde se buscou identificar novas formas de *splicing* alternativo em três tecidos diferentes de camundongo (músculo esquelético, fígado e tronco cerebral). O repositório proteico utilizado continha sequências proteicas canônicas e sequências obtidas de dados de RNA-Seq. Para cria-lo, o transcriptoma sequenciado foi traduzido nas seis fases de leitura onde apenas sequências com mais de 30 aminoácidos foram selecionadas. De acordo com os autores, para cada tecido, de dois a três peptídeos indicavam novas junções de *splice* que também eram confirmadas por ESTs. Apesar do pequeno número de novos peptídeos identificados, este estudo foi uma das primeiras iniciativas que instigaram outros trabalhos a unir dados de RNA-Seq e MS.

Sheynkman e colaboradores (2013) também utilizaram um repositório proteico personalizado para detectar variantes de *splicing*. Este foi construído a partir de sequências canônicas oriundas do Uniprot-Trembl e sequências de peptídeos provenientes de dados de RNA-Seq. Esses peptídeos que foram traduzidos *in silico*, correspondem às junções de *splice* identificadas no transcriptoma de uma linhagem de células T (Jurkat *Cells*). Os dados proteômicos para MS também foram obtidos a partir da origem, o que poderia favorecer uma correlação entres os dados de RNA-Seq e MS. Ao final, foram identificados 57 peptídeos correspondentes às junções de *splice* e que não estavam presentes nas sequências canônicas.

Em trabalho desenvolvido pelo nosso grupo, outro repositório personalizado foi criado a partir de sequências canônicas e peptídeos derivados de potenciais

variantes de *splicing*. O repositório então foi utilizado contra os mesmos dados de MS citados no parágrafo anterior com a proposta de identificar potenciais variantes de *splicing*. Foram identificados 54 peptídeos (Tavares *et al*., 2014), dos quais 10 eram idênticos aos encontrados por Sheynkman e colaboradores (2013). Este estudo faz parte da presente tese e seu desenvolvimento será melhor descrito nas próximas seções.

Demais estudos também utilizaram as abordagens com RNA-Seq/MS, porém as análises de MS foram feitas com repositórios tradicionalmente utilizados (Lundberg *et al*., 2010; Nagaraj *et al*., 2011; Higareda-Almaraz, *et al*., 2013). Apesar de trazerem segurança para os resultados encontrados, esses repositórios impossibilitam a descoberta de novas isoformas. Por outro lado, repositórios personalizados criam a possibilidade de encontrar novas variantes de *splicing* e demais proteínas derivadas outros processos celulares, que poderiam ser alvo para o desenvolvimento de fármacos (revisto por Tavares *et al*., 2015). A combinação entre a origem dos dados de RNA-Seq/MS, e a maneira como estes repositórios são construídos também podem auxiliar nessas descobertas (Figura 1.7).

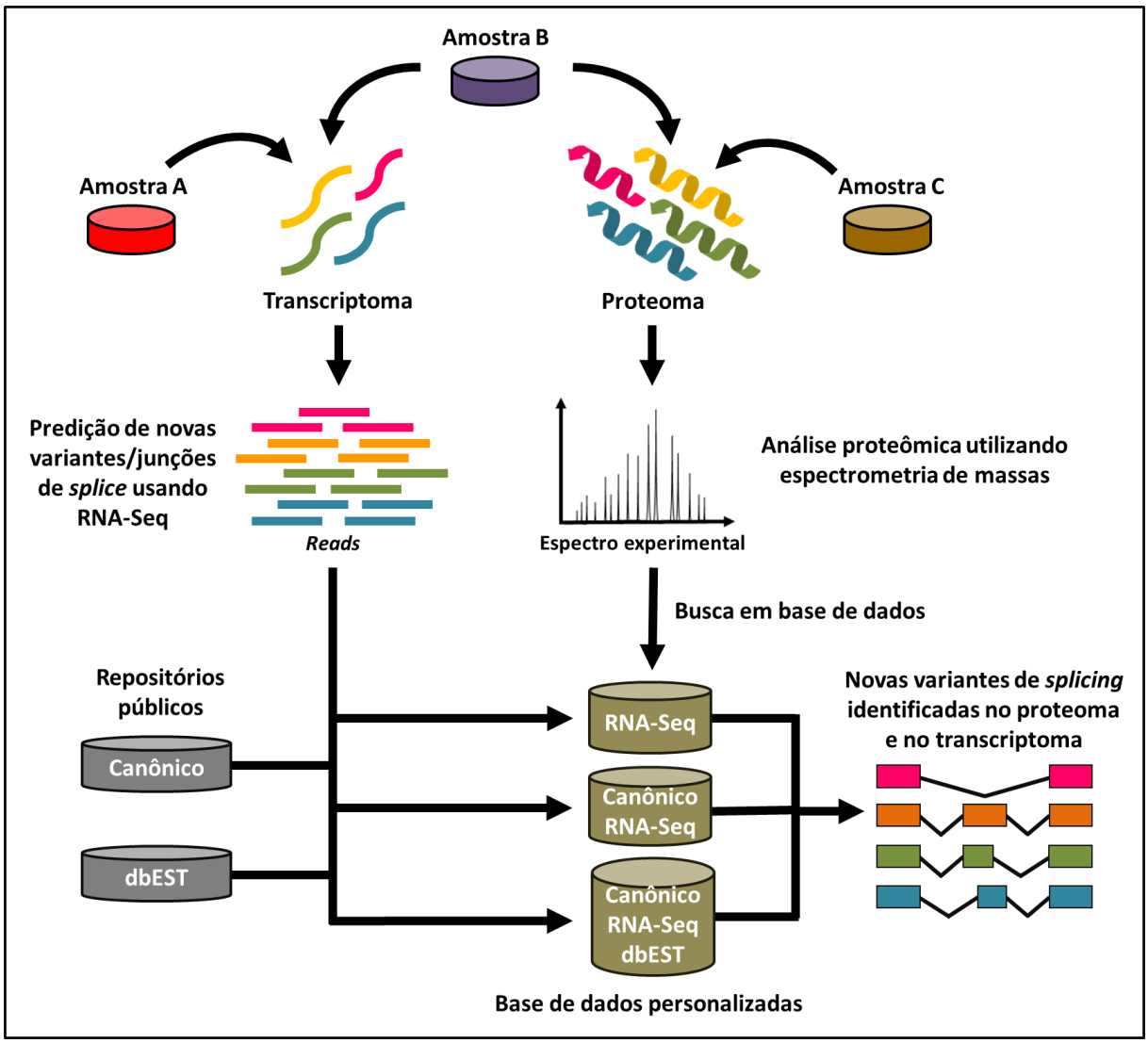

Figura 1.7 - Esquema representativo entre as origens das amostras e as bases de dados para a análise em proteogenômica (figura adaptada de Tavares *et al*., 2015). O transcriptoma e o proteoma podem ter origens diferentes, gerando dados a partir de RNA-Seq e de MS, respectivamente. Os dados de RNA-Seq podem ser associados à repositórios públicos gerando bases de dados personalizadas que serão utilizadas pela espectrometria de massas na busca por novas variantes de *splicing*.

Krug e colaboradores (2014) utilizaram dados de RNA-Seq na busca por polimorfismos de nucleotídeo único (SNP, do Inglês Single Nucleotide Polymorphism) não sinônimo. Os SNPs foram detectados através do programa SAM tools (Li *et al*., 2009) e o programa Cufflinks (Trapnell *et al*., 2010) foi utilizado para a montagem dos transcritos. Foram construídos três repositórios personalizados tendo como base sequências proteicas do Uniprot: o primeiro consistia na tradução dos transcritos nas seis fases de leitura. Para o segundo repositório, foi utilizado um software que inferia as regiões codificadoras dos transcritos. O terceiro e último repositório foi composto por transcritos que tiveram seus SNPs preditos computacionalmente por um software. Apesar do número de SNPs identificados pelos três repositórios terem sido aproximadamente o mesmo, mais da metade deles

foi identificado como falso positivo quando o primeiro repositório foi utilizado. Este resultado demonstra a importância de se buscar estratégias para a correta tradução dos transcritos a fim de evitar resultados derivados das metodologias utilizadas. Sheynkman e colaboradores também publicaram um artigo voltado para a busca por SNPs não sinônimos em uma linhagem de linfócitos T humanos. A partir de dados de RNA-Seq foram construídos repositórios personalizados, onde 70% dos SNPs identificados não estavam presentes na base de dados dbSNP (Sherry *et al*., 2001). Essa abordagem demonstra como novos dados de RNA-Seq são importantes para a identificação de novos SNPs.

A proteogenômica também avança na área da oncologia, como no estudo de Keerthikumar e colaboradores (2015), que analisaram o transcriptoma e o proteoma das vesículas extracelulares de linhagens celulares humanas de neuroblastoma. Foram criados dois repositórios para a elucidação do repertório proteico de tais vesículas: o primeiro consistia na utilização de um repositório tradicional (RefSeq) e o segundo foi formado pela tradução *in silico* de mRNAs de ectosomas e exomas. Este repositório possuía informações sobre INDELs e SNPs, e seria analisado por espectrometria de massas para a identificação de proteínas mutantes. Além de serem identificadas, tais proteínas possuíam o potencial para induzirem significativamente a proliferação e migração das células através do repositório customizado. Mertins e colaboradores (2016) analisaram 105 tumores de mama armazenados no TCGA e também criaram repositórios personalizados baseados em cada um deles. Os repositórios eram compostos também pelo repositório tradicional RefSeq, juntamente com dados de SNPs processados pelo programa QUILTS (Ruggles *et al*., 2015). Apesar da grande variabilidade de informação transcriptômica gerada, o número de transcritos com essas variações confirmados pela espectrometria de massas foi baixo.

Um dos projetos que caracterizam bem o crescimento da proteogenômica é o *Chromosome-based Human Proteome Project* (C-HUPO; Paik *et al*., 2012). Lançado em 2012, o projeto envolve diferentes laboratórios ao redor do mundo responsáveis por caracterizar o proteoma humano de acordo com cada cromossomo analisado. O Brasil participa deste projeto, tendo como foco o cromossomo 15, através de trabalhos que analisaram, por exemplo, o tecido cerebral (Martins-de-Souza *et al*., 2013). O C-HUPO também aborda trabalhos voltados o câncer, como publicado por Menon e colaboradores (2014). Utilizando RNA-seq e MS, o grupo focou na identificação de variantes de *splicing* do cromossomo 17, que possui genes

associados ao câncer como *BRCA1*, *BRCA2* e *TP53*. Utilizando três linhagens celulares de câncer de mama, 4.406 transcritos foram identificados utilizando RNA-seq, sendo 23,88% deles considerados variantes de *splicing*. Os resultados de MS indicaram novos peptídeos que estavam localizados em regiões não codificadoras de alguns genes, indicando um mecanismo de *splicing* desregulado ou uma incorreta anotação das regiões codificadoras e não-codificadoras deste gene. Fanayan e colaboradores também utilizaram a abordagem por RNA-Seq/MS ao analisar três linhagens celulares de câncer colorretal e correlacionar os níveis de transcritos expressos e proteínas codificadas em cada linhagem. Após a análise e comparados os níveis de expressão com outros trabalhos com resultados semelhantes, os autores sugeriram potenciais marcadores para cada câncer. Demais trabalhos desenvolveram *pipelines* automatizados e bases de dados inspirados no projeto C-HPP. Krasnov e colaboradores (2015) desenvolveram o *pipeline* automatizado “PPLine” que consegue detectar SNPs e variantes de *splicing*. Para demonstrar seu funcionamento, dados de RNA-Seq da linhagem de hepatócitos humanos HepG2 e de amostras teciduais de fígado foram utilizados juntamente com dados de espectrometria de massas. Foram identificados 659 transcritos contendo SNPs que não estavam anotados na base dbSNP além de 17 INDELS associados a modificações na fase de leitura. Jeong e colaboradores (2015) desenvolveram uma base de dados que reúne as proteínas já conhecidas pelo projeto C-HPP e explorá-las a partir de cada cromossomo.

Tendo em vista a associação de diferentes abordagens envolvendo tecnologias de larga-escala, a presente tese representa uma das iniciativas na proteogenômica e que buscou contribuir com o estado da arte nesta área.

2. OBJETIVOS

2.1 - Objetivo Geral

Desenvolver uma abordagem de proteogenômica para a análise de eventos de *splicing* alternativo em dados de transcriptoma e proteoma oriundos de homem e camundongo.

2.2 - Objetivos Específicos

a) Mapear dados de transcriptoma (dbEST e dados de RNA-Seq) de cada uma das duas espécies nas suas respectivas sequencias de genoma de referência.

b) Detectar eventos de *splicing* alternativo no homem e camundongo.

c) Criar um repositório contendo a tradução hipotética de variantes de *splicing* para cada uma das espécies.

d) Buscar variantes de *splicing* não descritas em outros repositórios de sequências de proteínas e presentes nas duas espécies que são foco de análise neste projeto

3. MATERIAL E MÉTODOS

A metodologia do presente trabalho está dividida em quatro partes: (a) informações sobre os dados utilizados; (b) construção das bases de dados, identificação e tradução das variantes de *splicing*; (c) desenvolvimento do *pipeline* para a montagem de transcriptoma; e (d) construção de repositórios proteicos e sua aplicação na anotação de experimentos de espectrometria massas.

3.1 - Dados utilizados

Todos os resultados que serão apresentados são referentes aos organismos *Homo sapiens* e *Mus musculus*. Para isso, genoma humano (versão hg-19) e de camundongo (versão mm9) foram obtidos através do sítio de FTP da Universidade da Califórnia, campus Santa Cruz (UCSC) (Karolchik *et al*., 2003).

Dados referentes aos transcritos e proteínas foram obtidos através do projeto RefSeq (Pruitt *et al*., 2009) onde foi possível obter dados de sequências de referência (prefixos NM e NP). Com o intuito de detectar mais variantes de *splicing*, foram utilizadas ESTs disponíveis no repositório dbEST (Boguski *et al*., 1993).

Duas fontes de sequenciamento de larga escala foram utilizadas. A primeira é proveniente da plataforma de sequenciamento 454/Roche obtida através do repositório SRA (*Sequence Read Archive*), hospedado pelo site do NCBI (*National Center for Biotechnology Information*). Esses dados estão relacionados a estudos de câncer de mama e melanoma, cujos identificadores do SRA do NCBI são: SRP000614, ERP000265, SRP003173 e SRP003483.

A segunda fonte foi obtida através do estudo de Lin e colaboradores (2014), onde se buscou ampliar a identificação de novas variantes de *splicing* através da montagem de transcriptoma. Nesse trabalho, as bibliotecas *paired-end* de RNAs mensageiros foram sequenciadas de treze tecidos diferentes tanto para homem quanto para camundongo. As amostras humanas de tecido adiposo, glândula adrenal, fígado, rim, pâncreas, pulmão, coração, intestino delgado, intestino grosso e baço foram obtidas através de doações enquanto as de cérebro, ovário e testículo foram obtidos comercialmente. As amostras dos tecidos de camundongo foram obtidas a partir da linhagem C57BL com idade entre seis e dez semanas de vida. Para cada tecido, foram obtidas duas réplicas técnicas com exceção do tecido ovariano, onde apenas uma réplica estava disponível para *download* para cada organismo, totalizando cinquenta amostras. Por fim, todas as corridas passaram por

um filtro de qualidade de 30 (*Phred score*) e remoção do adaptador através do software TrimGalore (www.bioinformatics.babraham.ac.uk/projects/trim_galore/; versão 0.3.7).

3.2 - Construção das bases de dados, identificação e tradução das variantes de *splicing*

3.2.1 - Construção das bases de dados

Utilizando o programa BLAT (Kent, 2002), os dados de RefSeq e EST foram alinhados contra os genomas dos respectivos organismos para a determinação das coordenadas gênicas. Em seguida, cada transcrito foi agrupado através do projeto Unigene (Wheeler *et al*., 2003) para *Homo sapiens* (versão 235) e *Mus musculus* (versão 194). Transcritos contendo apenas um éxon foram excluídos das bases de dados. Todas as informações processadas foram armazenadas em uma base de dados utilizando o sistema gerenciador de banco de dados (SGBD) PostgreSQL (versão 8.4.7).

3.2.2 - Identificação das variantes de *splicing*

Os transcritos de cada organismo em estudo foram analisados quanto aos seus mapeamentos e agrupados em variantes de *splicing* utilizando a metodologia de matrizes ternárias (pedido de patente requisitada no INPI sob o número PI 0703888-7 e concedida na Suíça sob número 699132). As matrizes ternárias são matrizes de tamanho NxM, onde N é o número total de quaisquer transcritos ou proteínas mapeados em uma região específica de um dado genoma, e M é o número total de éxons e íntrons encontrados na mesma região. A identificação dos eventos de *splicing* alternativo e demais características desta metodologia estão descritas de forma mais detalhada no artigo publicado durante este doutorado (Tavares *et al*., 2014).

3.2.3 - Tradução das variantes de *splicing*

Durante o mestrado deste estudante, foi realizado o teste de conceito que avaliou a tradução *in silico* dos programas Transeq (Rice *et al*., 2000), AUGUSTUS (Stanke *et al*., 2006) e PASA (Haas *et al*., 2003). Este teste consistia na tradução *in silico* de RNAs mensageiros de referência e seu posterior alinhamento contra a sua sequência proteica de referência (RefSeq) correspondente. Após verificar que o programa Transeq gerou os melhores resultados, todas as variantes de *splicing* da base de dados de homem (com base nos dados de ESTs, RefSeq e *reads* SRA) foram traduzidas, gerando o repositório proteico denominado *Splice*Prot (Tavares *et al*., 2014).

Da mesma forma, as variantes geradas pelas montagens de transcriptoma *de novo* foram submetidas à identificação de variantes de *splicing* e sua subsequente tradução *in silico* a partir do nossa metodologia previamente testada e publicada.

3.3 - Desenvolvimento do *pipeline* para montagem de transcriptoma

A montagem de transcriptoma dos dados de RNA-Seq de homem e camundongo foi realizada com o intuito de reconstruirmos as sequências completas dos transcritos, inseri-las em nosso sistema para a identificação das variantes de *splicing* e realizar sua posterior tradução. Resumidamente, o *pipeline* seleciona os *reads* não utilizados pela montagem com genoma de referência para que sejam conduzidos à montagem *de novo*. Foram utilizados os softwares de montagem com genoma de referência Cufflinks (Trapnell *et al*., 2010; versão 2.2.1) e montagem *de novo* Trinity (Grabherr *et al*., 2011; versão r20140717). Complementando o *pipeline*, o programa Novoalign (http://www.novocraft.com; versão 3.02.10) também foi utilizado para o alinhamento dos reads e o programa Samtools (Li *et al*., 2009; versão 0.1.19) foi aplicado nas etapas de seleção dos reads mapeados. Por fim, foram criados *scripts* em linguagem de programação Perl para a devida seleção dos *reads* de acordo com cada etapa da montagem. A Figura 3.1 mostra resumidamente as etapas deste processo.

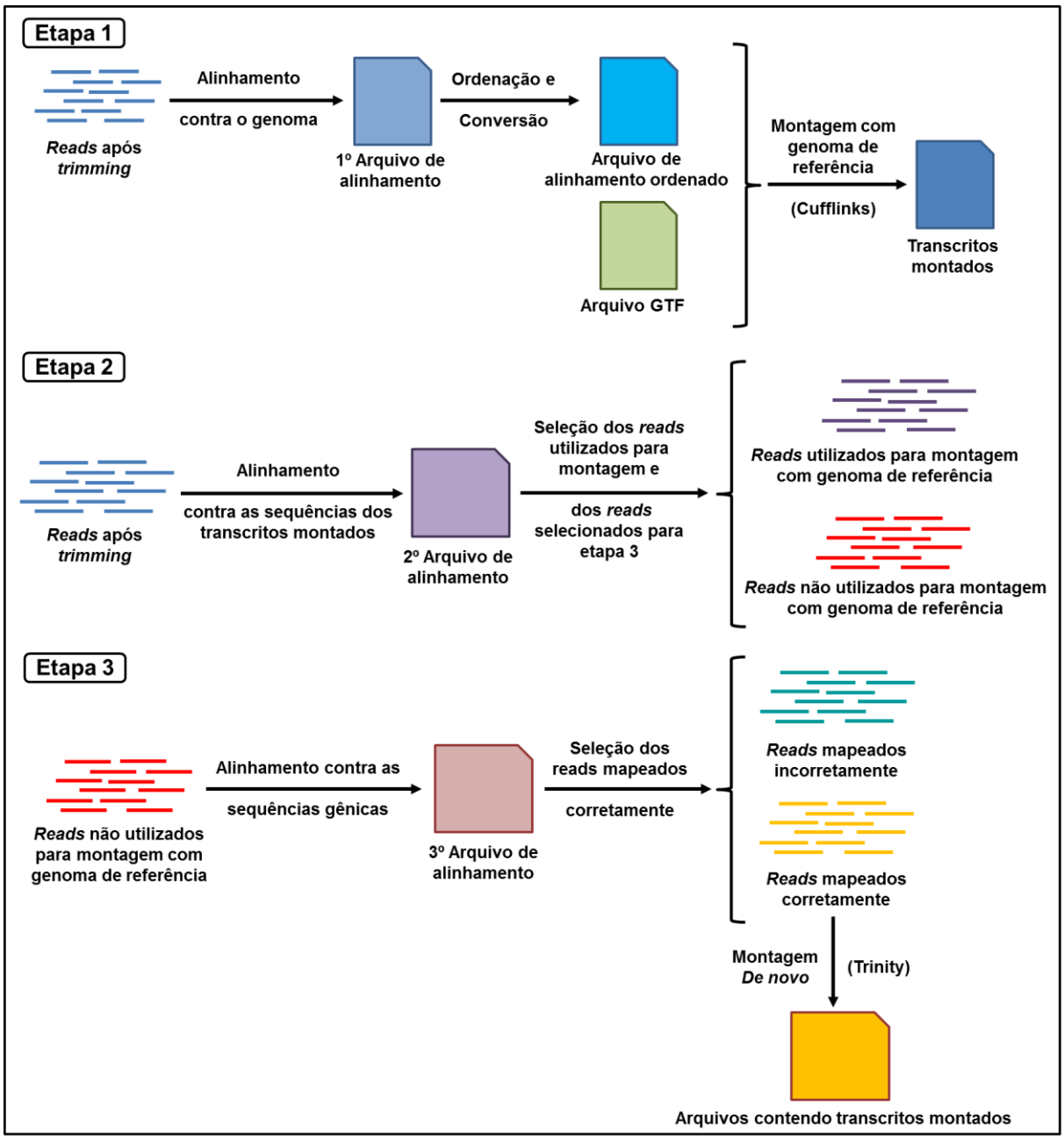

Figura 3.1 - Etapas executadas no *pipeline* para a montagem com genoma de referência e *de novo.* A etapa 1 realiza o primeiro alinhamento com os *reads* após o *trimming* e posterior montagem com genoma de referência utilizando o programa Cufflinks. A etapa 2 realiza o alinhamento dos *reads* utilizados na primeira etapa contra as sequências dos transcritos reconstruídos pelo Cufflinks. Em seguida, um *script* em Perl seleciona os *reads* que foram mapeados corretamente nos transcritos com o objetivo de identificar quais foram utilizados pela montagem com genoma de referência. A etapa 3 realiza o alinhamento dos *reads* não utilizados pelo Cufflinks contra as sequências de genes humanos. Em seguida, a partir de um *script* em Perl, aqueles *reads* mapeados corretamente em apenas um único gene foram direcionados para montagem *de novo* com o programa Trinity.

3.3.1 - Etapa 1: Alinhamento contra o genoma (1º Alinhamento) e montagem com genoma de referência

Após o devido tratamento dos *reads* mencionado ao final no item 3.1, cada uma das 50 corridas passou por um primeiro alinhamento contra o genoma de sua respectiva espécie. Em seguida, cada arquivo de alinhamento foi preparado para a montagem com genoma de referência com o programa Cufflinks. O preparo consistia na conversão de cada arquivo de alinhamento no formado SAM para o formato BAM e sua subsequente ordenação.

Além dos arquivos de alinhamento devidamente preparados, a montagem com genoma de referência necessita um arquivo no formato GTF (do Inglês, Gene Transfer Format) que contém as coordenadas dos éxons de transcritos já anotados. Desta forma, através do sítio da Universidade da Califórnia, campus Santa Cruz (UCSC), foram obtidos para cada organismo em estudo, dois arquivos (um para cada organismo) no formato mencionado contendo as informações de RNAs mensageiros de referência do projeto RefSeq.

Fornecidos os arquivos necessários, a montagem com genoma de referência foi executada com o objetivo reconstruir os transcritos de cada tecido de cada espécie e subsequentemente com a montagem *de novo*. Entretanto, o Cufflinks não informa quais *reads* foram utilizados para a montagem e quais foram descartados. Os arquivos de saída gerados indicam os transcritos foram montados, seus níveis de expressão e de seus genes. Sendo assim, a segunda etapa deste *pipeline* destina-se a identificação de tais *reads* para dar prosseguimento à montagem *de novo*.

3.3.2 - Etapa 2: Alinhamento contra as sequências dos transcritos montados (2º Alinhamento) e 1ª seleção de reads

Para identificarmos quais *reads* foram utilizados, foi estabelecido um contato com desenvolvedor do Cufflinks, Dr. Cole Trapnell, que sugeriu um segundo alinhamento dos *reads* originais contra a sequência dos transcritos identificados pelo seu programa. Desta forma, para cada arquivo que indicava os transcritos que foram montados, o programa Gffread (disponível no pacote de programas do Cufflinks) foi utilizado para a reconstrução de cada transcrito baseado na sequência genômica de cada organismo. Com as sequências reconstruídas, cada uma foi utilizada como referência para um segundo alinhamento usando os mesmos *reads* da primeira etapa.

Ao serem alinhados, os *reads* podem ser mapeados em regiões que não correspondem a sua verdadeira origem. Desta forma, com o auxílio do programa Samtools, foi possível filtrar os *reads* mapeados corretamente a partir do campo "Flag" no segundo arquivo de alinhamento. Para que os *reads* fossem considerados corretamente mapeados (orientação correta), estes deveriam apresentar um dos dois pares de "Flags" no arquivo de alinhamento: 99 e 147 ou 83 e 163. Identificados os *reads* mapeados corretamente, foi possível então selecionar aqueles não utilizados para a montagem com genoma de referência a partir de um *script* escrito na linguagem de programação Perl desenvolvido em nosso laboratório. Esses *reads* passaram por um último processamento na terceira etapa deste *pipeline* com o intuito de refinar a montagem *de novo* com o programa Trinity.

3.3.3 - Etapa 3: Alinhamento contra as sequências gênicas (3º Alinhamento), 2ª seleção de *reads* e montagem *de novo*

Os *reads* que não foram utilizados para a montagem com genoma de referência passaram por um terceiro alinhamento contra as sequências gênicas de cada organismo antes dar início à montagem *de novo*. Este terceiro alinhamento foi necessário a fim de aprimorar a montagem com o Trinity identificando quais *reads* foram alinhados corretamente em apenas um único gene. Desta forma, foi possível evitar que *reads* alinhados em diferentes regiões gênicas pudessem ser utilizados na reconstrução de um mesmo transcrito.

Com o auxílio do programa Samtools e utilizando os mesmos critérios para o campo *Flag* da etapa 2 deste *pipeline*, foi possível filtrar quais *reads* foram mapeados corretamente em apenas um único gene no terceiro arquivo de alinhamento. Em seguida, a partir de um segundo *script* em Perl (também desenvolvido em nosso laboratório), foram selecionados estes *reads* para que fossem agrupados em arquivos Fastq, de acordo com seu gene correspondente. Prontamente, a montagem *de novo* foi aplicada em cada gene utilizando esses arquivos Fastq contendo seus respectivos *reads*. Por fim, os transcritos montados foram incorporados às bases de dados de homem e camundongo (item 3.2) para que enfim fossem analisados na busca por novas variantes de *splicing*.

3.4 - Construção de repositórios proteicos e sua aplicação em experimentos de espectrometria massas

Foram construídos quarto repositórios proteicos (três para homem e um para camundongo) contendo as sequências polipeptídicas oriundas da tradução das variantes de *splicing* detectadas nos dados de transcriptoma. Cada um deles foi adaptado e aplicado em diferentes experimentos de espectrometria de massas a fim de se identificar potenciais novas variantes de *splicing* ou traçar o perfil do conjunto de transcritos e proteínas encontrados.

Em todos os experimentos de MS envolvendo os repositórios criados, foram aplicados os mesmos parâmetros para a identificação e validação das isoformas. Para a busca dos espectros, os parâmetros do programa Sequest disponível no sistema Proteome Discoverer (versão 1.3) foram configurados de forma a se aproximar da mesma configuração realizada no trabalho de Sheynkman e colaboradores (2013). Todas as análises descritas nesse texto que usaram o sistema Proteome Discoverer foram realizadas em colaboração com o a Dra. Adriana Franco Paes Leme e a Dra. Bianca Pauletti do Laboratório Nacional de Biociências (LNBio/Campinas-SP). Da mesma maneira, os critérios estatísticos para validar a identificação das proteínas também foram baseados no trabalho previamente mencionado, onde foram utilizados: “False Discovery Rate” (FDR) a 1% e 5% e “Posterior Error Probability” (PEP) a 1% e 5%.

O primeiro repositório foi submetido à dados de MS da linhagem de linfócitos T TIB-152 (Jurkat cells) utilizados no estudo de Sheynkman e colaboradores (2013). O repositório em questão foi construído selecionando sequências proteicas não redundantes do repositório UniProtKB/SwissProt em conjunto com peptídeos não redundantes oriundos da digestão *in silico* das sequências do *SpliceProt*. Apenas as sequências idênticas foram consideradas redundantes, sendo reduzidas a uma única sequência utilizando uma variável do tipo *hash* a partir de um *script* em Perl. A digestão computacional foi efetuada pelo programa Digest (pacote EMBOSS versão 6.3.1) que foi previamente modificado para que a digestão por tripsina clivasse somente os peptídeos lisina (L) e arginina (R), com exceção quando sucedidos por uma prolina (P). Os peptídeos gerados após esta digestão foram selecionados de forma que não pudessem estar presentes em qualquer sequência proteica do repositório UniProtKB/SwissProt. Em seguida, os peptídeos selecionados que possuíam sequências idênticas também foram reduzidos a uma única sequência,

removendo assim a sua redundância. Ao final, esse peptídeos foram adicionados ao repositório proteico-base composto pelo UniProtKB/SwissProt (Figura 3.2).

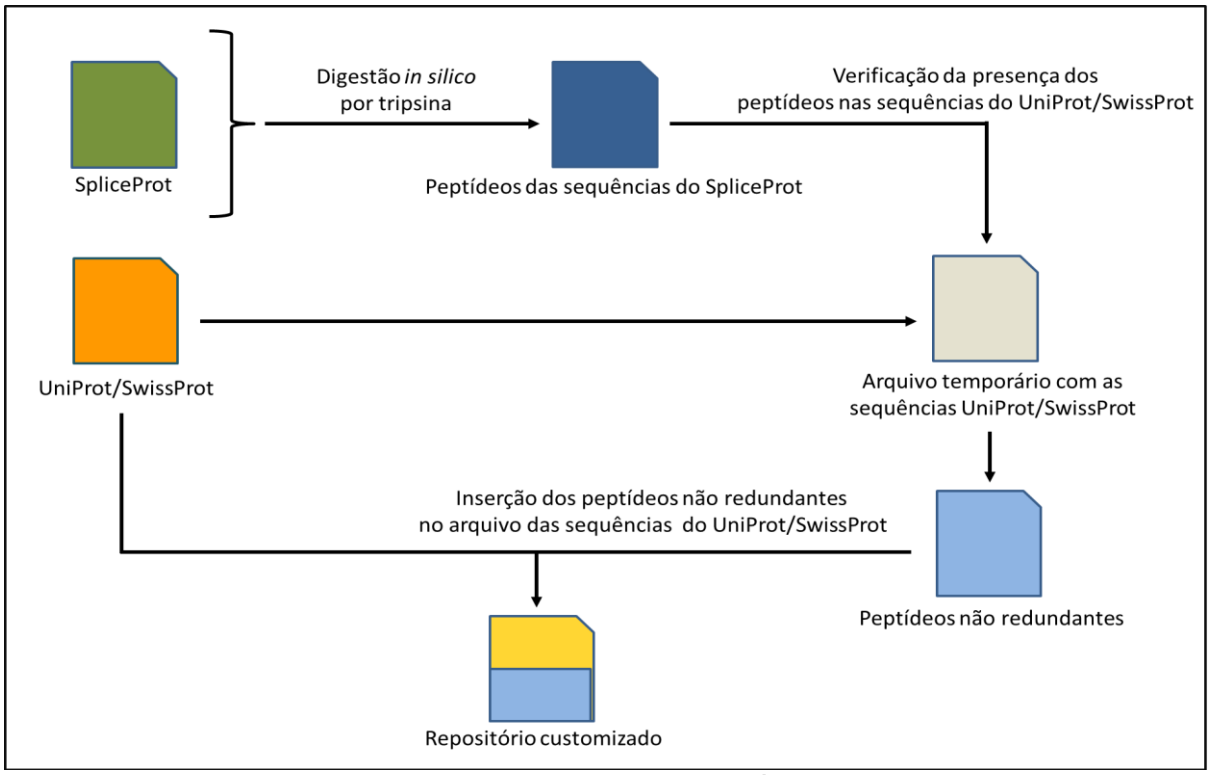

Figura 3.2 - Esquema para construção do primeiro repositório proteico personalizados.

O segundo repositório foi submetido a dados de MS da linhagem de oligondendrócitos humanos MO3.13, publicados por Iwata e colaboradores (2013). Para sua construção, as sequências canônicas do repositório Uniprot/SwissProt tiveram suas redundâncias removidas (da mesma maneira como efetuado para o primeiro repositório) e em seguida, foram copiadas para um arquivo temporário. As sequências identificadas como isoformas foram acrescidas a um arquivo a parte junto com as sequências do *Splice*Prot e, após a remoção das sequências redundantes, foi efetuada a digestão computacional pelo programa Digest (também modificado). Assim como feito para o primeiro repositório, os peptídeos gerados após a digestão foram selecionados de forma que não pudessem estar presentes em qualquer sequência canônica. Por fim, esses peptídeos selecionados tiveram também a sua redundância removida e, em seguida, foram acrescentados ao repositório proteico-base que continha as sequências canônicas, finalizando o segundo repositório (Figura 3.3). O estudo da busca por variantes de *splicing* alternativo no proteoma de MO3.13 foi feito em colaboração com o Dr. Daniel Martins-de-Souza (UNICAMP), Dra Juliana S. Cassoli (UNICAMP) e a Dra Patricia Savio de Araujo Souza (UFF e INCA).

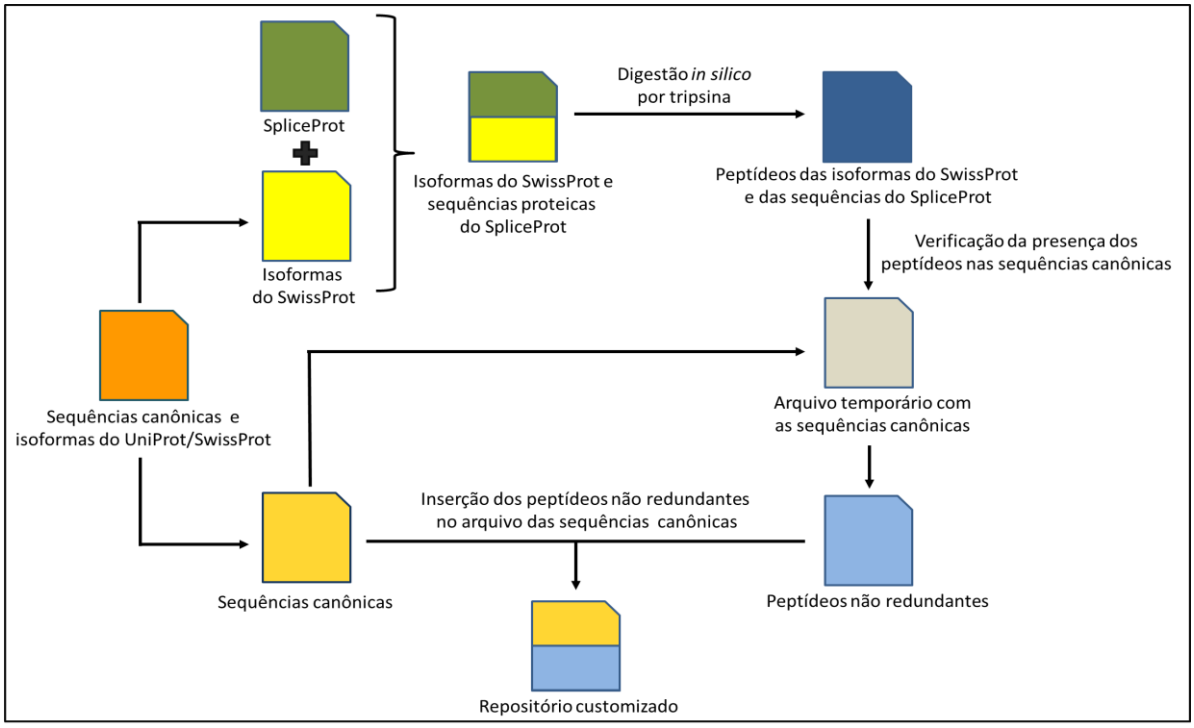

Figura 3.3 - Esquema para construção do segundo repositório proteico personalizado.

O terceiro repositório humano gerado foi submetido à dados de MS de corpo caloso (CC) e lobo temporal anterior (LTA) de humano publicados por Martins-de-Souza e colaboradores (2013). Para a realização desta análise, foi criado um repositório proteico personalizado a partir dos dados de RNA-Seq de cérebro humano que foram aplicados ao *pipeline* descrito no item 3.3. Para a construção deste repositório, as sequências proteicas de referência (RefSeq) correspondentes aos transcritos montados pelo Cufflinks foram obtidas em nossa base de dados. Em seguida, foi necessário identificar quais destes transcritos eram canônicos ou variantes de *splicing*. Desta forma, uma lista com transcritos classificados como canônicos foi obtida através do sítio da UCSC, como critério para a triagem destes transcritos. Aqueles identificados como canônicos tiveram a redundância de suas sequências removida da mesma forma como efetuado para o primeiro repositório e em seguida, copiados para um arquivo temporário. As isoformas também foram copiadas para outro arquivo temporário, acrescidas das sequências traduzidas *in silico* das variantes de *splicing* oriundas da montagem pelo Trinity que não foram confirmadas quaisquer ESTs ou sequência do RefSeq. Por sua vez, as sequências redundantes deste arquivo também foram removidas (como no primeiro repositório) e submetidas à digestão *in silico* pelo programa Digest (também modificado). Da mesma maneira como feito para o segundo repositório, os peptídeos gerados após

esta digestão foram selecionados de forma que não pudessem estar presentes em qualquer sequência dos transcritos canônicos. Ao final, esses peptídeos selecionados tiveram também as suas sequências redundantes removidas (assim como no primeiro repositório) e, por fim, foram acrescentados ao repositório proteico-base que continha as sequências canônicas, finalizando o terceiro repositório (Figura 3.4). Um quarto repositório também foi criado para analisar três réplicas biológicas de corpo caloso de cérebro de camundongo (Sharma *et al*., 2015). Sua construção seguiu os mesmos passos e critérios utilizados para a construção do terceiro repositório humano e seus resultados seriam utilizados para posterior comparação entre eles. O estudo da busca por variantes de *splicing* alternativo no proteoma de cérebro humano e de camundongo foi feito em colaboração com o Dr. Daniel Martins-de-Souza (UNICAMP).

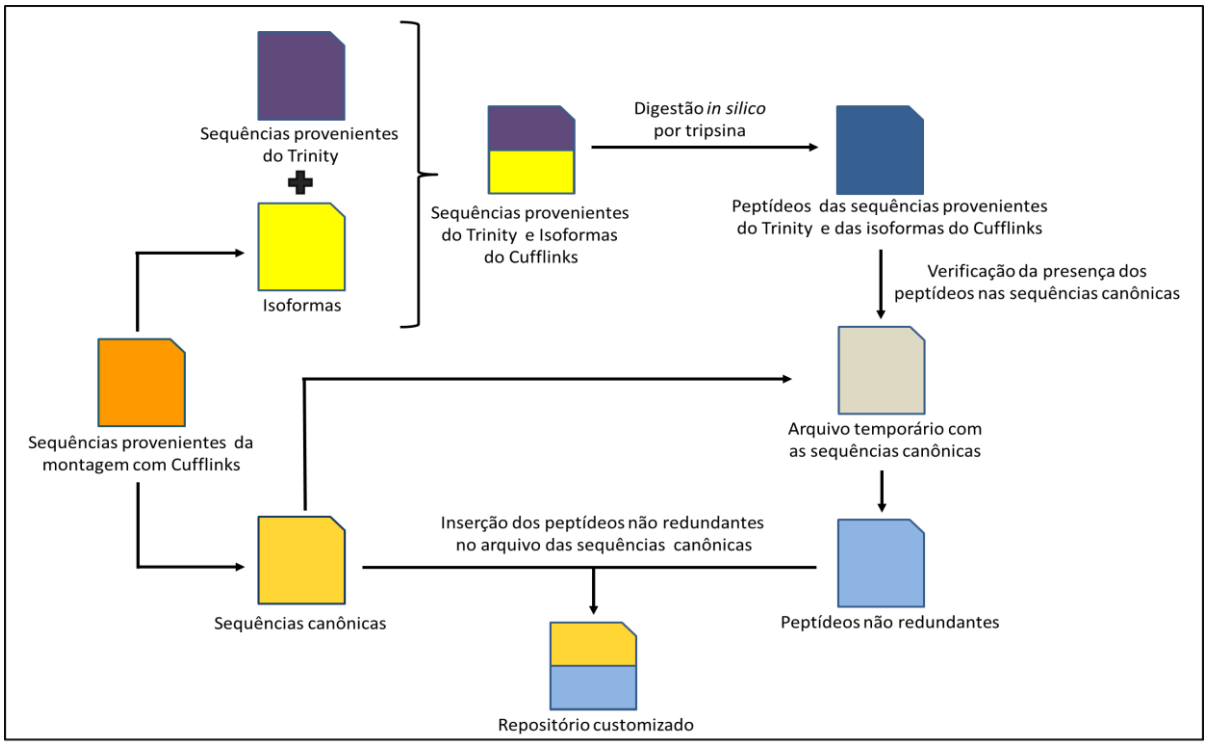

Figura 3.4 - Esquema para construção do terceiro repositório proteico personalizados.

3.5 - Dados experimentais de espectrometria de massa utilizados para a validação dos repositórios personalizados.

Dados de MS da linhagem de linfócitos T humano TIB-152 (Jurkat cells) utilizados no estudo de Sheynkman e colaboradores (2013) foram utilizados para a validação do primeiro repositório. Os dados estão disponíveis no sítio eletrônico do projeto Peptide Atlas (Desiere *et al*., 2006) sob o código “PASS00215”. Todos os 28

arquivos brutos foram agrupados e analisados ao mesmo tempo, gerando apenas uma única saída para cada critério estatístico mencionado no item 3.4.

Dados de MS da linhagem de oligondendrócitos humanos MO3.13, publicados por Iwata e colaboradores (2013), foram utilizados para a validação do segundo repositório. Os dados estão disponíveis no sítio eletrônico do projeto ProteomeXchange (Vizcaíno *et al*., 2014) sob o código “PXD000263”. Todos os 16 arquivos brutos foram agrupados e analisados ao mesmo tempo, gerando apenas uma única saída para cada critério estatístico.

Dados de MS de corpo caloso e de lobo temporal anterior humano, publicados por Martins-de-Souza e colaboradores (2013), foram utilizados para a validação do terceiro repositório. Os dados estão disponíveis no sítio eletrônico do projeto ProteomeXchange sobe o código “PXD000547” (corpo caloso) e “PXD000548” (lobo temporal anterior). As regiões mencionadas foram extraídas de oito indivíduos (três mulheres e cinco homens), totalizando 80 arquivos brutos (40 para cada região), que foram agrupados e analisados ao mesmo tempo, gerando apenas uma única saída para cada critério estatístico.

Os dados de camundongo utilizados para a avaliação do quarto repositório proteico foram publicados por Sharma e colaboradores (2015), onde várias regiões do cérebro deste organismo foram analisadas. Apenas três arquivos brutos correspondentes a três réplicas biológicas de corpo caloso foram analisados separadamente, gerando três saídas diferentes para cada critério estatístico.

3.6 - *Gene Set Enrichment Analysis* e interação entre proteínas

O *Gene Set Enrichment Analysis* (GSEA) é um método analítico que determina se um grupo de genes é estatisticamente significante em um conjunto de dados (Subramanian *et al*., 2005). Foi utilizado o modelo *Model-based Gene Set Analysis* (MGSA) em sua versão 1.13 (Bauer *et al*., 2010), através do pacote Bioconductor (versão 2.11; Reimers e Carey, 2006) e do programa estatístico R (versão 3.1.1). Esse método foi utilizado para a identificação dos processos biológicos com base no Gene Ontology (GO; Ashburner *et al*., 2000) para as isoformas detectadas no segundo e terceiro repositórios humanos personalizados. Apenas as probabilidades acima de 80% foram consideradas.

A interação entre as isoformas identificadas foi feita através de dados experimentais contidos na base *Ingenuity Pathways Knowledge Base* (IPKB) (http://www.ingenuity.com). Os genes correspondentes as isoformas identificadas pelo o segundo repositório humano foram submetidas a essa base com intuito de se detectar potenciais interações entre elas e demais proteínas.

3.7 - Validação experimental por RT-qPCR

Algumas das isoformas identificadas na linhagem de oligodendrócitos humanos (através do segundo repositório personalizado) foram validadas por PCR quantitativo (RT-qPCR). O RNA total da linhagem MO3.13 foi isolado utilizando o reagente Trizol LS (Invitrogen, Carlsbad, CA, USA) e transcrição reversa foi realizada utilizando os primers-Oligo(dT) (Superscript™ II Reverse Transcriptase, Invitrogen). Os cDNAs sintetizados foram submetidos à RT-qPCR. A amplificação dos mRNAs de β-actina foi utilizada como controle utilizando SYBR Green Master Mix (Applied Biosciences) e 7500 Real-Time PCR System (Applied Biosciences, Foster City, CA, USA). Todos os procedimentos foram realizados de acordo com as instruções do fabricante. Os *primers* das variantes de *splicing* foram desenhados para não alinharem com os transcritos canônicos.

4. RESULTADOS

Os resultados apresentados nesta tese englobam três grandes etapas desenvolvidas e aprimoradas ao longo do doutorado e que se destinavam à identificação de variantes de *splicing* no transcriptoma e sua posterior confirmação no proteoma por meio de experimentos de espectrometria de massas e RT-qPCR.

A primeira etapa destinou-se à construção do primeiro repositório proteico, denominado SpliceProt e culminou em uma publicação (Tavares *et al.*, 2014). Neste artigo, usamos o SpliceProt para analisar experimentos de MS a partir de dados oriundos de linfócitos T (linhagem Jurkat). Nesse primeiro momento, será apresentado o desenvolvimento do método para a identificação e tradução *in silico* de variantes de *splicing*, a construção de repositórios proteicos personalizados e a confirmação da presença dessas variantes no proteoma da linhagem utilizada.

A segunda etapa propõe a construção de um repositório proteico aprimorado, no qual as sequências do SpliceProt e as isoformas do Uniprot/SwissProt são utilizadas como estratégia para a caracterização das variantes de *splicing* numa linhagem de oligodendrócitos humanos (MO31.3). Este segundo momento também se destinou a aprimorar a compreensão dessas variantes no contexto celular da linhagem em estudo (Tavares *et al*., *no prelo*). Como reflexo do aprendizado no desenvolvimento de uma estratégia para análise de proteogenômica, foi publicado um artigo de revisão de literatura no qual apresentamos os desafios de proteogenômica no desenvolvimento de novas drogas (Tavares *et al*., 2015).

A terceira e última etapa utiliza a montagem de transcriptoma como estratégia para a reconstrução de transcritos em diferentes amostras de tecidos humanos e de camundongo. Por fim, a partir dos transcritos reconstruídos do tecido cerebral humano e de camundongo, dois repositórios proteicos foram construídos e aplicados em experimentos de MS de duas regiões do cérebro: lobo temporal anterior (apenas para homem) e de corpo caloso (para homem e camundongo).

4.1 - Primeira etapa: criação do repositório SpliceProt, primeiro repositório proteico personalizado e sua aplicação em experimentos espectrometria de massas oriundos de linhagem celular de linfócitos T.

4.1.1 - Criação do repositório SpliceProt e sua comparação com demais repositórios proteicos

A base de dados construída para *Homo sapiens* foi constituída por sequencias curadas de mRNAs do projeto RefSeq e ESTs da base dbEST. Após o mapeamento contra o genoma, essas sequências foram submetidas às matrizes ternárias que identificaram 267.632 variantes de *splicing*. Ao serem traduzidas com o programa TRANSEQ, essas variantes geraram 161.915 sequências polipeptídicas, das quais 159.719 não eram redundantes, formando assim o repositório SpliceProt. Este repositório foi comparado com mais três repositórios tradicionalmente utilizados como referência em experimentos de espectrometria de massas: RefSeq, ENSEMBL Gene e UniProtKB/Swiss-Prot.Foram consideradas para tal comparação, apenas sequências com pelo menos 24 aminoácidos, pois este era o tamanho da menor sequência contida no repositório do RefSeq. O SpliceProt contemplou 91,98% do repositório RefSeq, apresentando 23.770 sequências idênticas, 82,49% do repositório UniProtKB/Swiss-Prot, apresentando 16.660 sequências idênticas e 43,05% do repositório ENSEMBL Gene, apresentando 35.260 sequências idênticas (Tabela 4.1).

Tabela 4.1 - Comparação entre o número de sequências proteicas e peptídeos gerados pela digestão enzimática *in silico* dos repositórios SpliceProt, RefSeq, ENSEMBL Gene e UniProtKB/Swiss-Prot.

Comparação	Número de proteínas	Peptídeos produzidos pela digestão enzimática *in silico*			
		Tripsina	Lys-C	Glu-C_bicarb	Glu-C_phosph
SpliceProt vs. RefSeq					
Sequências idênticas	23.770	480.878	255.091	270.986	502.846
Sequências únicas do SpliceProt	135.949	219.614	107.804	106.122	173.993
Sequências únicas do RefSeq	2.071	5.654	2.910	3.371	5.784
SpliceProt vs. ENSEMBL Gene					
Sequências idênticas	35.260	513.678	271.712	288.629	531.885
Sequências únicas do SpliceProt	124.459	186.814	91.183	88.479	144.954
Sequências únicas do ENSEMBL Gene	46.640	64.461	37.155	39.755	58.928
SpliceProt vs. UniprotKB/Swiss-Prot					
Sequências idênticas	16.660	481.766	255.064	271.037	504.709
Sequências únicas do SpliceProt	143.059	218.726	107.831	106.071	172.130
Sequências únicas do UniprotKB/Swiss-Prot	3.535	24.552	12.562	13.501	24.861

4.1.2 - Digestão *in silico* do SpliceProt e demais repositórios proteicos

Com o objetivo de avaliar a contribuição do SpliceProt para experimentos de espectrometria de massas, a partir do programa Digest (pacote EMBOSS), o repositório foi submetido à digestão *in silico* pelas enzimas tripsina, a endoproteinase *Lys-C* e as glutamil endoproteinases utilizadas em meio com bicarbonato de amônio (*Glu-C bicarb*) e em meio com fosfato *(Glu-C phosph*). Considerando apenas peptídeos com tamanho entre 6 e 24 aminoácidos, a análise indicou que a enzima tripsina permitiu a maior produção de peptídeos únicos, seguida das endoproteinases *Glu-C phosph* e *Glu-C bicarb*, e *Lys-C*. Visando a comparação entre com o SpliceProt, os demais repositórios também passaram pela digestão *in silico*. O SpliceProt apresentou o maior número de peptídeos não redundantes (700.492), seguido por ENSEMBL Gene (578.139), UniProtKB/Swiss-Prot (506.318) e RefSeq (486.532).

Ao compararmos a digestão por tripsina, o SpliceProt possui 219.614 peptídeos exclusivos e 480.878 em comum com o repositório RefSeq. Contra o ENSEMBL Gene, foram identificados 513.678 peptídeos em comum, 64.461 peptídeos exclusivos do ENSEMBL Gene e 186.814 peptídeos exclusivos do repositório exclusivos do SpliceProt, indicando 2,8 peptídeos exclusivos a mais que repositório em comparação. Quando comparado com o UniProtKB/Swiss-Prot, foram identificados 24.552 peptídeos exclusivos deste repositório, 481.766 peptídeos em comum com o SpliceProt e 218.726 peptídeos exclusivos do nosso conjunto peptídeos, indicando desta vez, 8,9 peptídeos exclusivos a mais que UniProtKB/Swiss-Prot (Tabela 4.1). Através de um diagrama de Venn, a distribuição dos peptídeos indica que a maior parte dos peptídeos era compartilhada entre os repositórios analisados independentemente da enzima escolhida (Figura 4.1).

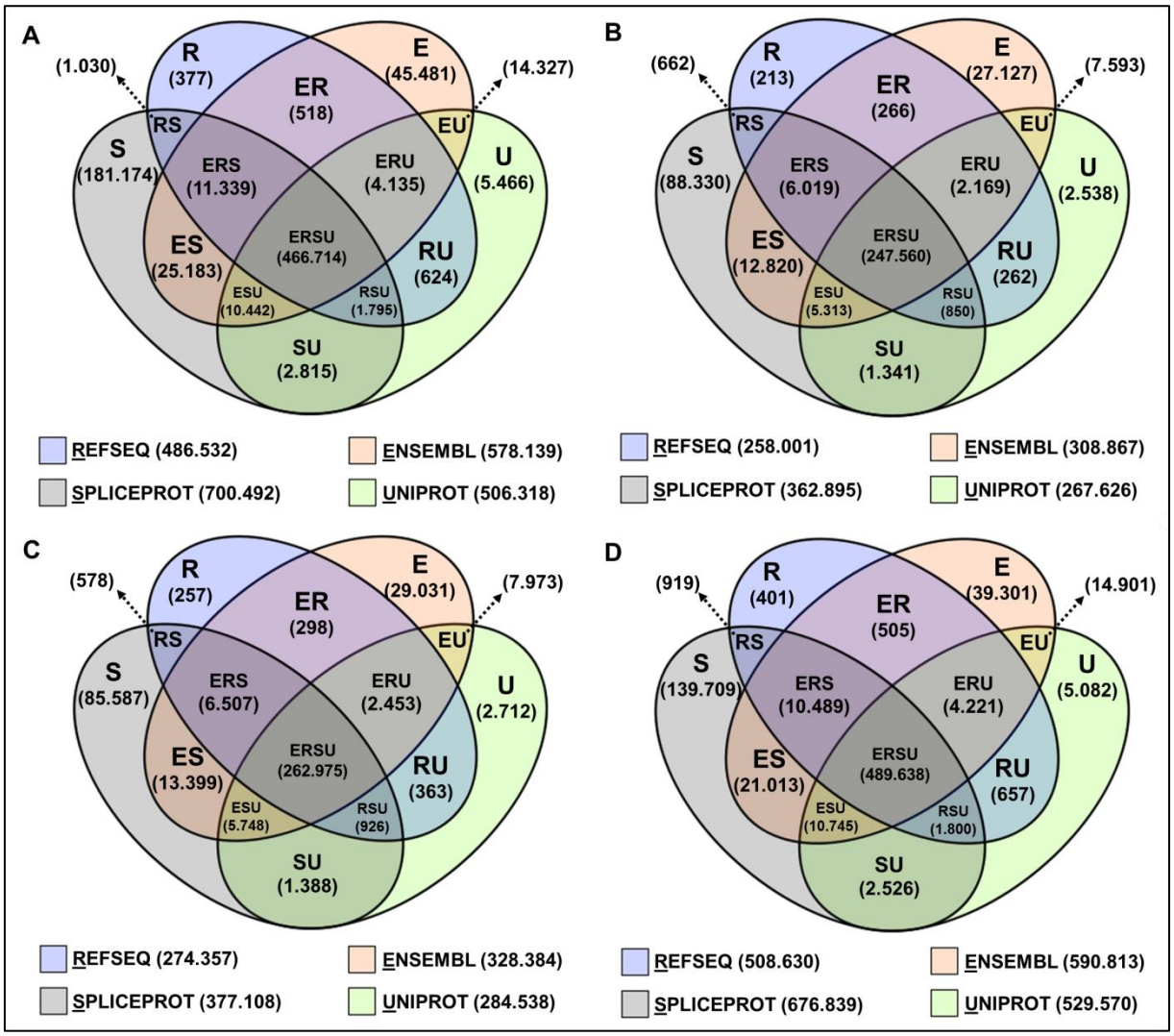

Figura 4.1. Distribuição dos peptídeos entre os repositórios analisados através da digestão *in silico* por tripsina (A), *Lys-C* (B), *Glu-C_bicarb* (C) e *Glu-C_phosph* (D).

4.1.3 - Aplicação do repositório personalizado em experimentos de espectrometria de massas

Dados de espectrometria de massas da linhagem de linfócitos T (Jurkat *cells*) foram utilizados para validar o repositório criado pela presente abordagem (Sheynkman *et al*., 2013). Para isso, os peptídeos digeridos por tripsina não redundantes do Spliceprot foram anexados ao arquivo contendo as sequências proteicas do UniProtKB/Swiss-Prot, produzindo um novo repositório personalizado. Usando esta mesma abordagem, foram criados outros repositórios personalizados utilizando ENSEMBL Gene e Refseq, para verificarmos a distribuição dos peptídeos anexados de acordo com cada enzima utilizada. É possível verificar que a maior parte dos peptídeos é compartilhada entre repositórios criados independentemente da enzima escolhida (Figura 4.2).

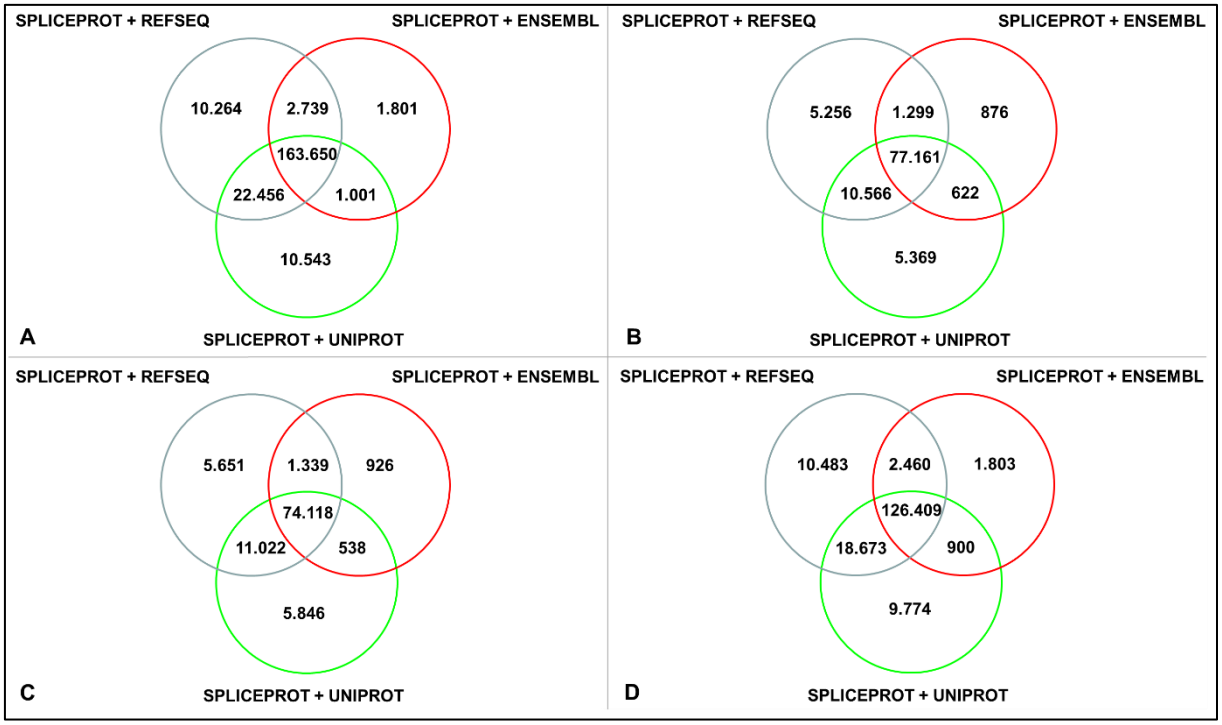

Figura 4.2 - Diagramas de Venn demonstrando a distribuição dos peptídeos do SpliceProt anexados aos repositórios personalizados para as enzimas: tripsina (A), *Lys-C* (B), Glu-C_bicarb (C) e Glu-C_phosph (D).

Foram identificadas 6.726 sequências proteicas usando FDR de 5%. Em análise adicional utilizando PEP de 1%, foram identificados 225 peptídeos oriundos do SpliceProt anexados ao repositório personalizado. Destes, 52 eram artefatos inerentes à metodologia utilizada para identificação de variantes de *splicing* e 173 representavam peptídeos encontrados no experimento. O repositório UniProtKB/TrEMBL foi utilizado com o intuito de refinar a identificação desses 173 peptídeos. Após a buscas desses peptídeos nesse repositório, 54 eram exclusivos

do SpliceProt e não seriam encontrados caso fosse utilizado somente o repositório UniProtKB/TrEMBL. Por fim, comparando os peptídeos encontrados pela nossa abordagem com os 57 encontrados no experimento realizado por Sheynkman e colaboradores (2013), 10 peptídeos eram idênticos para nas duas abordagens (Tabela 4.2).

Tabela 4.2 - Peptídeos encontrados pela abordagem utilizando o repositório personalizado e pela abordagem de Sheynkman e colaboradores (2013).

Peptídeo	UniprotKB/TrEMBL	Símbolo Gênico
AQEDALAQQAFEEAR	não	*CBFB*
EENPEGPPNANEDYR	não	*STK39*
SSTDSLPGELR	não	*NCOA5*
DGYELSPTAAANFTR	não	*PSMB2*
YISLIYTNYAGK	não	*GSTP1*
SEVADFEPER	não	*LETM1*
ATPEPGDEGEPGR	sim	*ARHGEF1*
LGPGGLDPVEELQK	sim	*CDC37*
MLDAEDIVGTLRPDEK	sim	*ACTN4*
GYATDESTVSSVQGSR	sim	*FXR1*

4.2 - Segunda etapa: construção do segundo repositório personalizado e sua aplicação em experimentos espectrometria de massas oriundos de linhagem celular de oligodendrócitos humanos

O segundo repositório proteico personalizado, constituído por sequências proteicas canônicas do UniProt/SwissProt e peptídeos não-redundantes oriundos das isoformas do SpliceProt e do UniProt/SwissProt, foi utilizado em experimentos de MS oriundos de linhagem celular de oligodendrócitos humanos (Iwata *et al*., 2013). O novo repositório personalizado era composto por 20.150 sequências canônicas e 204.294 peptídeos não redundantes, totalizando 224.453 sequências proteicas (Tavares *et al*., *no prelo*).

4.2.1 - Variantes de *splicing* encontradas no proteoma de oligodendrócitos humanos

O experimento com o repositório proteico personalizado resultou na identificação de 2.081 proteínas, confirmados por 12.990 peptídeos não redundantes, usando como critério de restrição estatística PEP de 1%. Entre os peptídeos encontrados, 45 correspondiam a 39 variantes e 38 genes identificados exclusivamente pelos peptídeos anexados ao repositório personalizado (Tabela 4.3). Resultados similares também foram obtidos utilizando outros *softwares* (Mascot e MaxQuant) que identificam proteínas em dados de espectrometria de massas, tanto para o número de isoformas encontradas quanto para o número de peptídeos identificados por ambos programas (Tabelas 4.4 e 4.5). A interação entre as isoformas é representada pela figura 4.3 e algumas delas tiveram seus resultados analisados de forma mais detalhada na discussão desta tese (Figura 4.4).

Tabela 4.3 - Lista das 39 isoformas encontradas com o segundo repositório proteico personalizado em oligodendrócitos humanos.

Símbolo gênico	Identificador Uniprot/Refseq ID	≥ *cut off**
DBNL	Q9UJU6-2	Sim
DNM2	P50570-3	Sim
EEF1D	P29692-3	Sim
GLS	O94925-3	Sim
HNRNPAB	Q99729-3	Sim
IKBIP	Q70UQ0-4	Sim
KRAS	P01116-2	Sim
PDLIM3	Q53GG5-2	Sim
RAVER1	Q8IY67-2	Sim
SDR39U1	Q9NRG7-2	Sim
SPTAN1	Q13813-3	Sim
SREK1	Q8WXA9-2	Sim
SUGT1	Q9Y2Z0-2	Sim
TPD52L2	O43399-2	Sim
TPM1	P09493-5	Sim
TPM3	P06753-2	Sim
HNRNPC	P07910-2	Sim
API5	Q9BZZ5-2	Não
CAPZB	P47756-2	Não
GANAB	Q14697-2	Não
GLOD4	Q9HC38-2	Não
HNRNPK	P61978-3	Não
HNRNPM	XP_005272536.1	Não
LOXL3	P58215-3	Não
MAP4	P27816-5	Não
MFF	Q9GZY8-2	Não
MRPL43	Q8N983-4	Não
NDUFV3	P56181-2	Não
NOL3	O60936-2	Não
NOLC1	Q14978-3	Não
NUDT4	Q9NZJ9-2	Não
PKM	P14618-2	Não
PNKD	Q8N490-2	Não
RAVER1	NP_597709.2	Não
SET	Q01105-2	Não
SF1	Q15637-6	Não
TOR1AIP1	Q5JTV8-3	Não
TPM2	P07951-2	Não
ZNF414	Q96IQ9-2	Não

*razão de PSMs e peptídeos únicos ≥ 1,4.

Tabela 4.4 - Comparação entre os programas Proteome Discoverer, MaxQuant e Mascot para a identificação de variantes de splicing utilizando o repositório personalizado.

Programa	Número de peptídeos não redundantes	Número de isoformas não redundantes	Número de genes não redundantes	Validação estatística
Proteome Discoverer	45	39	38	1% PEP
MaxQuant	48	42	40	1% FDR
Mascot	44	39	36	1% PEP

Tabela 4.5 - Número de peptídeos redundantes e não-redundantes entre o Proteome Discoverer e os programas MaxQuant e Mascot.

Comparação	Número de peptídeos
Proteome Discoverer vs. MaxQuant	
Peptídeos idênticos	33
Peptídeos únicos do Proteome Discoverer	12
Peptídeos únicos do MaxQuant	15
Total de peptídeos não-redundantes	60
Proteome Discoverer vs. Mascot	
Peptídeos idênticos	28
Peptídeos únicos do Proteome Discoverer	17
Peptídeos únicos do Mascot	16
Total de peptídeos não-redundantes	61

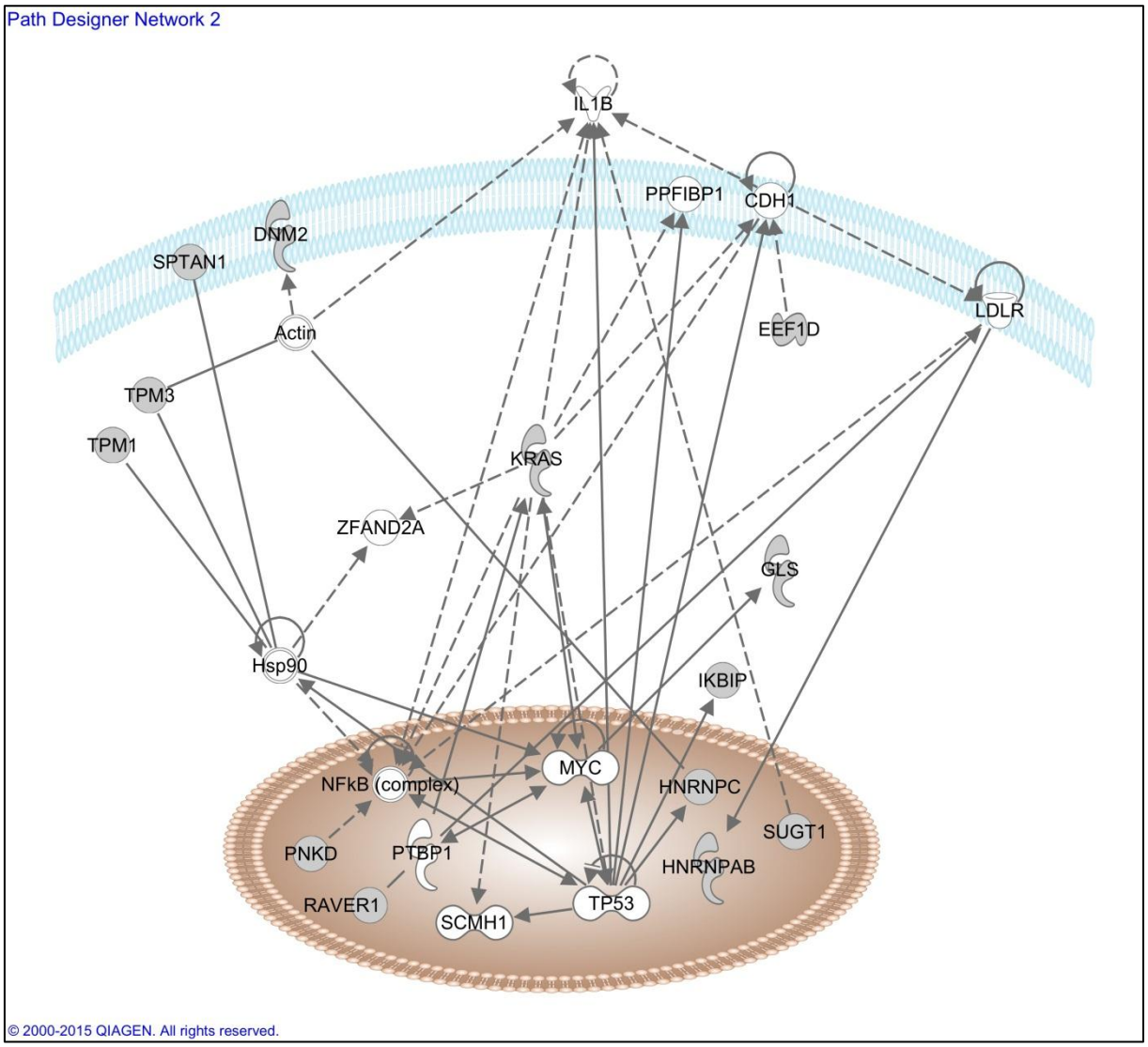

Figura 4.3 - Interação proteína-proteína entre as isoformas encontradas (proteínas em cinza) na linhagem de oligodendrócitos humanos e seus prováveis alvos de interação (proteínas em branco).

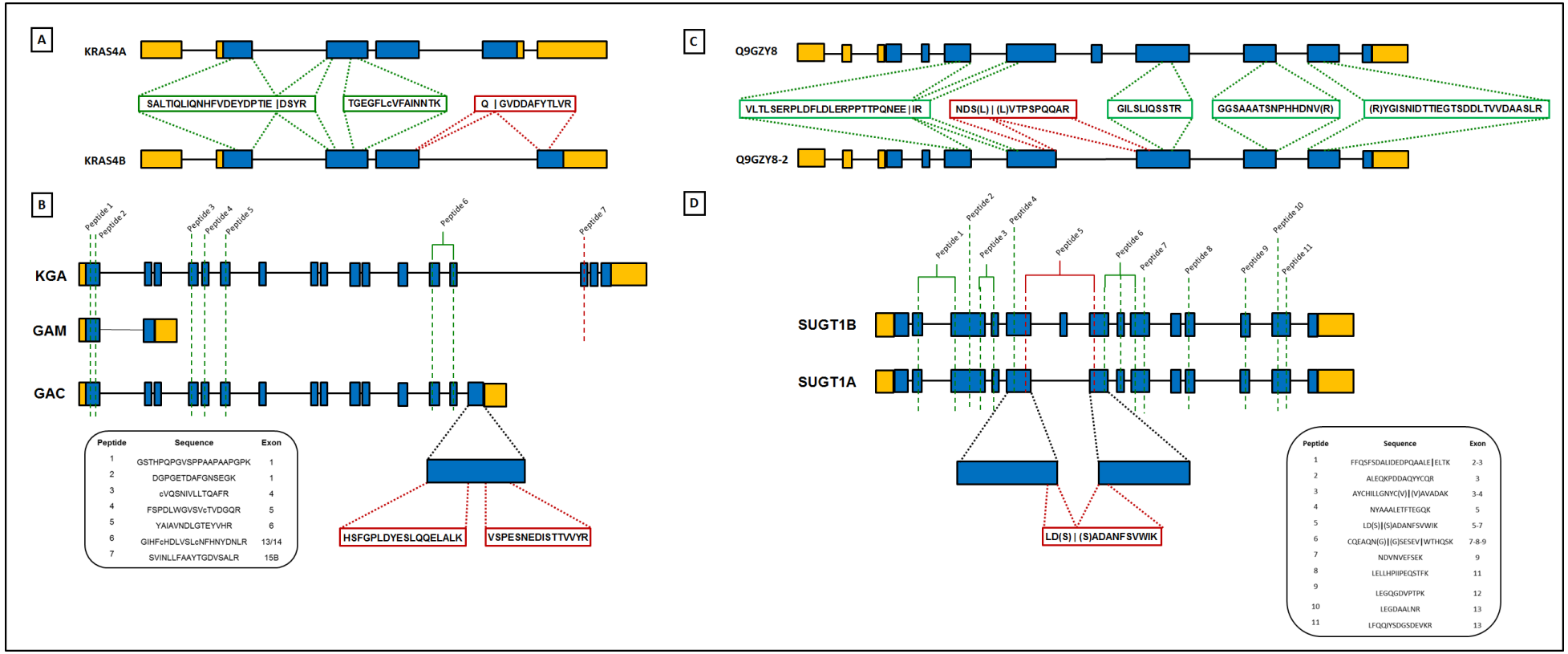

Figura 4.4 - Representação esquemática das variantes de *splicing* identificadas e seus respectivos peptídeos: (A) *KRAS*, (B) *GLS*, (C) *MFF* e (D) *SUGT1*. Éxons são representados pelos retângulos em amarelo (UTRs) e em azul (sequência codificadora) e íntrons são representados pelo símbolo “I” (do Inglês, *pipe*). Os peptídeos únicos são representados por linhas vermelhas tracejadas e pontilhadas e os peptídeos compartilhados por linhas verdes tracejadas e pontilhadas. Os aminoácidos entre parênteses indicam a fase do íntron.

4.2.2 - Proposta de um novo critério para atribuir confiabilidade aos peptídeos oriundos de eventos de splicing alternativo identificados em experimentos de espectrometria de massas

Além do número de peptídeos únicos (UP) e o número de espectros identificados (PSM) como modo para atribuir confiabilidade aos peptídeos detectados pelo espectrômetro de massas, foi proposto um novo critério baseado na razão entre o número de espectros por peptídeos únicos para identificação de variantes de *splicing*. Este cálculo foi feito a partir da razão entre PSM/UP das proteínas canônicas com pelo menos dois peptídeos únicos identificados, resultando em 1,4 PSM/UP em média para o conjunto de dados analisado (Figura 4.5). Se usássemos esse critério para a seleção das isoformas, 17 teriam um *cut off* ≥ 1,4 (Tabela 4.3).

Outra sugestão para a análise foi a contagem dos espectros das proteínas canônicas e das isoformas identificadas. Caso o número de espectros da isoforma fosse maior que o da canônica, é sugestivo que a isoforma está sendo mais expressa. Apenas um caso se enquadrou nessa condição representado pelo gene *SDR39U1* e suas proteínas Q9NRG7 e Q9NRG7-2 que compartilham um peptídeo com 3 PSMs. Porém, a isoforma possui 4 PSMs para o seu peptídeo exclusivo (Figura 4.6). Todas as análises desenvolvidas nesse tópico foram realizadas com a colaboração do estudante de doutorado do nosso grupo, Gabriel Wajnberg.

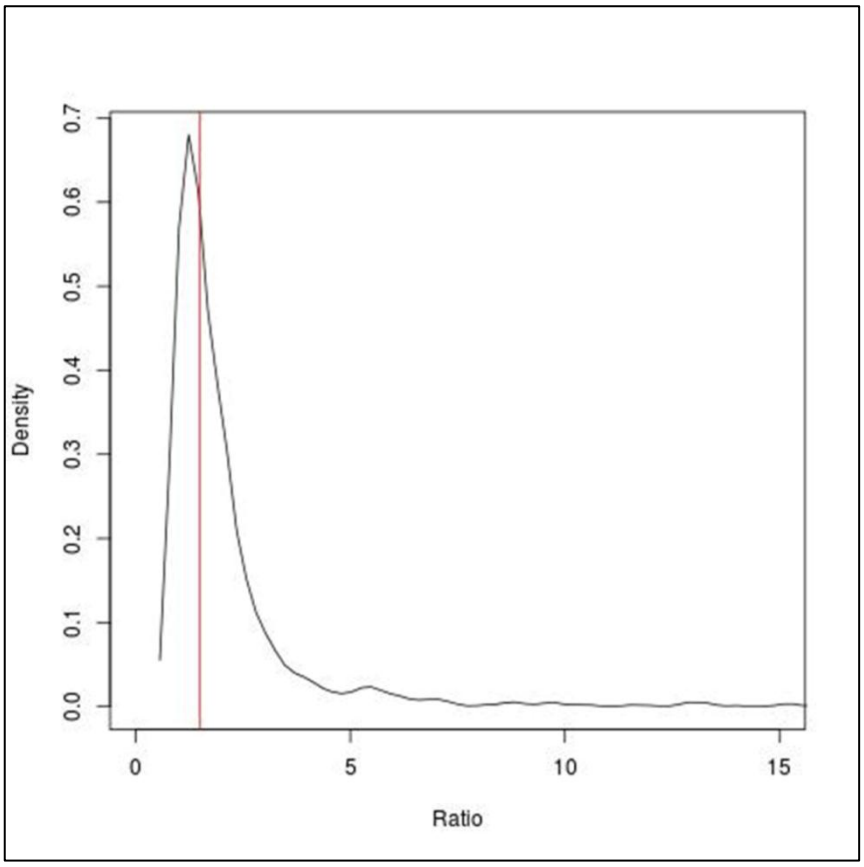

Figura 4.5 - Razão entre o número de espectros encontrados (PSMs) por peptídeos únicos (UP) para as proteínas canônicas encontrados nos dados de oligondendrócitos. A linha vermelha indica a média encontrada de 1,4 no conjunto de dados.

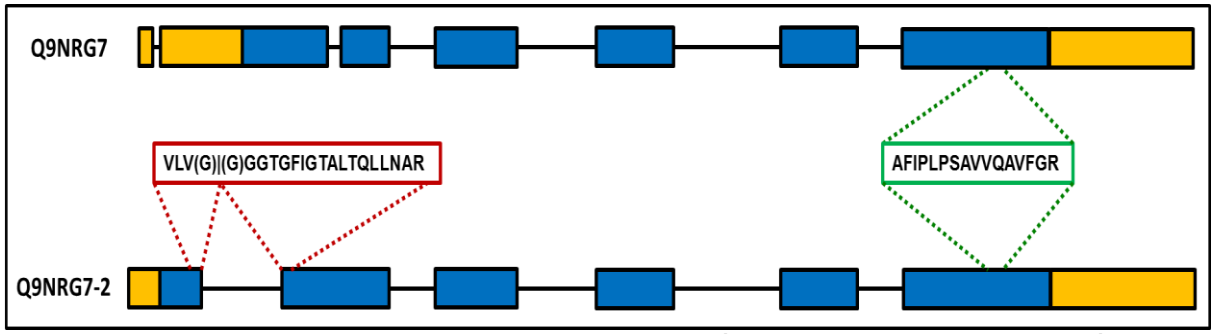

Figura 4.6 - Transcritos do gene *SDR39U1* onde o número de espectros do peptídeo (em vermelho) exclusivo da isoforma Q9NRG7-2 é maior que o número de espectros do peptídeo (em verde) compartilhado com a proteína canônica Q9NRG7.

4.2.3 - Validação da expressão de variantes de *splicing* encontradas na linhagem de oligodendrócitos humanos por RT-qPCR.

Com o intuito de detectar a expressão de mRNAs de algumas das variantes de *splicing* encontradas no proteoma da linhagem de oligodendrócitos humanos, cinco variantes (EEF1D, KRAS, MFF, SDR39U1, e SUGT1) foram selecionadas para validação experimental. Todas foram detectadas no nível de transcriptoma da linhagem MO3.13 (Figura 4.7). Estes experimentos foram realizados em colaboração com a Dra Patricia Savio de Araujo Souza (UFF e INCA).

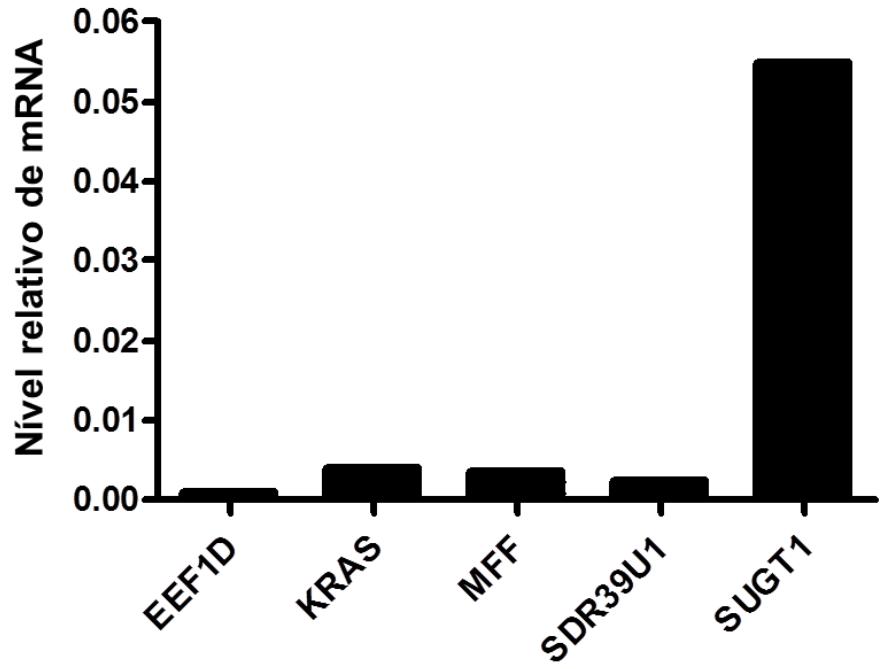

Figura 4.7 - Validação experimental de cinco isoformas (EEF1D, KRAS, MFF, SDR39U1, e SUGT1) encontradas na linhagem de oligodendrócitos humanos por RT-qPCR.

4.3 - Terceira etapa: utilização do pipeline de montagem de transcriptoma na construção de repositórios proteicos para aplicação em experimentos de MS de amostras cerebrais de homem e camundongo.

Nesta etapa serão apresentados dados relativos à montagem de dados de transcriptoma (RNA-Seq) de homem e camundongo. Os resultados seguem os três passos ilustrados na metodologia, seguidos pela identificação das variantes de *splicing* e sua tradução computacional.

4.3.1 - Alinhamento contra o genoma (1º Alinhamento) e montagem com genoma de referência com Cufflinks.

Os *reads* originais passaram por filtro de qualidade e remoção dos adaptadores (*trimming)* antes do alinhamento contra o genoma de *Homo sapiens* e *Mus musculus*. Foi observada média de aproveitamento de aproximadamente 90% e 95% para as corridas de homem e camundongo, respectivamente (Tabelas 4.6 e 4.7).

Tabela 4.6 - Número de *reads* originais e tratados após o *trimming* e sua porcentagem de aproveitamento para o organismo *Homo sapiens*.

Tecido	*Reads* originais	*Reads* após *trimming*	Porcentagem de aproveitamento
Cérebro-RT1	15.674.315	15.252.976	97,31%
Cérebro-RT2	28.523.713	26.816.007	94,01%
Fígado-RT1	37.803.015	33.740.075	89,25%
Fígado-RT2	26.773.844	25.416.053	94,93%
Baço-RT1	12.534.187	10.838.181	86,47%
Baço-RT2	24.511.680	23.544.936	96,06%
Pâncreas-RT1	11.313.824	11.052.520	97,69%
Pâncreas-RT2	26.087.337	24.660.792	94,53%
Tecido Adiposo-RT1	21.851.075	20.151.108	92,22%
Tecido Adiposo-RT2	34.014.252	31.597.032	92,89%
Glândula Adrenal-RT1	68.399.538	62.683.085	91,64%
Glândula Adrenal-RT2	44.343.116	42.771.030	96,45%
Ovário-RT1	52.282.560	47.874.719	91,57%
Pulmão-RT1	18.971.601	17.315.531	91,27%
Pulmão-RT2	24.893.990	23.718.384	95,28%
Intestino Grosso-RT1	41.887.084	37.104.934	88,58%
Intestino Grosso-RT2	29.918.829	27.590.262	92,22%
Intestino Delgado-RT1	52.215.140	47.385.894	90,75%
Intestino Delgado-RT2	29.794.253	28.423.132	95,40%
Rim-RT1	14.083.463	11.878.324	84,34%
Rim-RT2	28.492.976	25.836.709	90,68%
Coração-RT1	50.084.789	44.838.320	89,52%
Coração-RT2	33.438.516	32.247.500	96,44%
Testículo-RT1	20.723.459	18.429.570	88,93%
Testículo-RT2	30.834.929	29.600.597	96,00%

Tabela 4.7 - Número de *reads* originais e tratados após o *trimming* e sua porcentagem de aproveitamento para o organismo *Mus musculus*.

Tecido	*Reads* originais	*Reads* após *trimming*	Porcentagem de aproveitamento
Cérebro-RT1	23.895.350	23.491.467	98,31%
Cérebro-RT2	25.879.769	23.186.277	89,59%
Fígado-RT1	32.182.775	30.694.632	95,38%
Fígado-RT2	26.185.733	24.882.163	95,02%
Baço-RT1	40.069.205	39.089.562	97,56%
Baço-RT2	24.363.823	22.775.245	93,48%
Pâncreas-RT1	17.333.609	16.413.095	94,69%
Pâncreas-RT2	27.677.997	26.087.828	94,25%
Tecido Adiposo-RT1	28.938.530	27.799.563	96,06%
Tecido Adiposo-RT2	28.095.734	26.941.233	95,89%
Glândula Adrenal-RT1	36.616.875	35.401.101	96,68%
Glândula Adrenal-RT2	22.398.475	21.534.889	96,14%
Ovário-RT1	31.557.381	30.532.813	96,75%
Pulmão-RT1	21.023.171	20.059.980	95,42%
Pulmão-RT2	23.749.496	22.346.829	94,09%
Intestino Grosso-RT1	38.051.335	35.955.709	94,49%
Intestino Grosso-RT2	31.369.371	28.411.929	90,57%
Intestino Delgado-RT1	38.286.584	36.776.044	96,05%
Intestino Delgado-RT2	32.803.922	31.534.589	96,13%
Rim-RT1	37.729.075	34.539.530	91,55%
Rim-RT2	27.023.549	25.872.580	95,74%
Coração-RT1	29.892.315	28.727.380	96,10%
Coração-RT2	25.489.392	24.611.316	96,56%
Testículo-RT1	52.230.061	49.557.011	94,88%
Testículo-RT2	30.611.590	29.332.481	95,82%

O alinhamento dos *reads* de cada corrida gerou um arquivo de alinhamento que foi submetido à montagem com genoma de referência através do programa Cufflinks. A partir do critério de corte de FPKM > 0, foi observada média de 30.232 e 22.269 transcritos montados para as corridas de homem e camundongo (Tabela 4.8), gerados por uma média de 15.595 genes em homem e 15.258 camundongo (Tabela 4.9). Entretanto, a nossa base de dados não possuía a referência de uma pequena fração dos transcritos reconstruídos (Tabela 4.10).

Tabela 4.8 - Número de transcritos montados pelo programa Cufflinks para os organismos *Homo sapiens* e *Mus musculus*.

Tecido	Homem	Camundongo
Cérebro-RT1	27.405	21.593
Cérebro-RT2	31.874	23.778
Fígado-RT1	28.551	20.236
Fígado-RT2	30.134	21.358
Baço-RT1	26.908	22.606
Baço-RT2	31.362	18.872
Pâncreas-RT1	29.134	20.938
Pâncreas-RT2	30.028	21.468
Tecido Adiposo-RT1	30.358	23.595
Tecido Adiposo-RT2	32.232	24.851
Glândula Adrenal-RT1	31.546	21.281
Glândula Adrenal-RT2	32.110	22.847
Ovário-RT1	31.351	22.078
Pulmão-RT1	28.173	21.362
Pulmão-RT2	31.692	23.360
Intestino Grosso-RT1	29.148	21.933
Intestino Grosso-RT2	32.162	23.720
Intestino Delgado-RT1	26.580	21.456
Intestino Delgado-RT2	29.166	23.506
Rim-RT1	27.130	21.308
Rim-RT2	31.543	22.838
Coração-RT1	29.303	19.772
Coração-RT2	30.627	22.311
Testículo-RT1	32.702	24.469
Testículo-RT2	34.596	25.206

Tabela 4.9 - Número de genes atingidos pela montagem com Cufflinks com FPKM FPKM > 0 para as corridas de homem e camundongo.

Tecido	Homem	Camundongo
Cérebro-RT1	14.294	14.836
Cérebro-RT2	16.158	16.126
Fígado-RT1	14.847	13.898
Fígado-RT2	15.301	14.472
Baço-RT1	14.095	15.394
Baço-RT2	15.958	13.068
Pâncreas-RT1	15.177	14.338
Pâncreas-RT2	15.614	14.556
Tecido Adiposo-RT1	15.694	16.406
Tecido Adiposo-RT2	16.413	17.007
Glândula Adrenal-RT1	16.177	14.669
Glândula Adrenal-RT2	16.275	15.463
Ovário-RT1	16.248	15.278
Pulmão-RT1	14.799	14.798
Pulmão-RT2	16.089	15.816
Intestino Grosso-RT1	15.360	15.136
Intestino Grosso-RT2	16.343	16.109
Intestino Delgado-RT1	13.976	14.815
Intestino Delgado-RT2	14.877	15.968
Rim-RT1	14.266	14.676
Rim-RT2	16.116	15.450
Coração-RT1	15.193	13.518
Coração-RT2	15.501	15.077
Testículo-RT1	17.228	17.148
Testículo-RT2	17.886	17.422

Tabela 4.10 - Número de transcritos com informação sobre seu gene de origem em nossa base de dados.

Tecido	Homem		Camundongo	
	Número de transcritos com gene identificado	Número de transcritos sem gene identificado	Número de transcritos com gene identificado	Número de transcritos sem gene identificado
Cérebro-RT1	27.367	38	21.569	24
Cérebro-RT2	31.820	54	23.745	33
Fígado-RT1	28.499	52	20.215	21
Fígado-RT2	30.080	54	21.332	26
Baço-RT1	26.862	46	22.581	25
Baço-RT2	31.296	92	18.862	10
Pâncreas-RT1	29.079	55	20.916	22
Pâncreas-RT2	29.973	55	21.444	24
Tecido Adiposo-RT1	30.302	56	23.556	39
Tecido Adiposo-RT2	32.163	69	24.805	46
Glândula Adrenal-RT1	31.486	60	21.254	27
Glândula Adrenal-RT2	32.037	73	22.815	32
Ovário-RT1	31.276	75	22.039	39
Pulmão-RT1	28.121	52	21.340	22
Pulmão-RT2	31.631	61	23.331	29
Intestino Grosso-RT1	29.094	54	21.902	31
Intestino Grosso-RT2	32.094	68	23.686	34
Intestino Delgado-RT1	26.524	56	21.426	30
Intestino Delgado-RT2	29.103	63	23.465	41
Rim-RT1	27.085	45	21.286	22
Rim-RT2	31.478	65	22.806	32
Coração-RT1	29.245	58	19.755	17
Coração-RT2	30.571	56	22.286	25
Testículo-RT1	32.610	92	24.371	98
Testículo-RT2	34.477	119	25.119	87

4.3.2 - Alinhamento contra as sequências dos transcritos (2º Alinhamento) e 1ª seleção de *reads*.

Após a montagem dos transcritos usando a anotação de transcritos de referência, foi realizada a primeira seleção de *reads* para identificar aqueles que foram utilizados por esta montagem e selecionar aqueles que não foram utilizados. Foi encontrada média de aproximadamente de 70% e 60% de *reads* não utilizados pelo Cufflinks para a montagem (Tabelas 4.11 e 4.12).

Tabela 4.11 - Número de *reads* após o *trimming*, utilizados e não utilizados pelo programa Cufflinks para o organismo *Homo sapiens*.

Tecido	*Reads* após o *trimming*	Utilizados pelo Cufflinks	Não utilizados pelo Cufflinks
Cérebro-RT1	15.252.976	5.068.177	10.184.799
Cérebro-RT2	26.816.007	9.184.496	17.631.511
Fígado-RT1	33.740.075	11.416.693	22.323.382
Fígado-RT2	25.416.053	9.191.195	16.224.858
Baço-RT1	10.838.181	3.354.070	7.484.111
Baço-RT2	23.544.936	7.120.821	16.424.115
Pâncreas-RT1	11.052.520	4.107.485	6.945.035
Pâncreas-RT2	24.660.792	8.449.218	16.211.574
Tecido Adiposo-RT1	20.151.108	5.059.398	15.091.710
Tecido Adiposo-RT2	31.597.032	10.928.599	20.668.433
Glândula Adrenal-RT1	62.683.085	22.029.469	40.653.616
Glândula Adrenal-RT2	42.771.030	15.727.158	27.043.872
Ovário-RT1	47.874.719	16.254.196	31.620.523
Pulmão-RT1	17.315.531	6.983.287	10.332.244
Pulmão-RT2	23.718.384	9.269.610	14.448.774
Intestino Grosso-RT1	37.104.934	8.030.615	29.074.319
Intestino Grosso-RT2	27.590.262	6.913.698	20.676.564
Intestino Delgado-RT1	47.385.894	12.704.281	34.681.613
Intestino Delgado-RT2	28.423.132	8.787.962	19.635.170
Rim-RT1	11.878.324	3.345.830	8.532.494
Rim-RT2	25.836.709	7.459.726	18.376.983
Coração-RT1	44.838.320	10.955.725	33.882.595
Coração-RT2	32.247.500	10.802.857	21.444.643
Testículo-RT1	18.429.570	6.025.888	12.403.682
Testículo-RT2	29.600.597	10.416.829	19.183.768

Tabela 4.12 - Número de *reads* após o *trimming*, utilizados e não utilizados pelo programa Cufflinks para o organismo *Mus musculus*.

Tecido	*Reads* após o *trimming*	Utilizados pelo Cufflinks	Não utilizados pelo Cufflinks
Cérebro-RT1	23.491.467	8.584.244	14.907.223
Cérebro-RT2	23.186.277	10.693.233	12.493.044
Fígado-RT1	30.694.632	12.092.656	18.601.976
Fígado-RT2	24.882.163	13.545.040	11.337.123
Baço-RT1	39.089.562	17.172.219	21.917.343
Baço-RT2	22.775.245	2.850.649	19.924.596
Pâncreas-RT1	16.413.095	6.743.643	9.669.452
Pâncreas-RT2	26.087.828	10.973.540	15.114.288
Tecido Adiposo-RT1	27.799.563	11.095.577	16.703.986
Tecido Adiposo-RT2	26.941.233	13.493.497	13.447.736
Glândula Adrenal-RT1	35.401.101	9.494.238	25.906.863
Glândula Adrenal-RT2	21.534.889	10.711.894	21.534.889
Ovário-RT1	30.532.813	12.333.065	18.199.748
Pulmão-RT1	20.059.980	10.336.990	9.722.990
Pulmão-RT2	22.346.829	12.604.033	9.742.796
Intestino Grosso-RT1	35.955.709	12.218.609	23.737.100
Intestino Grosso-RT2	28.411.929	13.792.516	14.619.413
Intestino Delgado-RT1	36.776.044	12.173.103	24.602.941
Intestino Delgado-RT2	31.534.589	16.641.847	14.892.742
Rim-RT1	34.539.530	10.937.598	23.601.932
Rim-RT2	25.872.580	12.901.335	12.971.245
Coração-RT1	28.727.380	3.378.937	25.348.443
Coração-RT2	24.611.316	8.339.668	16.271.648
Testículo-RT1	49.557.011	22.198.392	27.358.619
Testículo-RT2	29.332.481	186.719	29.145.762

4.3.3 - Alinhamento contra as sequências gênicas (3º Alinhamento), 2ª seleção de reads e montagem *de novo* com Trinity

Os *reads* não utilizados para montagem de novos transcritos pelo programa Cufflinks, foram alinhados contra as sequências gênicas de seus respectivos organismos (3º Alinhamento). Esta etapa teve como intuito identificar aqueles *reads* que mapearam corretamente em apenas um único gene. Após este terceiro mapeamento, foi observada média de aproveitamento de aproximadamente 3% e 5% de *reads* não utilizados pelo programa Cufflinks e que seguiram o critério mencionado para homem e camundongo, respectivamente (Tabelas 4.13 e 4.14).

Tabela 4.13 - Número de *reads* não utilizados pelo Cufflinks, número de *reads* selecionados e não selecionados no o terceiro alinhamento para organismo *Homo sapiens*.

Tecido	*Reads* não utilizados pelo Cufflinks	*Reads* não selecionados no o 3º alinhamento	Selecionados no 3º alinhamento
Cérebro-RT1	10.184.799	9.858.160	326.639
Cérebro-RT2	17.631.511	17.101.811	529.700
Fígado-RT1	22.323.382	21.797.925	525.457
Fígado-RT2	16.224.858	15.868.750	356.108
Baço-RT1	7.484.111	7.213.153	270.958
Baço-RT2	16.424.115	15.939.566	484.549
Pâncreas-RT1	6.945.035	6.711.513	233.522
Pâncreas-RT2	16.211.574	15.877.463	334.111
Tecido Adiposo-RT1	15.091.710	14.737.523	354.187
Tecido Adiposo-RT2	20.668.433	20.004.275	664.158
Glândula Adrenal-RT1	40.653.616	39.539.581	1.114.035
Glândula Adrenal-RT2	27.043.872	26.244.733	799.139
Ovário-RT1	31.620.523	30.374.594	1.245.929
Pulmão-RT1	10.332.244	10.053.484	278.760
Pulmão-RT2	14.448.774	14.067.373	381.401
Intestino Grosso-RT1	29.074.319	28.458.551	615.768
Intestino Grosso-RT2	20.676.564	20.242.094	434.470
Intestino Delgado-RT1	34.681.613	33.496.748	1.184.865
Intestino Delgado-RT2	19.635.170	19.047.560	587.610
Rim-RT1	8.532.494	8.277.304	255.190
Rim-RT2	18.376.983	17.879.599	497.384
Coração-RT1	33.882.595	33.128.178	754.417
Coração-RT2	21.444.643	20.940.841	503.802
Testículo-RT1	12.403.682	11.705.269	698.413
Testículo-RT2	19.183.768	18.212.300	971.468

Tabela 4.14 - Número de *reads* não utilizados pelo Cufflinks, número de *reads* selecionados e não selecionados no terceiro alinhamento para organismo *Mus musculus*.

Tecido	*Reads* não utilizados pelo Cufflinks	*Reads* não selecionados no o 3º alinhamento	Selecionados no 3º alinhamento
Cérebro-RT1	14.907.223	14.137.329	769.894
Cérebro-RT2	12.493.044	11.775.655	717.389
Fígado-RT1	18.601.976	17.696.085	905.891
Fígado-RT2	11.337.123	10.613.753	723.370
Baço-RT1	21.917.343	20.286.429	1.630.914
Baço-RT2	19.924.596	19.478.474	446.122
Pâncreas-RT1	9.669.452	8.974.083	695.369
Pâncreas-RT2	15.114.288	13.802.896	1.311.392
Tecido Adiposo-RT1	16.703.986	15.679.143	1.024.843
Tecido Adiposo-RT2	13.447.736	12.567.185	880.551
Glândula Adrenal-RT1	25.906.863	25.420.038	486.825
Glândula Adrenal-RT2	21.534.889	20.400.572	1.134.317
Ovário-RT1	18.199.748	17.478.441	721.307
Pulmão-RT1	9.722.990	9.274.887	448.103
Pulmão-RT2	9.742.796	9.279.927	462.869
Intestino Grosso-RT1	23.737.100	22.579.525	1.157.575
Intestino Grosso-RT2	14.619.413	13.615.894	1.003.519
Intestino Delgado-RT1	24.602.941	23.763.733	839.208
Intestino Delgado-RT2	14.892.742	14.015.202	877.540
Rim-RT1	23.601.932	22.961.917	640.015
Rim-RT2	12.971.245	12.357.092	614.153
Coração-RT1	25.348.443	25.063.054	285.389
Coração-RT2	16.271.648	15.809.794	461.854
Testículo-RT1	27.358.619	23.351.915	4.006.704
Testículo-RT2	29.145.762	25.977.051	3.168.711

Os *reads* que mapearam corretamente em apenas um único gene foram submetidos à montagem *de novo* com programa Trinity, para que reconstrução dos transcritos pudesse ser executada. Foi observado aproveitamento de 93% dos *reads* utilizados pelo Trinity para ambos os organismos (Tabelas 4.15 e 4.16).

Tabela 4.15 - Número de *reads* mapeados corretamente nas sequências gênicas de *Homo sapiens* e que foram selecionados para a montagem *de novo* e, número e *reads* que foram utilizados pelo Trinity.

Tecido	*Reads* mapeados corretamente e selecionados para montagem com Trinity	*Reads* utilizados na montagem pelo Trinity
Cérebro-RT1	326.639	306.739
Cérebro-RT2	529.700	504.470
Fígado-RT1	525.457	486.107
Fígado-RT2	356.108	327.932
Baço-RT1	270.958	236.041
Baço-RT2	484.549	443.201
Pâncreas-RT1	233.522	207.489
Pâncreas-RT2	334.111	304.046
Tecido Adiposo-RT1	354.187	323.293
Tecido Adiposo-RT2	664.158	627.924
Glândula Adrenal-RT1	1.114.035	1.071.314
Glândula Adrenal-RT2	799.139	758.552
Ovário-RT1	1.245.929	1.192.284
Pulmão-RT1	278.760	257.229
Pulmão-RT2	381.401	355.278
Intestino Grosso-RT1	615.768	572.620
Intestino Grosso-RT2	434.470	399.998
Intestino Delgado-RT1	1.184.865	1.152.891
Intestino Delgado-RT2	587.610	558.975
Rim-RT1	255.190	230.856
Rim-RT2	497.384	462.773
Coração-T1	754.417	711.362
Coração-T2	503.802	473.522
Testículo-T1	698.413	638.114
Testículo-T2	971.468	917.454

Tabela 4.16 - Número de *reads* mapeados corretamente nas sequências gênicas de *Mus musculus* e que foram selecionados para a montagem *de novo* e, número e *reads* que foram utilizados pelo Trinity.

Tecido	*Reads* mapeados corretamente e selecionados para montagem com Trinity	Reads utilizados na montagem pelo Trinity
Cérebro-RT1	769.894	733.794
Cérebro-RT2	717.389	673.532
Fígado-RT1	905.891	878.134
Fígado-RT2	723.370	697.427
Baço-RT1	1.630.914	1.601.750
Baço-RT2	446.122	314.154
Pâncreas-RT1	695.369	517.270
Pâncreas-RT2	1.311.392	1.201.544
Tecido Adiposo-RT1	1.024.843	987.042
Tecido Adiposo-RT2	880.551	845.096
Glândula Adrenal-RT1	486.825	454.312
Glândula Adrenal-RT2	1.134.317	1.113.056
Ovário-RT1	721.307	683.124
Pulmão-RT1	448.103	414.564
Pulmão-RT2	462.869	431.106
Intestino Grosso-RT1	1.157.575	1.123.987
Intestino Grosso-RT2	1.003.519	970.031
Intestino Delgado-RT1	839.208	808.508
Intestino Delgado-RT2	877.540	846.127
Rim-RT1	640.015	605.653
Rim-RT2	614.153	577.070
Coração-RT1	285.389	260.762
Coração-RT2	461.854	435.750
Testículo-RT1	4.006.704	3.966.406
Testículo-RT2	3.168.711	3.131.563

Para *Homo* sapiens, o número de transcritos montados pelo Trinity variou entre 2.894 e 29.316, sendo observada média aproximada de 10.000 transcritos por corrida. Para *Mus musculus*, o número de transcritos montados pelo Trinity variou entre 3.904 e 124.871, sendo observada média aproximada de 22.000 transcritos por corrida. O número de genes com pelo menos um transcrito montado variou entre 1.360 e 6.163, apresentando média aproximada de 2.700 genes por corrida para humano. Para camundongo, o número de genes com pelo menos um transcrito montado variou entre 1.537 e 8.322, apresentando média aproximada de 3.600 genes por corrida (Tabela 4.17; Figuras 4.8 a 4.11).

Tabela 4.17 - Número de transcritos montados e número de genes com pelo menos um transcrito montado em homem e camundongo.

Tecido	Número de transcritos montados		Número de genes com pelo menos um transcrito montado	
	Homem	Camundongo	Homem	Camundongo
Cérebro-RT1	4.044	26.678	1.788	4.099
Cérebro-RT2	5.276	15.789	2.339	3.580
Fígado-RT1	11.206	13.820	2.541	2.549
Fígado-RT2	7.900	10.065	2.099	2.248
Baço-RT1	6.991	45.946	2.292	4.879
Baço-RT2	9.694	4.328	2.797	2.368
Pâncreas-RT1	5.639	4.494	1.987	1.715
Pâncreas-RT2	3.313	3.903	1.555	1.537
Tecido Adiposo-RT1	7.682	26.964	2.346	4.772
Tecido Adiposo-RT2	12.717	22.381	3.184	4.342
Glândula Adrenal-RT1	13.151	12.172	3.328	2.955
Glândula Adrenal-RT2	9.820	13.467	2.964	4.351
Ovário-RT1	29.316	18.561	4.539	3.817
Pulmão-RT1	4.057	7.872	1.732	2.676
Pulmão-RT2	4.596	6.683	2.089	2.601
Intestino Grosso-RT1	8.997	34.260	2.897	4.598
Intestino Grosso-RT2	8.573	22.387	2.869	3.854
Intestino Delgado-RT1	22.807	21.853	3.511	3.607
Intestino Delgado-RT2	12.153	17.072	2.620	3.286
Rim-RT1	2.894	14.151	1.360	3.029
Rim-RT2	4.955	10.202	2.200	2.845
Coração-RT1	10.728	8.210	2.788	2.031
Coração-RT2	8.625	10.392	2.516	2.552
Testículo-RT1	19.575	124.871	5.147	8.322
Testículo-RT2	27.447	57.899	6.163	7.363

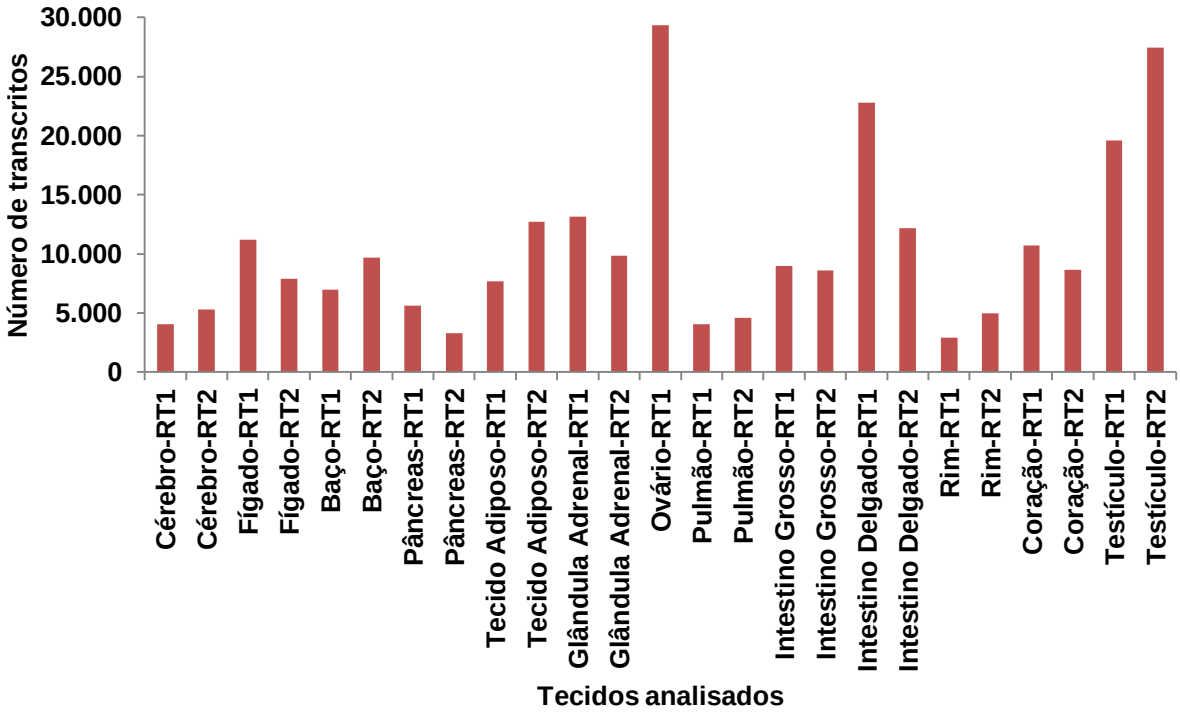

Figura 4.8 - Número de transcritos montados pelo Trinity em homem.

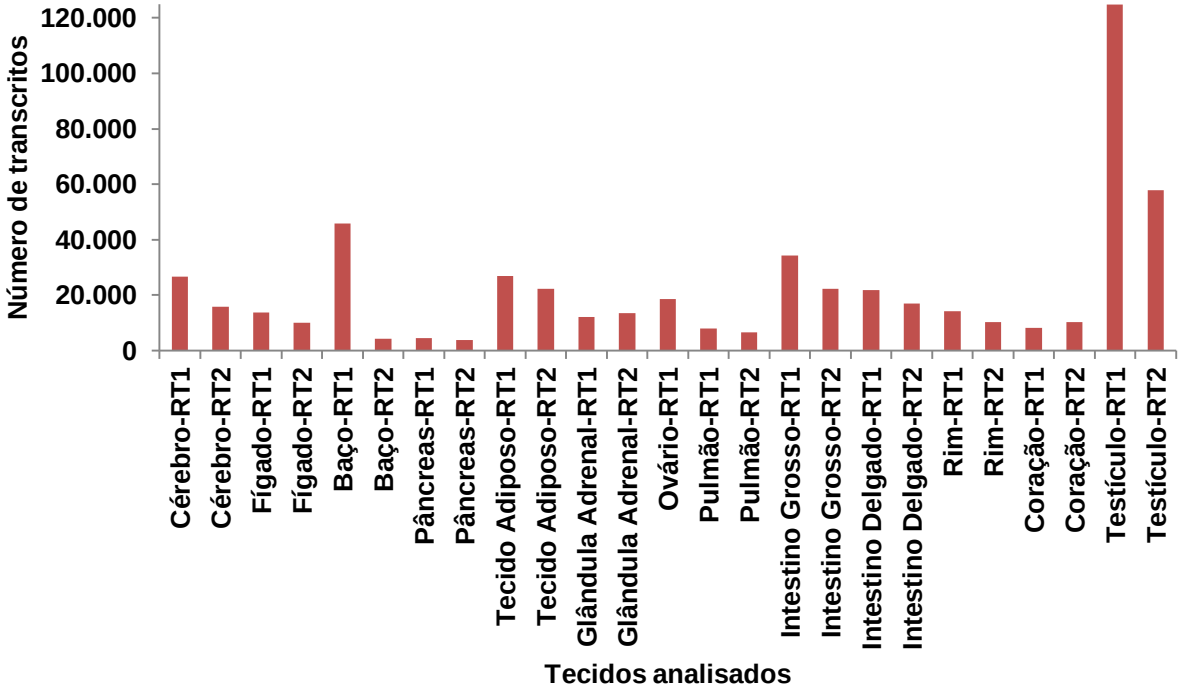

Figura 4.9 - Número de transcritos montados pelo Trinity em camundongo.

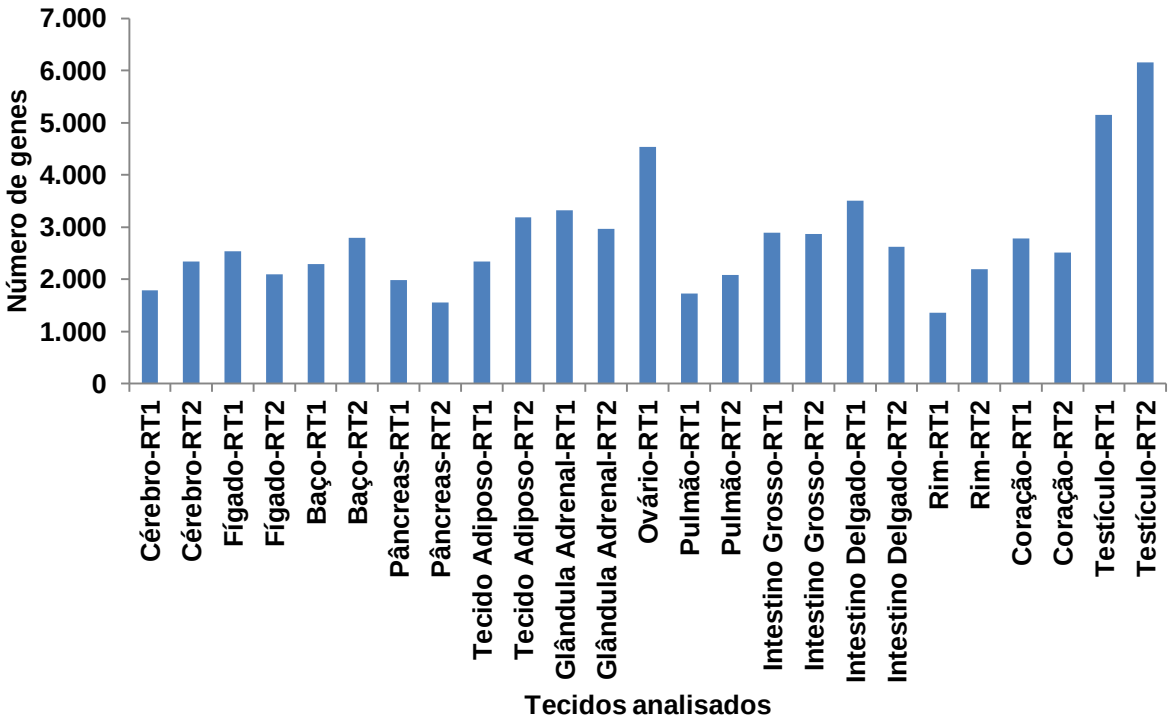

Figura 4.10 - Número de genes com pelo menos um transcrito montado pelo Trinity em homem.

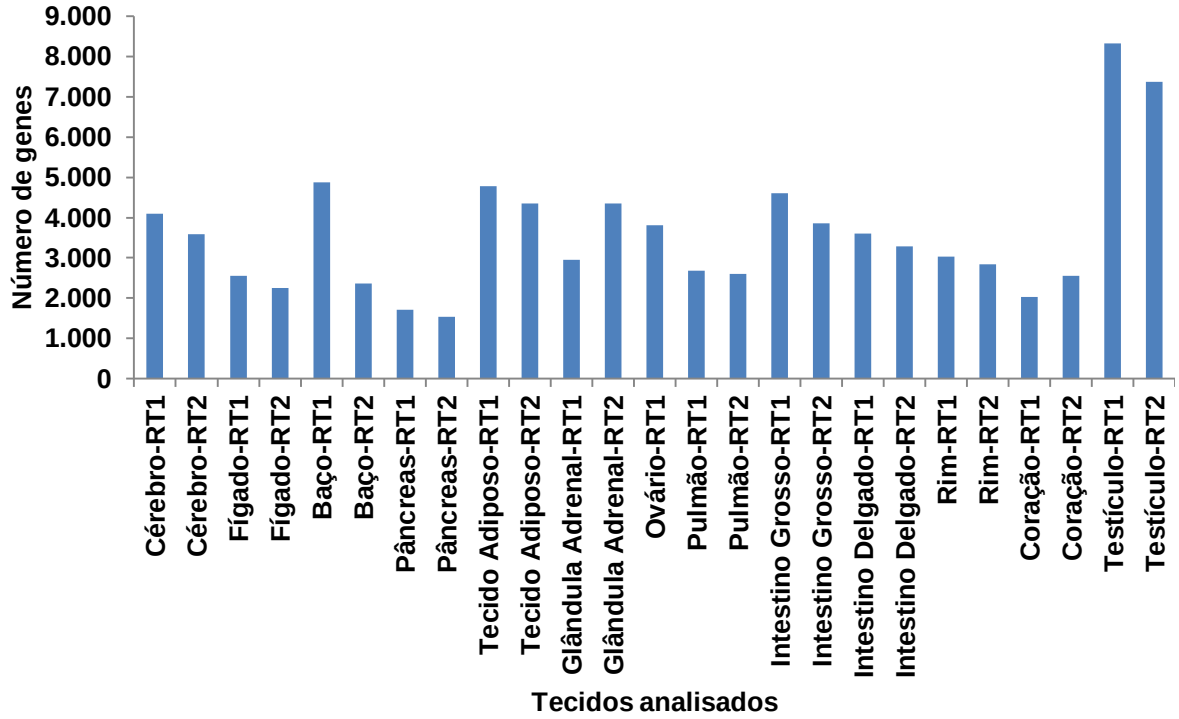

Figura 4.11 - Número de genes com pelo menos um transcrito montado pelo Trinity em camundongo.

4.3.4 - Identificação e tradução das variantes de *splicing* de camundongo e humano

Os transcritos reconstruídos pela montagem *de novo* foram incorporados às bases de dados de humano e camundongo para posterior identificação das variantes de *splicing*. Em *Homo sapiens*, a maioria das variantes de *splicing* geradas pelo programa Trinity poderiam ser confirmadas também por ESTs ou por sequências curadas (RefSeq). O número de variantes geradas exclusivamente pelo Trinity variou entre 66 (rim – primeira réplica técnica) e 969 (testículo – segunda réplica técnica) (Tabela 4.18; Figura 4.12).

Tabela 4.18 - Número de variantes de *splicing* geradas pelo Trinity e sua confirmação por demais dados de transcriptoma (ESTs e RefSeqs) em homem.

Tecido	Trinity, RefSeq e ESTs	Trinity e RefSeq	Trinity e ESTs	RefSeq e ESTs	Apenas Trinity
Cérebro-RT1	131	0	403	689	100
Cérebro-RT2	242	0	718	1.134	260
Fígado-RT1	111	0	468	611	157
Fígado-RT2	132	0	534	702	235
Baço-RT1	56	0	318	427	92
Baço-RT2	143	0	693	776	202
Pâncreas-RT1	67	0	322	420	81
Pâncreas-RT2	59	0	289	355	76
Tecido Adiposo-RT1	125	1	570	814	121
Tecido Adiposo-RT2	223	1	794	1.101	265
Glândula Adrenal-RT1	224	0	749	1.118	305
Glândula Adrenal-RT2	251	3	818	1.195	296
Ovário-RT1	177	0	672	909	234
Pulmão-RT1	100	0	397	581	113
Pulmão-RT2	155	1	600	874	197
Intestino Grosso-RT1	101	0	476	635	151
Intestino Grosso-RT2	166	0	657	929	219
Intestino Delgado-RT1	82	0	466	496	170
Intestino Delgado-RT2	143	0	637	773	303
Rim-RT1	66	0	264	456	66
Rim-RT2	155	0	531	864	154
Coração-RT1	139	0	540	798	182
Coração-RT2	165	0	656	919	213
Testículo-RT1	184	0	995	1.059	651
Testículo-RT2	267	1	1.373	1.325	969

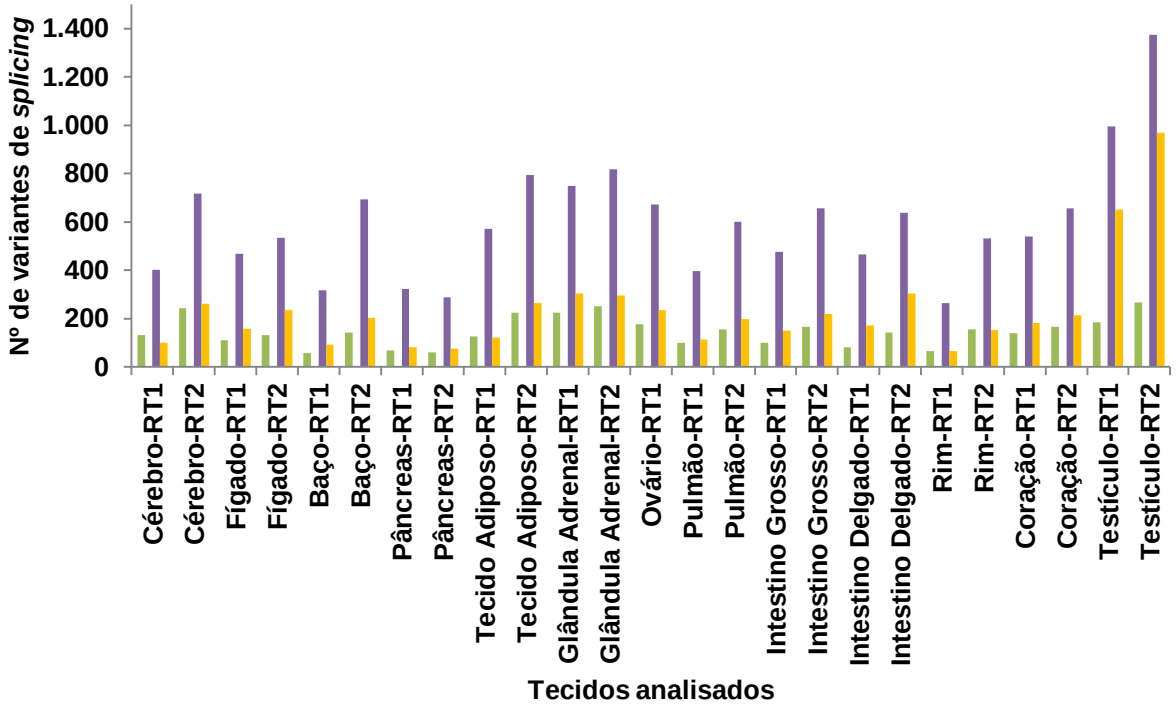

Figura 4.12 - Número de variantes de *splicing* geradas pelo Trinity confirmadas por RefSeq e ESTs (em verde), confirmadas apenas por ESTs (em roxo), confirmadas apenas por RefSeq (em vermelho) e exclusivas do Trinity (em amarelo) em humano.

Em *Mus musculus*, o mesmo comportamento foi observado e o número de variantes geradas exclusivamente pelo Trinity variou entre 35 (pâncreas – primeira réplica técnica) e 1.739 (segunda réplica técnica do tecido testicular) (Tabela 4.19; Figura 4.13).

Tabela 4.19 - Número de variantes de *splicing* geradas pelo Trinity e sua confirmação por demais dados de transcriptoma (ESTs e RefSeqs) em camundongo.

Tecido	Trinity, RefSeq e ESTs	Trinity e RefSeq	Trinity e ESTs	RefSeq e ESTs	Apenas Trinity
Cérebro-RT1	12	0	404	105	122
Cérebro-RT2	30	0	726	175	241
Fígado-RT1	11	0	351	72	136
Fígado-RT2	20	0	510	129	187
Baço-RT1	17	0	779	102	391
Baço-RT2	0	0	53	4	50
Pâncreas-RT1	4	0	214	67	35
Pâncreas-RT2	8	0	244	68	44
Tecido Adiposo-RT1	7	0	574	116	226
Tecido Adiposo-RT2	28	0	867	178	365
Glândula Adrenal-RT1	8	0	388	101	100
Glândula Adrenal-RT2	36	0	1597	189	349
Ovário-RT1	11	0	411	105	102
Pulmão-RT1	9	0	310	63	86
Pulmão-RT2	23	0	557	152	169
Intestino Grosso-RT1	10	1	463	63	134
Intestino Grosso-RT2	21	0	790	168	300
Intestino Delgado-RT1	4	0	330	71	100
Intestino Delgado-RT2	24	0	672	158	271
Rim-RT1	11	0	393	66	115
Rim-RT2	21	0	624	136	225
Coração-RT1	3	0	268	51	59
Coração-RT2	17	0	558	130	176
Testículo-RT1	16	0	915	66	1003
Testículo-RT2	40	2	2393	150	1739

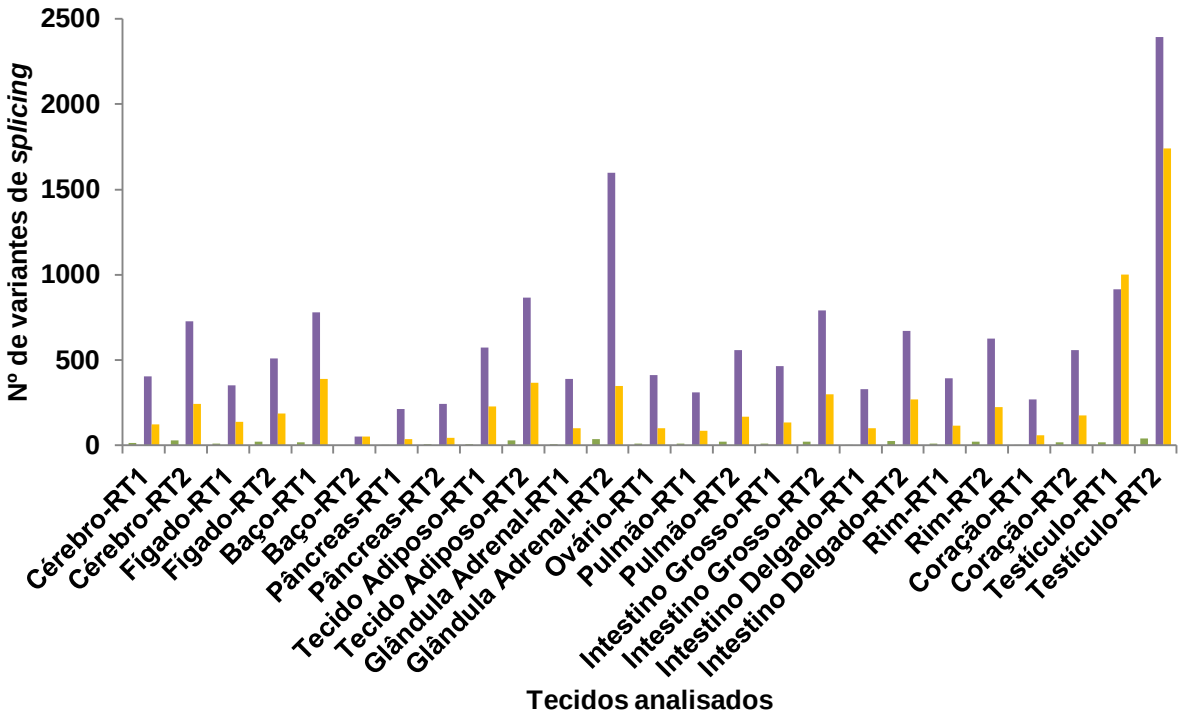

Figura 4.13. Número de variantes de *splicing* geradas pelo Trinity confirmadas por RefSeq e ESTs (em verde), confirmadas apenas por ESTs (em roxo), confirmadas apenas por RefSeq (em vermelho) e exclusivas do Trinity (em amarelo) em camundongo.

As variantes exclusivas geradas pelo Trinity foram traduzidas, variando entre 42 (pâncreas – segunda réplica técnica) e 564 (testículo – segunda réplica técnica) sequências proteicas para as corridas de *Homo sapiens*. Para *Mus musculus*, as variantes exclusivas variaram entre 9 (baço – segunda réplica técnica) e 939 (testículo – segunda réplica técnica) sequências proteicas (Tabela 4.20; Figuras 4.14 e 4.15).

Tabela 4.20 - Número de variantes exclusivas do Trinity e suas sequências proteicas para *Homo sapiens* e *Mus musculus*.

Tecido	Variantes de *splicing* exclusivas do Trinity		Sequências proteicas obtidas a partir das variantes de *splicing* exclusivas do Trinity	
	Humano	Camundongo	Humano	Camundongo
Cérebro-RT1	100	122	77	63
Cérebro-RT2	260	241	173	132
Fígado-RT1	157	136	104	58
Fígado-RT2	235	187	131	93
Baço-RT1	92	391	57	192
Baço-RT2	202	50	116	9
Pâncreas-RT1	81	35	47	18
Pâncreas-RT2	76	44	41	22
Tecido Adiposo-RT1	121	226	91	103
Tecido Adiposo-RT2	265	365	158	185
Glândula Adrenal-RT1	305	100	204	47
Glândula Adrenal-RT2	296	349	192	211
Ovário-RT1	234	102	136	58
Pulmão-RT1	113	86	74	49
Pulmão-RT2	197	169	126	95
Intestino Grosso-RT1	151	134	89	67
Intestino Grosso-RT2	219	300	124	158
Intestino Delgado-RT1	170	100	97	46
Intestino Delgado-RT2	303	271	177	142
Rim-RT1	66	115	48	60
Rim-RT2	154	225	103	123
Coração-RT1	182	59	117	26
Coração-RT2	213	176	140	95
Testículo-RT1	651	1003	375	480
Testículo-RT2	969	1739	564	939

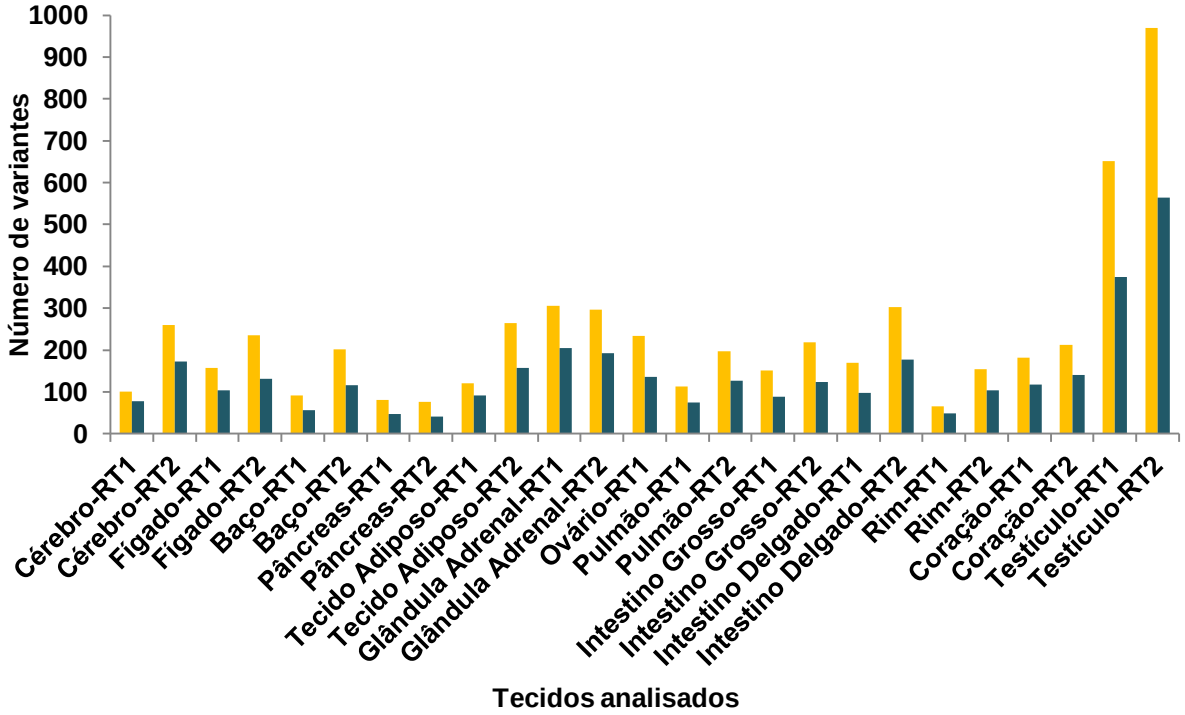

Figura 4.14 - Número de variantes exclusivas do Trinity (em amarelo) e suas sequências proteicas (em azul) para *Homo sapiens*.

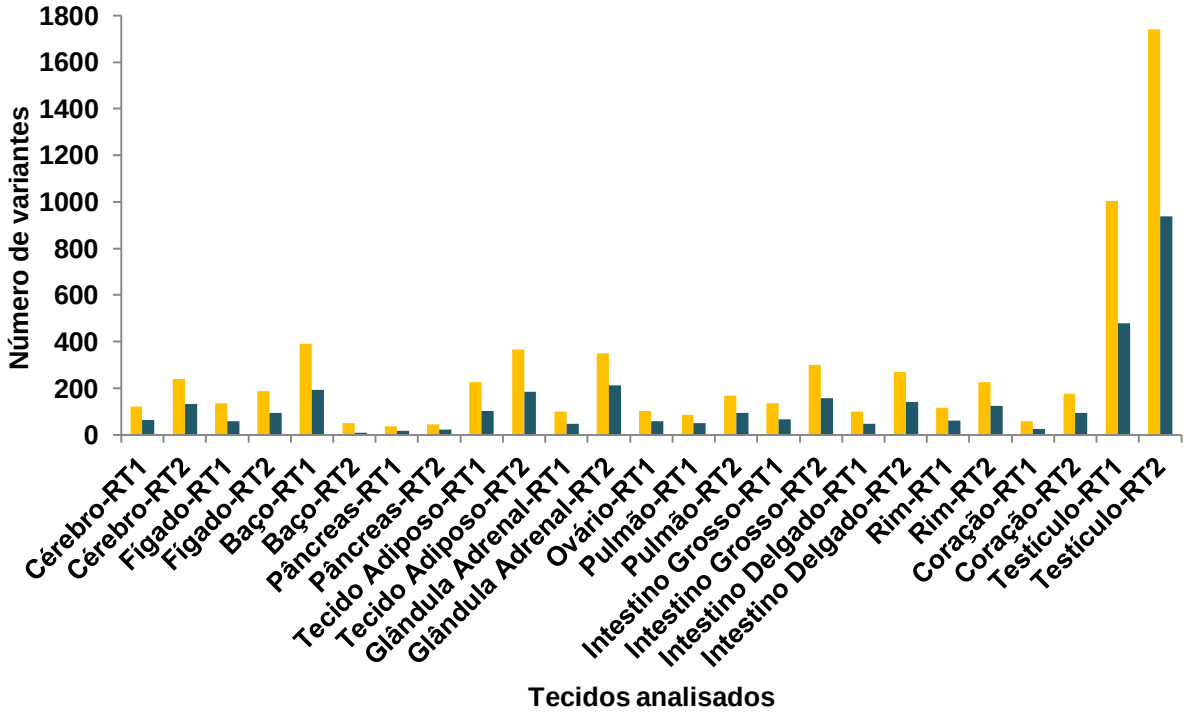

Figura 4.15 - Número de variantes exclusivas do Trinity (em amarelo) e suas sequências proteicas (em azul) para *Mus musculus*.

Ao unirmos os resultados das réplicas técnicas de cada tecido, foram contabilizadas 3.399 e 3.382 sequências proteicas não redundantes oriundas exclusivamente do programa Trinity, distribuídas em 2.654 e 2.718 genes não redundantes em homem e camundongo, respectivamente (Figura 4.16 e Figura 4.17).

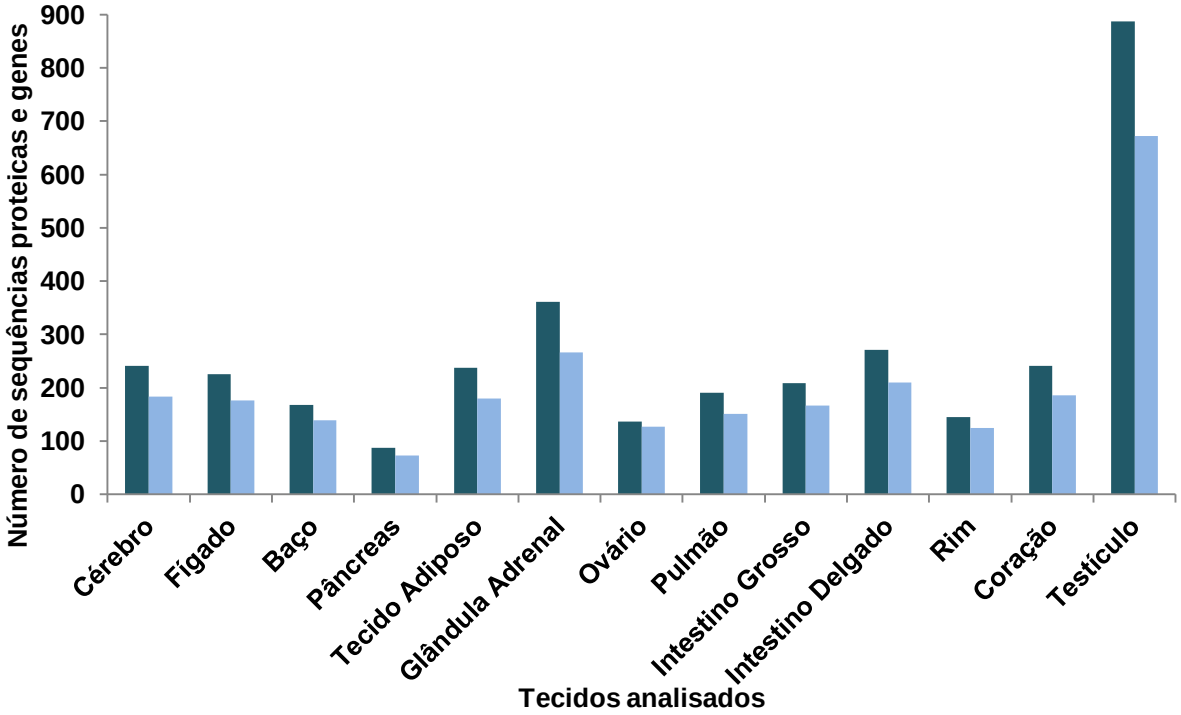

Figura 4.16 - Número de sequência proteicas (em azul escuro) e genes (em azul claro) não redundantes obtidos após a tradução das variantes de *splicing* exclusivas geradas pelo Trinity para homem.

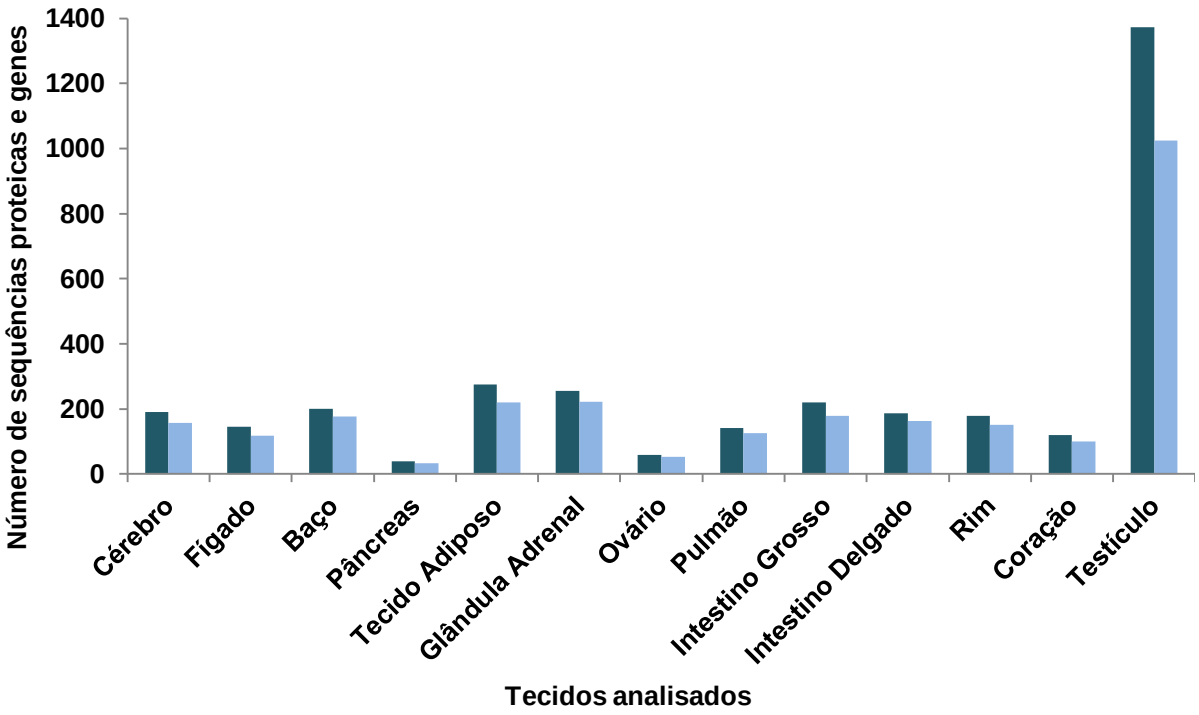

Figura 4.17 - Número de sequência proteicas (em azul escuro) e genes (em azul claro) não redundantes obtidos após a tradução das variantes de *splicing* exclusivas geradas pelo Trinity para camundongo.

4.3.5 - Aplicação de repositórios personalizados em experimentos de espectrometria de massas de tecido cerebral humano e de camundongo.

O terceiro repositório, contendo 15.447 sequências canônicas e 28.102 peptídeos digeridos *in silico* (provenientes das isoformas reconstruídas pela montagem de transcriptoma), foi aplicado para a anotação de experimentos de MS de amostras do lobo temporal anterior (ATL) e corpo caloso (CC) humano. Para a identificação das proteínas canônicas, pelo menos dois peptídeos únicos deveriam ser considerados.

Foram identificadas 1.240 proteínas canônicas em ATL, confirmadas por 8.028 peptídeos e derivadas de 1.240 genes. Em CC, foram identificadas 927 proteínas canônicas, confirmadas por 6.179 e derivadas de 927 genes. O número de isoformas identificadas em ATL foi de 136, confirmadas por 390 peptídeos, derivadas de 129 genes. Em CC, foram identificadas 109 isoformas, confirmadas por 295 peptídeos e derivadas de 102 genes (Tabela 4.21).

A razão PSM/UP calculada a partir das proteínas canônicas de cada região cerebral foi de 4,72 para ATL e 5,85 para CC. Ao aplicarmos esse critério para as isoformas, seriam consideradas 59 em ATL e 19 em CC.

Tabela 4.21 - Número de proteínas identificadas e seus respectivos números de genes e peptídeos em ATL e CC.

Região cerebral	Lobo temporal anterior			Corpo caloso		
	Genes	Proteínas	Peptídeos	Genes	Proteínas	Peptídeos
Canônicas	1.240	1.240	8.028	927	927	6.179
Isoformas	129	136	390	102	109	295
Total	1.348	1.376	8.418	1.011	1.036	6.474

Os genes em ATL indicam que 1.219 expressam somente proteínas canônicas, 108 apenas isoformas e 21 tanto proteínas canônicas quanto suas isoformas. Em CC, 909 expressam somente proteínas canônicas, 84 apenas isoformas e 18 tanto proteínas canônicas quanto suas isoformas (Tabela 4.22).

Tabela 4.22 - Perfil da expressão proteica dos genes em ATL e CC.

Expressão proteica dos genes	ATL	CC
Somente canônicas	1.219	909
Somente isoformas	108	84
Canônicas e isoformas	21	18
Total	1.348	1.011

A análise das isoformas também revelou que de um total de 163 isoformas distintas, 82 eram expressam em ATL e CC, 54 eram expressas exclusivamente em ATL e 27 eram expressas exclusivamente em CC. Quando comparadas somente as isoformas com a razão superior ao calculado para cada região, foram totalizadas 58 isoformas distintas, sendo 13 expressas em ATL e CC, 6 exclusivas de CC e 39 exclusivas de ATL. Entre as canônicas, 716 eram expressam em ATL e CC, 524 eram exclusivas de ATL e 211 eram exclusivas de CC (Tabela 4.23).

Tabela 4.23 - Distribuição das proteínas canônicas e isoformas com e sem *cut off* em ATL e CC.

Comparação	Proteínas		
ATL vs. CC	Canônicas	Isoformas (sem *cut off*)	Isoformas (com *cut off*)
Número de proteínas idênticas	716	82	13
Número de proteínas exclusivas de ATL	524	54	39
Número de proteínas exclusivas de CC	211	27	6
Total	1451	163	58

Entre as isoformas encontradas, duas foram detectadas somente pelo uso do programa Trinity. Uma delas foi encontrada somente em ATL e pertencia ao gene *KIF1A*. O peptídeo detectado no espectrômetro de massas também era compartilhado por outra isoforma (NP_004312) que tem origem em um mRNA de 46 éxons (NM_004321) (Figura 4.18).

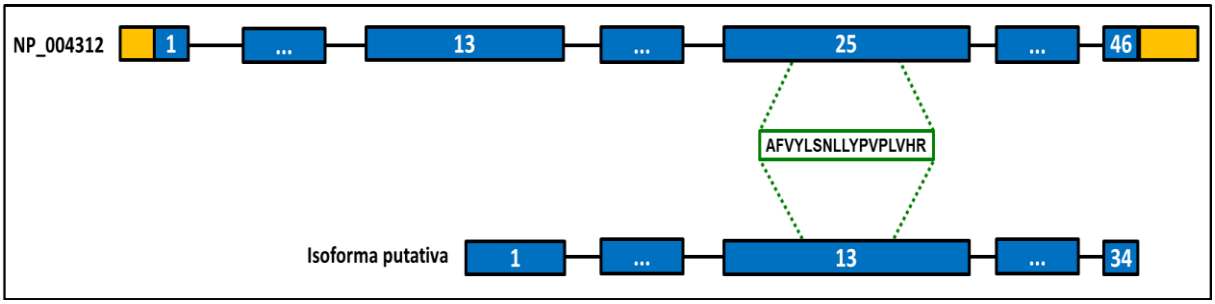

Figura 4.18 - Comparação entre a isoforma NP_004312 e a isoforma hipotética gerada pela nossa abordagem. Éxons são representados pelos retângulos em amarelo (UTRs) e em azul (sequência codificadora). As reticências simbolizam todos os éxons que intervalam os éxons numerados. O peptídeo encontrado é representado pelo retângulo verde.

Entre os genes que expressavam as proteínas canônicas e as isoformas, apenas em CC foram observados dois genes *NUDT16* e *NEBL*. A isoforma (NP_689608) do gene *NUDT16* foi confirmada por 4 peptídeos (7 PSMs) e a proteína canônica por (NP_001165377) foi confirmada por dois peptídeos (3 PSMs; Figura 4.19A). A isoforma (NP_998734) do gene *NEBL* foi confirmada por 6 peptídeos (11 PSMs) e a proteína canônica por (NP_006384) foi confirmada por 3 peptídeos (8 PSMs; Figura 4.19B).

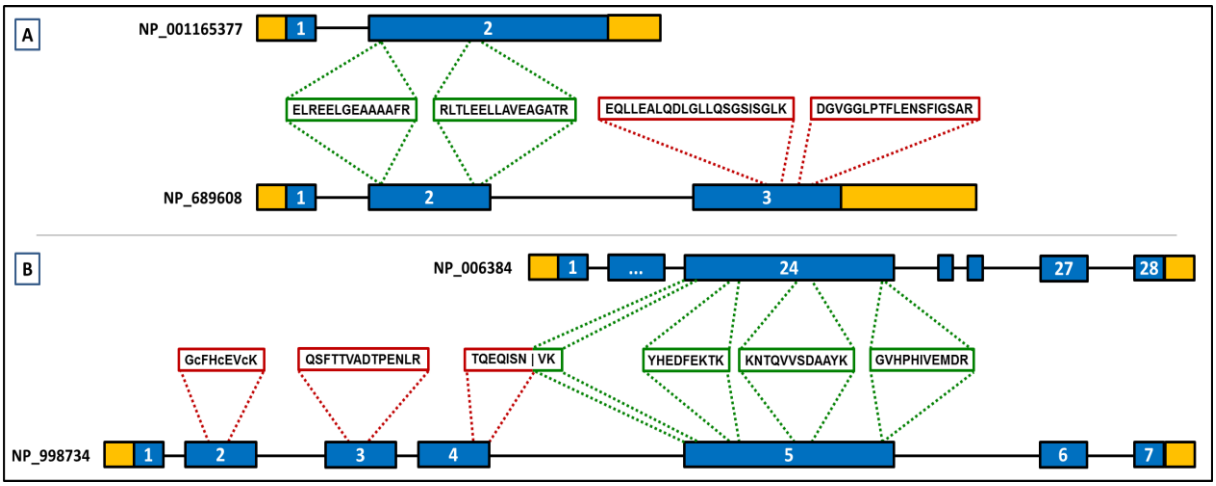

Figura 4.19 - Representação esquemática das variantes de *splicing* identificadas e seus respectivos peptídeos: (A) *NUDT16*, (B) *NEBL*. Éxons são representados pelos retângulos em amarelo (UTRs) e em azul (sequência codificadora) e íntrons são representados pelo símbolo "I". As reticências simbolizam todos os éxons que intervalam os éxons numerados. Os peptídeos únicos são representados por linhas vermelhas tracejadas e pontilhadas. Os peptídeos compartilhados são representados por linhas verdes tracejadas.

O estudo que gerou os dados de ATL e CC identificou 5 proteínas que ainda não haviam sido evidenciadas por MS. Foram encontrados dois peptídeos (“VTVAESSSDGR” com 1 PSM e “AADAVGEILLSLSYLPTAER” 3 PSMs) referentes a proteína canônica NP_808878, que pertence ao gene *SYT12*. Ao alinharmos as quatro proteínas que são expressas por este gene, observamos que os dois são compartilhados entre elas (Figura 4.20).

Figura 4.20 - Alinhamento entre a proteína canônica NP_808878 e as demais isoformas do gene *SYT12*. Os peptídeos encontrados estão destacados pelas linhas vermelhas.

O quarto repositório (camundongo) consistia em 19.518 proteínas canônicas e 9.301 peptídeos digeridos *in silico* a partir das isoformas provenientes da montagem de transcriptoma. Esse repositório foi destinado à análise de três réplicas biológicas de corpo caloso de camundongo. Para a identificação das proteínas canônicas, pelo menos dois peptídeos únicos deveriam ser considerados.

Foram identificadas 4.050 proteínas canônicas na primeira réplica biológica, 4.300 na segunda réplica biológica e 3.955 na terceira réplica biológica, confirmadas por 25.899, 35.111 e 26.671 peptídeos distintos, respectivamente. Essas proteínas eram derivadas de 4.050 genes na primeira réplica biológica, 4.300 genes na segunda réplica biológica e 3.955 genes na terceira réplica biológica (Tabela 4.24).

Foram identificadas 79 isoformas na primeira réplica biológica, 75 na segunda réplica biológica e 70 na terceira réplica biológica, confirmadas por 25.899, 35.111 e 26.671 peptídeos distintos, respectivamente. Essas isoformas eram derivadas de 75 genes na primeira réplica biológica, 73 genes na segunda réplica biológica e 67 genes na terceira réplica biológica (Tabela 4.24).

Tabela 4.24 - Número de proteínas identificadas e seus respectivos números de genes e peptídeos nas réplicas biológicas de corpo caloso de camundongo.

Réplica	Canônicas	Isoformas	Total
Réplica 1			
Genes	4.050	75	4.085
Proteínas	4.050	79	4.129
Peptídeos	25.899	210	26.109
Réplica 2			
Genes	4.300	73	4.341
Proteínas	4.300	75	4.375
Peptídeos	35.111	223	35.334
Réplica 3			
Genes	3.955	67	3.992
Proteínas	3.955	70	4.025
Peptídeos	26.671	184	26.855

Os genes na primeira réplica indicam que 4.010 expressam somente proteínas canônicas, 35 apenas isoformas e 40 tanto proteínas canônicas quanto suas isoformas. Na segunda réplica, 4.268 expressam somente proteínas canônicas, 41 apenas isoformas e 32 tanto proteínas canônicas quanto suas isoformas. Já na terceira réplica, 3.925 expressam somente proteínas canônicas, 37 apenas isoformas e 30 tanto proteínas canônicas quanto suas isoformas. (Tabela 4.25).

Tabela 4.25 - Perfil da expressão proteica dos genes em ATL e CC.

Expressão proteica dos genes	Réplica 1	Réplica 2	Réplica 3
Somente canônicas	4.010	4.268	3.925
Somente isoformas	35	41	37
Canônicas e isoformas	40	32	30
Total	4.085	4.341	3.992

Unindo as proteínas canônicas de todas as três réplicas, foram contabilizadas 5.246 proteínas canônicas distintas e 110 isoformas distintas. A distribuição entre as réplicas indica que a maior parte das proteínas canônicas e das isoformas era compartilhado entre as três réplicas (Figuras 4.21 e 4.22).

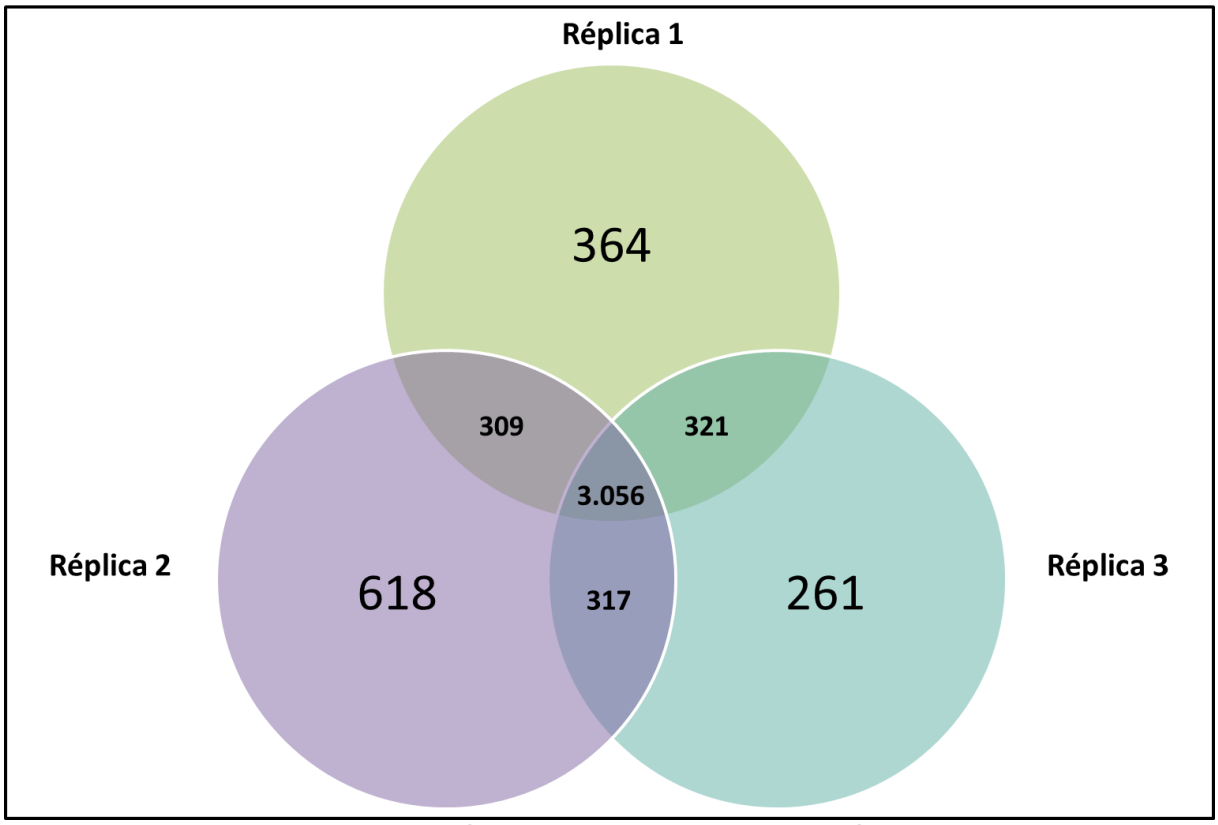

Figura 4.21 - Distribuição das proteínas canônicas entre as três réplicas.

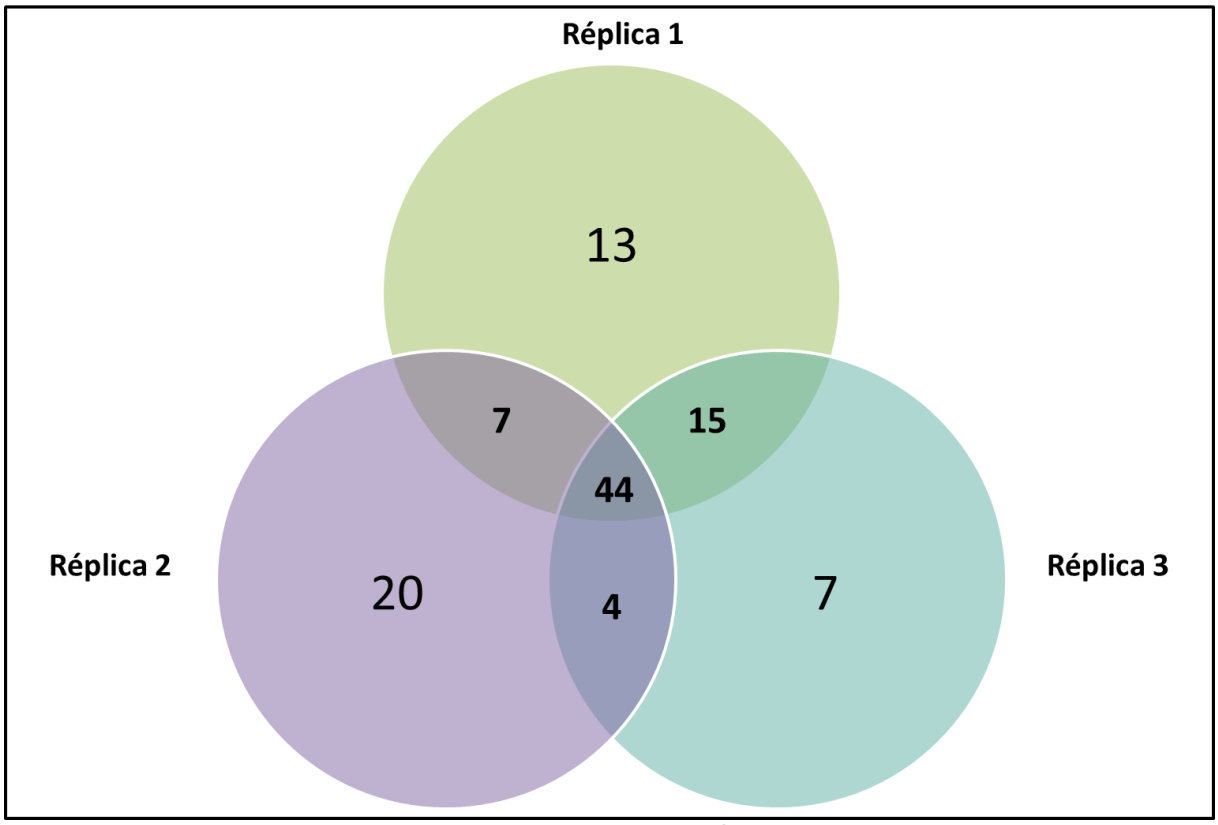

Figura 4.22 - Distribuição das isoformas entre as três réplicas.

A razão PSM/UP obtida através das proteínas canônicas foi calculada em aproximadamente 2 para as três réplicas. Quando este critério foi aplicado para a seleção das isoformas, foram selecionadas 24, 18 e 17 isoformas para as réplicas 1, 2 e 3, respectivamente. Ao uni-las, foram contabilizadas 35 isoformas distintas que, distribuídas entre as três réplicas, indica que a réplica 1 detém a maior parte das isoformas (Figura 4.23).

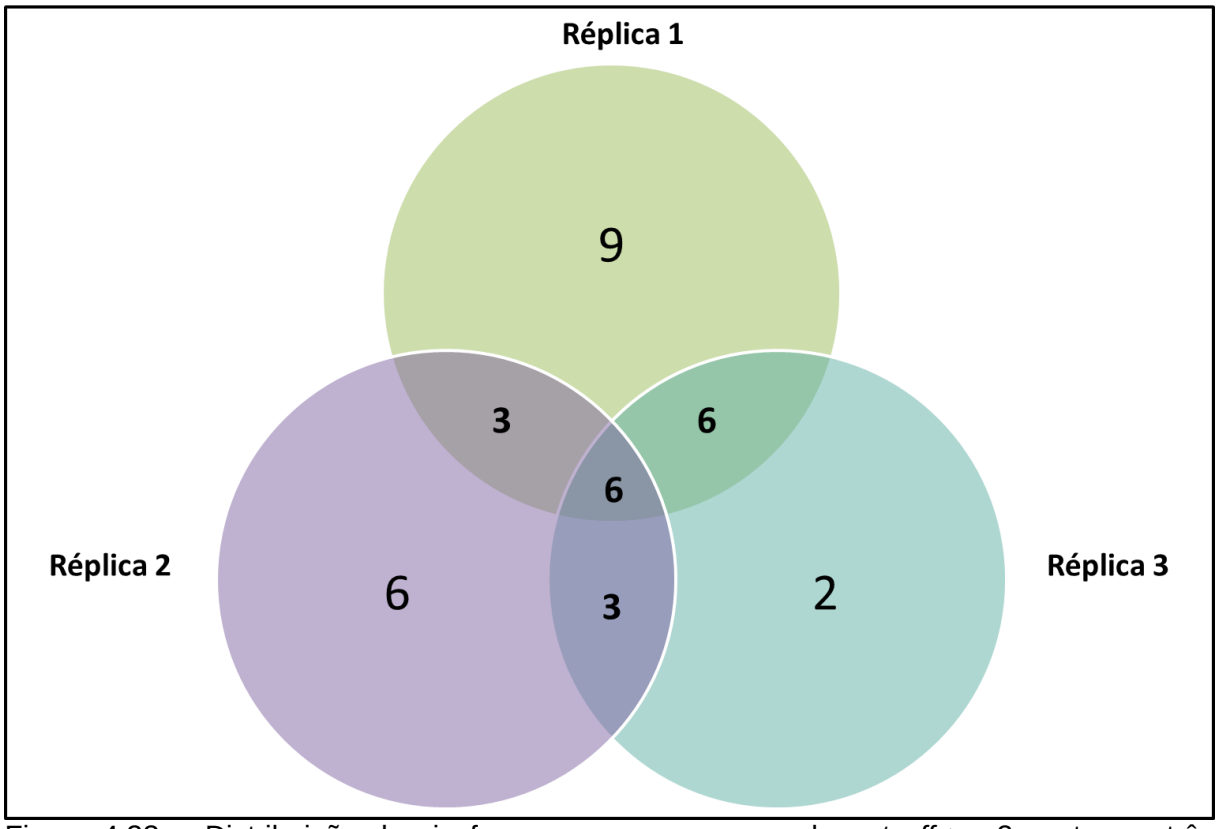

Figura 4.23 -. Distribuição das isoformas que passaram pelo *cut off* >= 2 entre as três réplicas.

Entre as isoformas encontradas, 8 eram derivadas exclusivamente do programa Trinity, sendo duas exclusivas de uma única réplica. A primeira pertence ao gene Slc4a10. O alinhamento entre a isoforma hipotética gerada e uma das isoformas curadas do mesmo gene (NP_001229310), demonstra que o peptídeo encontrado (NGQVVSPQSAPACAENK) é compartilhado entre elas e as porções inicial e final das sequências são diferentes na composição dos aminoácidos (Figura 4.24). A segunda isoforma pertence ao gene Ralgapa1 e comparando-a com uma das isoformas curadas do mesmo gene (NP_001273192) é possível observar que o peptídeo encontrado (SIGECALPSAYIR) está é compartilhado entre as duas proteínas (Figura 4.25).

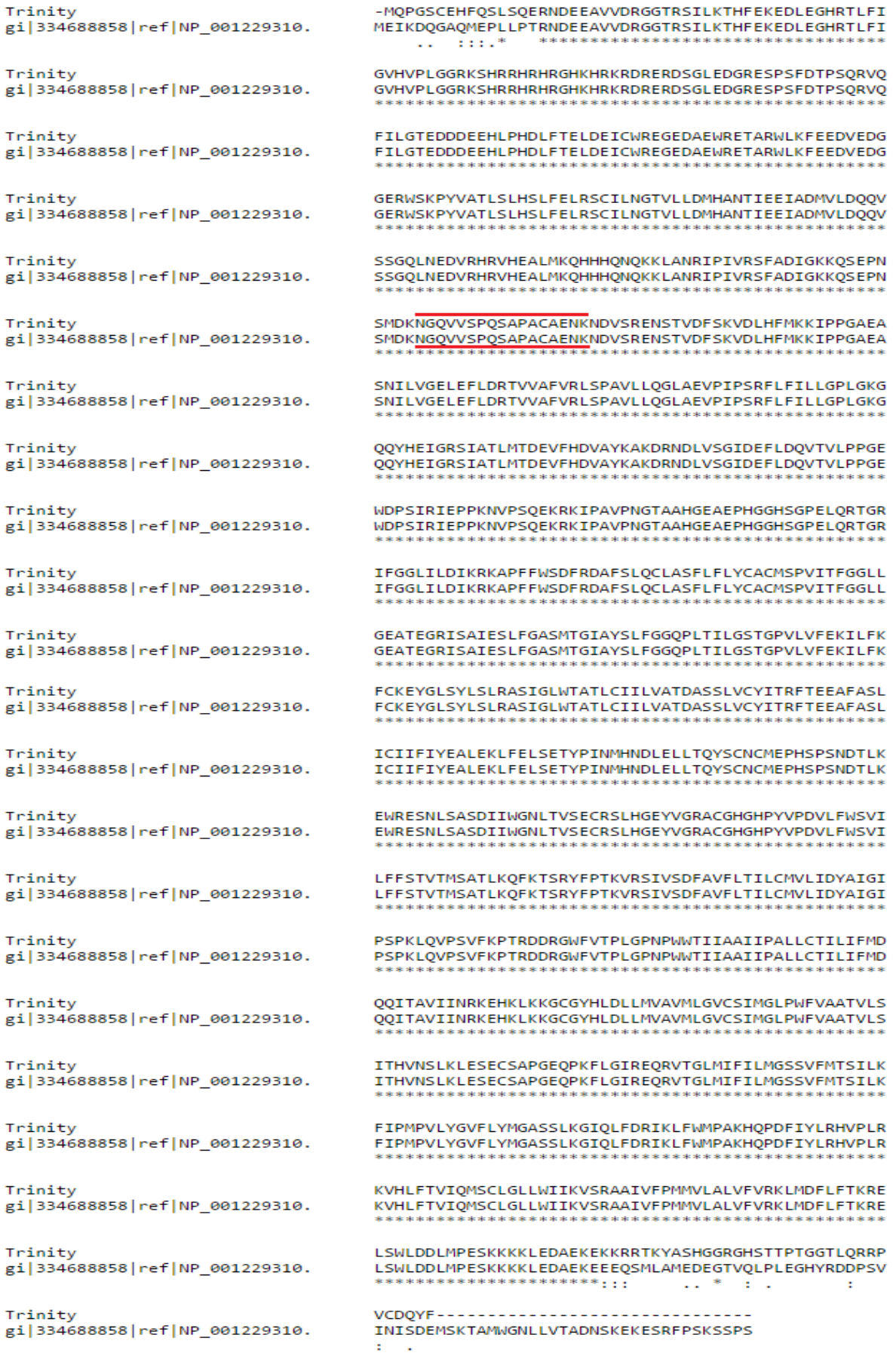

Figura 4.24 - Alinhamento entre a primeira sequência proteica gerada a partir do transcrito gerado pelo Trinity com uma isoforma curada do mesmo gene. O peptídeo encontrado está destacado pelas linhas vermelhas.

Figura 4.25 - Alinhamento parcial entre a segunda sequência proteica gerada a partir do transcrito gerado pelo Trinity com uma isoforma curada do mesmo gene. O peptídeo encontrado está destacado pelas linhas vermelhas.

Entre os genes que apresentavam a proteína canônica e a isoforma expressa, três apresentaram a contagem do número de espectros de suas isoformas maior que os espectros canônicos: Mical3, Dnm1l e Nup98. Foram identificados dez peptídeos para a isoforma (NP_001257404; 12 PSMs) e cinco para a proteína canônica (NP_700445; 5 PSMs) do gene Mical3 (Figura 4.26).

Figura 4.26 - Alinhamento entre as isoforma e a proteína canônica do gene Mical3. Os peptídeos encontrados estão destacados pelas linhas vermelhas (continua).

Figura 4.26. Alinhamento entre as isoforma e a proteína canônica do gene Mical3. Os peptídeos encontrados estão destacados pelas linhas vermelhas (continua).

Figura 4.26. Alinhamento entre as isoforma e a proteína canônica do gene Mical3. Os peptídeos encontrados estão destacados pelas linhas vermelhas (conclusão).

Já para o gene Dnm1l, um único peptídeo com 4 PSMs foi atribuído à duas isoformas (NP_001263270; NP_001263269) e dois peptídeos com 2 PSMs foram atribuídos a proteína canônica (NP_690029) (Figura 4.27A). Por sua vez, o gene Nup98 teve uma isoforma (NP_001274093) identificada por cinco peptídeos (5 PSMs) e sua proteína canônica (NP_075355) foi confirmada por três peptídeos (3 PSMs; Figura 4.27B).

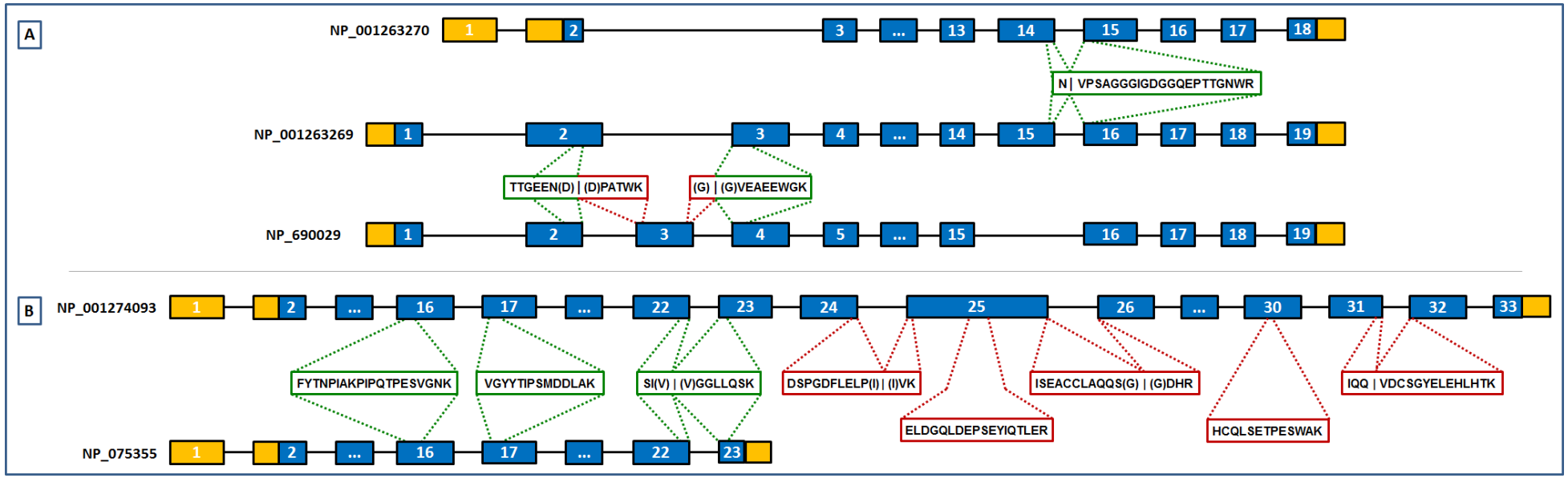

Figura 4.27 - Representação esquemática das isoformas identificadas e seus respectivos peptídeos: (A) Dnm1l, (B) Nup98. Éxons são representados pelos retângulos em amarelo (UTRs) e em azul (sequência codificadora), e íntrons são representados pelo símbolo “I”. As reticências simbolizam todos os éxons que intervalam os éxons numerados. Os peptídeos únicos são representados por linhas vermelhas tracejadas. Os peptídeos compartilhados são representados por linhas verdes tracejadas.

Ao buscarmos por genes ortólogos de homem e camundongo identificados nos experimentos de MS de corpo caloso das respectivas espécies, foram encontrados 835 genes ortólogos referentes às proteínas canônicas e 7 referentes as isoformas (Tabela 4.26).

Tabela 4.26 - Genes ortólogos das isoformas identificadas nos experimentos de MS de corpo caloso de homem e camundongo.

Homem		Camundongo	
Gene ID	*Gene Symbol*	*Gene ID*	*Gene Symbol*
998	*CDC42*	12540	Cdc42
7170	*TPM3*	59069	Tpm3
1936	*EEF1D*	66656	Eef1d
5315	*PKM*	18746	Pkm
989	*SEPT7*	235072	Sept7
6418	*SET*	56086	Set
22902	*RUFY3*	52822	Rufy3

Em humano, o gene *CDC42* teve apenas um peptídeo (NVFDEAILAALEPPETQPK) encontrado em nossa análise que pertencia a isoforma NP_426359 (Figura 4.28A). Para o gene ortólogo em camundongo (Cdc42) foi detectada a presença da isoforma NP_001230698 através dos peptídeos “YVECSALTQR” e “NVFDEAILAALEPPETQPK”. A título de comparação, a figura 4.28B monstra o alinhamento entre esta isoforma e uma segunda proteína também expressa pelo mesmo gene e a figura 4.28C compara as duas isoformas encontradas em homem e camundongo.

Figura 4.28. Isoformas dos genes *CDC42* em homem e Cdc42 em camundongo. (A) Alinhamento entre as isoformas geradas pelo gene humano. (B) Alinhamento entre as isoformas geradas pelo gene em camundongo. (C) Comparação entre a isoformas de homem (NP_426359) e camundongo (NP_001230698) identificadas nos experimentos de MS. As linhas em vermelho indicam os peptídeos encontrados.

Para o gene *SEPT7* (em humano) foram encontrados seis peptídeos atribuídos a duas isoformas (NP_001011553 e NP_001779) expressas por ele (Figura 4.29). Em camundongo, apenas um peptídeo foi encontrado (SVNCGTMAQPK) e atribuído a isoforma NP_001192296. A título de comparação, a sequência desta isoforma foi comparada com outra proteína expressa pelo mesmo gene (Figura 4.30).

Figura 4.29 - Comparação entre as duas isoformas do gene humano *SEPT7* identificadas no experimento de MS. As linhas em vermelho indicam os peptídeos encontrados.

Figura 4.30 - Alinhamento entre as duas isoformas expressas pelo gene Sept7 em camundongo. A linha em vermelho indica o peptídeo encontrado.

Para o gene *RUFY3* (humano), três peptídeos foram identificados e atribuídos a duas isoformas do gene *RUFY3* (NP_055776 e NP_00112418). Em camundongo, apenas um peptídeo (INSLQLEVEALTR) foi atribuído a isoforma NP_001276703 do gene ortólogo Rufy3 (Figura 4.31). A título de comparação, esta isoforma foi comparada com as demais do mesmo gene (Figura 4.32).

Figura 4.31 - Alinhamento entre as isoformas encontradas do gene *RUFY3* no experimento de MS de humano. As linhas em vermelho indicam os peptídeos encontrados.

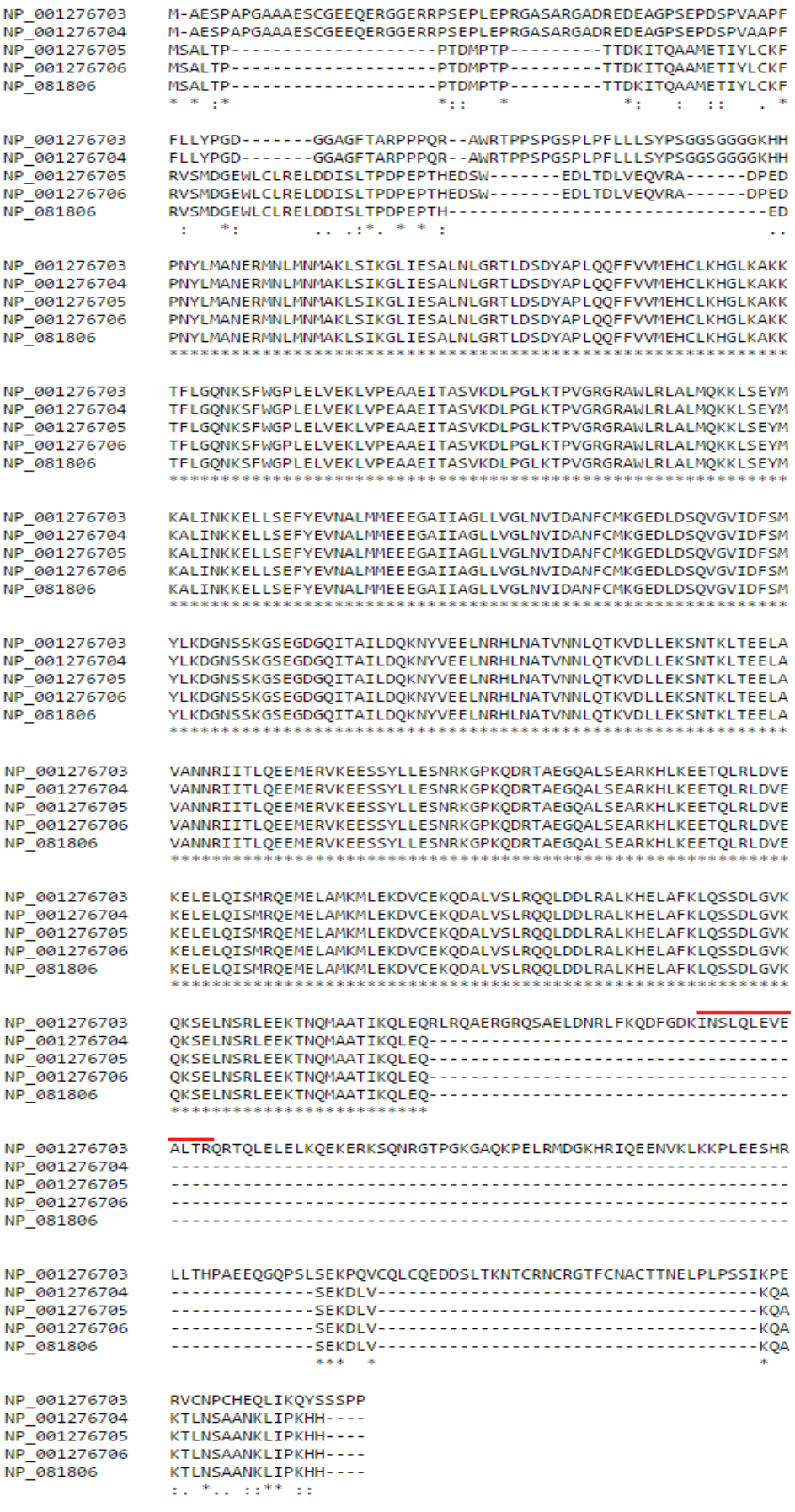

Figura 4.32 - Alinhamento entre a isoforma encontrada (NP_001276703) no experimento de MS de camundongo comparada com as demais proteínas do gene Rufy3. A linha em vermelho indica o peptídeo encontrado.

5. DISCUSSÃO

O presente trabalho apresentou o desenvolvimento de abordagens voltadas para a área da proteogenômica ao longo do doutoramento, sendo parte publicada em um trabalho original (Tavares *et al*., 2014), outro aceito para publicação (Tavares *et al*., *no prelo*) e uma revisão que considerou aspectos e possibilidades na proteogenômica (Tavares *et al*., 2015).

O primeiro repositório criado, SpliceProt, foi a primeira iniciativa do nosso grupo que reuniu a tradução *in silico* de variantes de *splicing* identificadas em dados de transcriptoma gerados em larga escala. Em comparação aos demais repositórios (RefSeq, ENSEMBL Gene e UniProtKB/Swiss-Prot) utilizados para comparação, o SpliceProt apresentou a menor porcentagem de sequências redundantes (1,36%) e o maior número de sequências únicas disponíveis.

A redundância entre os repositórios também foi avaliada, onde grande parte das sequências do RefSeq (91,98%) estava contida no repositório desenvolvido neste estudo. Importante salientar que era esperado que todas das sequências do RefSeq fossem contempladas já que este repositório foi utilizado para a confecção do SpliceProt. Entretanto, dois fatores poderiam explicar tal diferença: primeiro, as sequências dos transcritos conduzidos à tradução *in silico* são oriundas do genoma, o que poderia acarretar em pequenas modificações na sequência proteica. E, em segundo lugar, dependendo do transcrito, a tradução tem seu início a partir da primeira ou segunda metionina encontrada pelo maquinário ribossomal. De maneira semelhante, a forma como são geradas as traduções para o ENSEMBL Gene poderia explicar o baixa porcentagem (43,05%) de sequências redundantes com o SpliceProt. Todavia, não foram encontradas informações na literatura disponível que pudessem elucidar o processo de tradução e compará-lo com o desenvolvido no presente estudo. As 3.535 sequências do UniProtKB/Swiss-Prot que não foram contempladas pelo SpliceProt poderia ser explicada pela composição dos repositórios. Enquanto o primeiro é composto por sequências com validação experimental, o segundo contém sequências preditas computacionalmente.

A digestão computacional por tripsina demonstrou que SpliceProt gera a maior quantidade de peptídeos não redundantes (700.492), seguido pelo ENSEMBL Gene (578.139), UniProtKB/Swiss-Prot (506.318) e RefSeq (486.532). Em comparação com RefSeq, 219.614 eram exclusivos do SpliceProt e 480.878 eram compartilhados entre eles. Em comparação com o ENSEMBL Gene, 64.461

peptídeos eram exclusivos deste repositório, 186.814 eram exclusivos do SpliceProt e 513.678 eram compartilhados entre eles. Desta forma, o repositório criado pela presente abordagem oferta 2,8 vezes mais peptídeos únicos digeridos por tripsina que o ENSEMBL Gene. Em comparação com o UniProtKB/Swiss-Prot, 24.552 peptídeos eram exclusivos deste repositório, 218.726 eram exclusivos do SpliceProt e 481.766 eram compartilhado entre eles. Assim, o repositório criado oferta 8,9 vezes mais peptídeos únicos digeridos por tripsina que o UniProtKB/Swiss-Prot. De uma forma geral, independente da enzima utilizada, o SpliceProt ofertou a maior quantidade de peptídeos únicos que poderiam ser utilizados para a identificação de potenciais variantes de *splicing*.

Observada a grande quantidade de peptídeos únicos provenientes do SpliceProt e inspirado nos resultados de Sheykman e colaboradores (2013), três repositórios personalizados foram criados. Cada um deles tinha como base as sequências não redundantes das proteínas dos repositórios RefSeq, ENSEMBL Gene e UniProtKB/Swiss-Prot, acrescidos dos peptídeos não redundantes das sequências do SpliceProt que foram digeridos computacionalmente por tripsina e as endoproteinases *Lys-C*, *Glu-C bicarb* e *Glu-C phosph* A comparação entre esses peptídeos acrescentados aos repositórios-base indicou que a maior parte deles eram compartilhados entre os repositórios personalizados independente da enzima utilizada. Ademais, os repositórios RefSeq e UniProtKB/Swiss-Prot apresentaram uma maior quantidade de peptídeos exclusivos em comparação ao ENSEMBL Gene.

O repositório personalizado UniProtKB/Swiss-Prot/SpliceProt foi utilizado em dados de espectrometria de massas obtidos a partir da linhagem de células T (*Jurkat cells*) com o intuito de identificarmos potenciais variantes de *splicing*. Os 54 peptídeos identificados são provenientes daqueles oriundos do SpliceProt acrescidos ao repositório UniProtKB/Swiss-Prot. Além disso, esses mesmos peptídeos não foram encontrados no repositório UniProtKB/TrEMBL, que possui a tradução de sequências que ainda não foram anotadas ou com validação experimental. Desta forma, a identificação das 54 variantes foi possível através da inserção desses dados e da maneira como este repositório foi construído.

Das 57 variantes identificadas pelo trabalho de Sheykman e colaboradores (2013), apenas 10 foram identificadas também pelo presente trabalho. Uma das possíveis explicações para esta diferença deve-se ao fato dos dados de transcriptoma e proteoma terem sido gerados a partir da mesma origem (linhagem

de linfócitos T), ao passo que o SpliceProt têm origem em diferentes conjuntos de dados (como ESTs, dados de sequenciamento de larga escala e sequências de referência). Ao final desta primeira análise, é possível afirmar que o primeiro repositório personalizado obteve êxito na identificação variantes de *splicing* em dados de MS e que a utilização de repositórios personalizados mostrou ser uma alternativa em relação aos repositórios tradicionalmente empregados para a anotação de dados de MS.

Apesar da primeira inciativa de construir um repositório personalizado, não houve a oportunidade de analisarmos detalhadamente as variantes de *splicing* encontradas no contexto da amostra estudada. Assim, um segundo repositório personalizado foi construído e usado para anotar dados de MS da linhagem de oligodendrócitos humanos. O repositório em questão foi construído utilizando como base as sequências das proteínas canônicas do UniProtKB/Swiss-Prot, acrescido de peptídeos não redundantes digeridos computacionalmente das sequências das isoformas do UniProtKB/Swiss-Prot e do SpliceProt. Importante salientar que os peptídeos inseridos no repositório não estavam presentes em quaisquer das sequências canônicas, pois desta forma, a identificação das proteínas canônicas e das isoformas seria facilitado após a análise.

As 2.081 proteínas canônicas identificadas apresentaram uma razão de 1,4 PSMs/Peptídeos únicos que, quando aplicada como critério para seleção de isoformas, foi possível selecionar 17 das 39 isoformas encontradas. O cálculo entre os espectros da proteína canônica e da isoforma do gene *SDR39U1* também foi uma proposta com o objetivo de sugerir que a proteína alternativa estaria sendo mais expressa que a canônica.

Além da utilização da espectrometria de massas para a identificação das isoformas no protema, confirmamos a presença dos mRNAs das variantes de *splicing* de *EEF1D*, *KRAS*, *MFF*, *SDR39U1* e *SUGT1* por experimentos de RT-qPCR na linhagem celular de oligodendrócitos MO3.13 (experimentos realizados pela Dra Patricia Savio de Araujo Souza). Desta maneira, foi possível confirmar algumas das variantes que identificamos no proteoma também no transcriptoma, faltando apenas a investigação de todas a isoformas e sua possível atuação dentro do contexto do tecido cerebral. A seguir, apresentaremos uma descrição das informações funcionais de algumas das variantes de *splicing* que identificamos nos dados de proteômica da linhagem MO3.13.

O gene *KRAS* têm um papel regulador na proliferação celular. O domínio C-terminal da variante de *splicing* KRAS4B é responsável por se ancorar à membrana plasmática e ativar a proteína (Welman *et al*., 2000). Como demonstrado na figura 4.3A, a KRAS4B (Uniprot P01116-2) difere da proteína canônica (Uniprot P01116-1) no domínio C-terminal pela presença do éxon codificador 4A (McGrath *et al*., 1983). O peptídeo encontrado em nosso estudo mapeia exatamente na junção de *splice* do éxon 3 e o éxon 4B, sendo exclusivo da variante KRAS4B. Outros dois peptídeos detectados pela nossa abordagem também foram atribuídos ao gene *HRAS*, um parálogo do *KRAS*, apresentando uma identidade maior que 80% entre as suas sequências. Assim, é importante destacar que devido a esta alta identidade, estes mesmos peptídeos também mapeiam nos éxons 1 e 2 do gene *KRAS*, que são responsáveis pela formação do região N-termnal dessas proteínas.

A glutaminase (*GLS*) catalisa a deaminação hidrolítica da glutamina em glutamato e amônia e tem papel na regulação do neurotransmissor glutamato no cérebro (Márquez *et al*., 2006). Até o momento, três variantes deste gene são conhecidas (Elgadi *et al*., 2014) e o presente trabalho identificou duas delas. O gene *GLS* possui 19 éxons codificadores e três transcritos: a proteína canônica KGA (Uniprot O94925-1) e as variantes GAM (Uniprot O94925-2) e GAC (Uniprot O94925-3). A figura 4.3B demonstra a distribuição dos peptídeos identificados de acordo com suas respectivas proteínas. As variantes KGA e GAC já foram detectadas em cérebro e outros tecidos e a alta expressão de GAC foi observada em tumores cerebrais, incluindo oligodendrogliomas, astrocitomas, ganglioglioma e ependimomas (Szeliga *et al*., 2008).

Os eventos de fissão mitocondrial são regulados por diversos elementos e mecanismos (Okamoto *et al*., 2005; Chan *et al*., 2006) que levam à fragmentação da mitocôndria. O gene MFF (do Inglês, mitochondrial fission factor) é um componente utilizado na fissão constitutiva e induzida da mitocôndria e peroxissomos. As nove variantes codificadas por este gene diferem na presença ou ausência do éxon 1 e nas combinações entre os éxons 5, 6 e 7, que são responsáveis por codificar a região central da proteína (Gandre-Babbe *et al*., 2008). Nossa análise detectou a isoforma Q9GZY8-2 (que não apresenta o éxon 7) e a proteína canônica Q9GZY8-2 (Figura 4.3C).

O gene humano *SUGT1* está associado ao complexo de proteínas do cinetócoro e está envolvida na fixação do centrossomo durante o ciclo celular. Este gene codifica a proteína canônica (SUGT1A; Uniprot Q9Y2Z0) e a isoforma

(SUGT1B; Uniprot Q9Y2Z0-2), descritos previamente por Niikura e Kitagawa (2003). As sequências de ambas as proteínas são 91% idênticas e sua diferença encontra-se na codificação de 33 aminoácidos a mais da região entre os éxons 5 e 6 da variante SGT1B e na ausência de uma serina na posição 110 da proteína SGT1A. Esta inserção acarreta no motivo de repetições de tetratricopeptídeos, quem modula a interação específica de proteína-proteína. Nossa análise indicou a presença de ambas as proteínas, como demonstrado na figura 4.3D.

Os genes da família das ribonucleoproteínas heterogênas (hnRNP) atuam no metabolismo de mRNAs e foram previamente descritos em oligodendrócitos (Iwata *et al*., 2013). Elevados níveis dos membros de HRNPAB e HNRNPC podem levar a superexpressão do gene MYC, que atual como fator de transcrição no câncer e é um marcador molecular nos casos de meduloblastoma (Staal *et al*., 2015). O gene HNRNPC está associado ao transporte e processamento do pré-mRNA (Park *et al*., 2012) e codifica duas proteínas distintas que foram identificadas pela nossa análise: C1 com 293 aminoácidos (responsável por ligar-se ao mRNA de *TP53* durante a apoptose; Christian *et al*., 2008) e C2 com 306 aminoácidos (Burd *et al*., 1989). A interação física dessas duas proteínas parece estar relacionada com o controle da agressividade de gliobastomas através da baixa expressão do gene PDCD4 (Park *et al*., 2012). As hnRNPs também estão associadas à esquizofrenia (Martins-de-Souza *et al*., 2009) e podem ter um papel central na disfunção de oligindendrócitos nesta doença (Iwata *et al*., 2011).

A proteína canônica e a variante por *splicing* alternativo do gene piruvato quinase (*PKM*) também foram detectadas no proteoma de oligodendrócitos. O gene *PKM* possui dois éxons (9 e 10) mutuamente exclusivos, produzidos respectivamente por dois transcritos: M1 (alternativo) e M2 (canônico). A proteína M2 é superexpressa em células proliferativas e tumorais (Staal *et al*., 2015), quando comparado com a M1 (Christofk *et al*., 2008) e sua expressão é regulada pela hnRNP A1/A2 e pela proteína de ligação ao trato de polipirimidinas (PTB; Clower *et al*., 2008).

A variante canônica e uma variante de *splicing* do gene *EEF1D*, foi detectada nas nossas análises. Este gene está localizado no cromossomo humano 8 e codifica quatro proteínas no cérebro humano (Cao *et al*., 2014) classificadas a partir do seu tamanho: a isoforma maior conhecida como eEF1BδL (647 aminoácidos; Uniprot P29692-2) e mais três proteínas menores. Estas proteínas alternativas são nomeadas como eEF1Bδ (proteína canônica com 281 aminoácidos; Uniprot P29692-

1), isoforma 3 (257 aminácidos; Uniprot P29692-3) e a isoforma 4 (262 aminoácidos; Uniprot P29692-4). As proteínas eEF1Bδ and eEF1BδL têm funções e localização celular distintas. A primeira é geralmente localizada no citoplasma e possui papel importante como fator de alongamento ao se ligar com outros membros do complexo eEF1 (eEF1Bα e eEF1Bγ). A segunda está localizada no núcleo e induz a transcrição dos genes *HSPA6*, *CRYAB*, *DNAJB1*, e *HMOX1* (Kaitsuka *et al*., 2015). Bartoli e colaboradores (2006) demonstraram que o surgimento do meduloblastoma está associado com a superexpressão de EEF1D, sugerindo seu potencial como biomarcador.

O gene *TOR1AIP1* codifica o polipeptídeo associado à lâmina 1 (LAP1), que é uma proteína interna da membrana nuclear (Senior e Gerace, 1988). Existem três proteínas potencialmente expressas por esse gene: a proteína canônica Q5JTV8-1, a isoforma LAP1B e isoforma LAP1C, esta última descoberta recentemente por Santos e colaboradores (2014). As proteínas LAP1 e LAP1B diferem na inserção do trinucleotídeo CAG no sítio de *splice* 3' do íntron 2 de *LAP1B*, resultando em uma alanina adicional na sequência codificadora. Santos e colaboradores (2014) identificaram a isoforma LAP1C numa linhagem derivada do tumor de neuroblastoma usando bioinformática, espectrometria de massas e técnicas de biologia molecular para validação. Na presente análise foram identificadas apenas as proteínas LAP1 e LAP1B.

O gene *PDLIM3* está associado à distrofia miotônica (DM) e codifica três proteínas identificadas pela nossa análise e que variam de acordo com o uso dos éxon 4, 5 e 6 (Uniprot Q53GG5-1, Q53GG5-2, e Q53GG5-3). A associação foi confirmada por Ohsawa e colaboradores (2011), que demonstrou a expressão predominante da variante Q53GG5-2 em pacientes com essa doença. A isoforma mencionada possui o éxon 4 e a ausência dos éxons 5 e 6, que podem influenciar na ligação da proteína com a α-actinina 2, já que a interação física entre elas ocorre através do domínio PDZ, localizado no éxon 6.

A análise de enriquecimento de vias (GSEA) dos 38 genes das isoformas identificadas na linhagem celular MO3.13 indicou uma participação relevante de tropomiosinas através do termo *muscle thin filament tropomyosin* (GO:0005862). De fato, três genes de da família de tropomiosinas foram encontrados em nossa análise: TPM1, TPM2, e TPM3. As tropomiosinas são componentes dos filamentos de actina encontrados nos dendritos de células neuronais e em células musculares (Santos *et al*., 2014). Em humano, elas são codificadas por quatro genes conhecidos como

TPM1 (localizado no cromossomo 15), *TPM2* (localizado no cromossomo 9), *TPM3* (localizado no cromossomo 1), e *TPM4* (localizado no cromossomo 19). Existem dez diferentes isoformas do gene humano *TPM1*, três isoformas do gene *TPM2* e sete isoformas do gene *TPM3*. No cérebro humano, algumas isoformas já foram descritas, sendo elas: isoformas 3 e 5 da *TPM1*, isoforma 1 de *TPM2* e isoformas 4 de *TPM3* (Lin *et al*., 2008). Nosso estudo identificou as isoformas 2 e 5 do gene *TPM1*, isoformas 2 e 3 de *TPM2*, e as isoformas 2, 3, 4, 5 e 6 de *TPM3*. É importante destacar que apesar de termos identificados uma família de tropomiosinas, a linhagem de oligodendrócitos MO3.13 é resultante de oligodendrócitos com células musculares de rabdomiossarcoma. Desta forma, a identificação dessas isoformas foi questionada se poderiam ser resultado dessa característica intrínseca da linhagem utilizada e somente com dados experimentais de proteômica de cérebro humano poderíamos confirmar a presença destas variantes.

Fizemos também a análise da interação proteína-proteína das 39 variantes de *splicing*. Encontramos a interação de algumas delas com as proteínas do complexo NFκB, MYC e TP53. Campos-Sandoval e colaboradores (2015) relacionaram o gene da glutaminase GLS e suas variantes de *splicing* na regulação das isoformas de NFκB, MYC e TP53 em tumores cerebrais. Demais trabalhos apontam que o gene TP53 é responsável por manter a estabilidade genômica das células percussoras de oligondendrócitos (OPC, do Inglês oligodendrocyte precursor cells; Tokumoto *et al*., 2001). Os genes MYC e NFκB são responsáveis por outro grupo de genes associados a mielinização dessas células, desempenhando importante função durante sua diferenciação (Blank e Prinz *et al*., 2014).

Como exposto, a análise feita com o segundo repositório personalizado possibilitou a investigação das isoformas na linhagem de oligodendrócitos. Entretanto, era necessário criar uma estratégia para a busca por potenciais variantes de *splicing* ainda não descritas na literatura para homem e camundongo. A montagem computacional de transcriptoma foi utilizada como recurso para a identificação de proteínas já anotadas e potenciais variantes de *splicing* no transcriptoma de homem e camundongo. Os programas Cufflinks e Trinity foram escolhidos por apresentarem os melhores resultados em artigos que compararam o desempenho dos programas de reconstrução de transcritos com um genoma de referência e *de novo*. Desta forma, um *pipeline* integrando as duas estratégias foi desenvolvido em três etapas que tinham como pontos principais a seleção dos *reads*

utilizados na montagem com genoma de referência e daqueles entregues para a montagem *de novo*.

A primeira etapa consistia no tratamento dos *reads* originais a partir da qualidade das corridas e a remoção de seus adaptadores. O aproveitamento dos *reads* foi considerado satisfatório, pois aproximadamente 90% e 95% dos *reads* nas corridas de humano e de camundongo, respectivamente, foram aproveitadas após o tratamento. Em seguida, todos os *reads* selecionados pela etapa anterior foram alinhados contra o genoma de seus respectivos organismos, gerando um arquivo de alinhamento. Cada arquivo foi preparado e juntamente com um arquivo GTF de cada espécie foi utilizado pelo programa Cufflinks para a montagem com genoma de referência. Dos 39.006 mRNAs anotados no arquivo GTF de humano, o Cufflinks foi capaz de reconstruir em média 30 mil transcritos independente do seu nível de expressão. Em camundongo, dos 29.735 mRNAs anotados em seu arquivo GTF, foram reconstruídos em média 22 mil transcritos independente do seu nível de expressão para este organismo. Porém, é importante salientar que a reconstrução desses transcritos está limitada ao número de transcritos com anotação. Logo, dependendo da origem ou versão do arquivo GTF utilizado, esses números poderão variar conforme novos transcritos sejam anotados nas bases de dados como RefSeq e UCSC. Desta forma, optamos por usar a base de dados RefSeq para a associação de transcritos e proteínas, uma vez que o uso do repositório Uniprot/Swiss-Prot aumentaria o grau de complexidade das análises, pois não encontramos dados que permitissem a associação dos seus identificadores com sequências de mRNAs.

O número de genes identificados pela montagem obtida pelo programa Cufflinks foi de aproximadamente 15 mil genes nas duas espécies. Uma fração pequena de transcritos oriundos de RNA-Seq não teve seu gene encontrado em nossas bases de dados. Este problema tem origem provável na anotação dos arquivos obtidos através da base de dados do RefSeq e Unigene.

A segunda etapa consistia numa preparação para a identificação dos *reads* utilizados e não utilizados pelo Cufflinks já que este programa não habilita esta opção. Assim, iniciamos uma colaboração com o Dr. Cole Trapnell, desenvolvedor do programa Cufflinks, para criar uma abordagem para a identificação e seleção dos *reads* usados pelo referido programa. Desta forma, os *reads* da primeira etapa foram alinhados contra as sequências dos transcritos reconstruídos pelo programa gffread (integrante do programa Cufflinks). Aqueles *reads* corretamente mapeados foram considerados como utilizados para a montagem e os demais foram conduzidos para

a terceira etapa do *pipeline*. Aproximadamente 70% dos *reads* dos dados de RNA-Seq de humano e 60% de camundongo não foram utilizados para a reconstrução dos transcritos.

A terceira etapa teve como proposta inicial um refinamento na seleção dos *reads* que seriam conduzidos para a montagem *de novo* com o programa Trinity. Para isso, os *reads* que não foram utilizados pelo Cufflinks foram alinhados contra as sequências gênicas de seus respectivos organismos. Desta forma, este alinhamento serviria como "filtro" para evitarmos a reconstrução de transcritos quiméricos. Os resultados apontaram que uma pequena porcentagem dos *reads* (3% em homem e 5% em camundongo) foi mapeada corretamente em um único gene. Sabendo que é necessária uma grande quantidade de *reads* para a montagem *de novo* (revisto por Martin e Wang, 2011), esses resultados inicialmente colocaram em dúvida se o Trinity conseguiria reconstruir os transcritos com esta quantidade pequena de *reads* resultantes. Entretanto, após a montagem *de novo*, foi possível verificar que além de conseguir realizar a montagem, o programa Trinity aproveitou mais de 90% dos *reads* para ambos os organismos. Isto sugere que o terceiro alinhamento e a seleção dos *reads* mapeados corretamente em um único gene obteve êxito. Desta forma, decidimos investigar qualitativa e quantitativamente a contribuição de variantes de *splicing* detectadas exclusivamente pelo uso do programa Trinity, uma vez que já havíamos mostrado que a estratégia de usar o Uniprot/Swiss-Prot com dados de ESTs é eficaz na detecção de isoformas de *splicing* (Tavares *et al.*, 2014; Tavares *et al.*, *no prelo*).

O número de transcritos montados pelo programa Trinity variou significativamente entre os organismos. Entre os dados de RNA-Seq de homem, o número máximo de transcritos montados por uma corrida foi de aproximadamente 30 mil (réplica 1 - tecido ovariano), enquanto em camundongo, o número máximo de transcritos montados foi de aproximadamente 124 mil (réplica 1 - tecido testicular). Esses números podem ser explicados pela diferença na complexidade dos organismos analisados (Lin *et al*., 2014) e que se reflete na montagem *de novo*. Este fato também é relatado no trabalho original do programa Trinity onde a porcentagem de transcritos montados para a levedura (*Schizosaccharomyces pombe*) era maior quando comparado com camundongo.

A identificação das variantes de *splicing* em nossa base de dados foi de fundamental importância para sabermos se os transcritos montados pelo Trinity poderiam ser confirmados por outros dados de transcriptoma. Em humano, a maior

parte dos transcritos montados em cada corrida pode ser confirmada por ESTs e sequências curadas do RefSeq ou ESTs. Já em camundongo, a maior parte dos transcritos foi confirmada por ESTs. Essas confirmações por outras fontes de dados experimentais indicam que a montagem realizada pelo Trinity conseguiu reconstruir possíveis transcritos derivados de eventos de *splicing* alternativo mesmo após todas as etapas do *pipeline*. As demais variantes geradas pelo programa Trinity que não foram confirmadas por nenhum dado experimental de transcriptoma tinham o potencial de indicar novos transcritos derivados de eventos de *splicing* alternativo.

A partir do sistema de tradução desenvolvido para criação do repositório SpliceProt, as variantes geradas pelo programa Trinity e sem confirmação por outras sequências foram traduzidas a fim de serem confirmadas futuramente por experimentos de espectrometria de massas. Para a maioria dos dados de RNA-Seq em ambos os organismos, pelo menos metade das variantes foi traduzida com sucesso. As variantes que não foram traduzidas tiveram um número excessivo de códons de parada ou sequências proteicas de tamanho muito pequeno.

Com o intuito de confirmar os transcritos reconstruídos pela montagem de transcriptoma e potenciais variantes de *splicing*, dois repositórios proteicos personalizados foram criados a partir das réplicas oriundas de RNA-Seq de tecido cerebral de homem e camundongo. Ambos repositórios foram construídos com as sequências de proteínas canônicas e peptídeos digeridos *in silico* e não redundantes resultantes das isoformas obtidas pelos programas Cufflinks e Trinity.

Para homem, duas regiões cerebrais distintas foram analisadas: lobo temporal anterior (ATL) e corpo caloso (CC). A primeira região é localizada no córtex e está relacionada com a doença de Alzheimer (Domoto-Reilly *et al*., 2012). Já a segunda, encontra-se na parte interna do cérebro, interligando os hemisférios cerebrais e está relacionada com a esquizofrenia (Paul *et al*., 2007). O estudo original de tais amostras não explorou as potenciais isoformas derivadas de eventos de *splicing* alternativo no proteoma das regiões analisadas. Assim, decidimos investigar a contribuição do uso de dados de RNA-Seq para a detecção de variantes de *splicing* usando o programa Trinity. A classificação entre proteínas canônicas e isoformas foi obtida através do sítio da UCSC, onde foi obtida uma lista de transcritos canônicos tanto para homem quanto para camundongo. Usando o critério de 1% de PEP, foram identificadas 136 isoformas em ATL e 109 em CC, sendo contabilizadas 163 isoformas distintas onde a maior parte delas era compartilhada por ambas as regiões. Ao aplicarmos o GSEA para as isoformas encontradas tanto

em ATL quanto em CC, o termo “muscle thin filament tropomyosin” foi novamente encontrado, indicando um perfil de expressão de variantes de *splicing* associado a genes do citoesqueleto. Em uma análise comparativa, foi possível verificar que das 39 isoformas encontradas no estudo em oligodendrócitos, 11 também foram encontradas em ATL e CC. Dessas 11, 2 eram exclusivas de CC (genes *EEF1D* e *PDLIM3*), uma exclusiva de ATL (gene *PNKD*) e 8 eram compartilhadas entre as duas regiões (genes *KRAS*, *PKM*, *TPM1*, *TPM3*, *SET*, *SPTAN1*, *TPD52L2* e *CAPZB*). Outra observação relevante que esta análise revelou foi a presença da família de tropomiosinas (*TPM1*, *TPM3* e *TPM4*) nessas regiões analisadas. Isso sugere que os tipos celulares encontrados no cérebro possam expressar com certa regularidade os membros dessa família e que a linhagem híbrida de oligodendrócitos não influenciou nos resultados encontrados naquele estudo. Entretanto, é importante ressaltar que as variantes de *splicing* identificadas neste estudo refletem aquelas descritas previamente pelo projeto RefSeq ou inéditas e propostas pelo programa Trinity. Uma parcela não desprezível das variantes de *splicing* detectadas usando dados de ESTs não foi analisada no contexto desta tese.

Entre as isoformas encontradas e que eram derivadas exclusivamente do programa Trinity, podemos observar que o peptídeo encontrado para a isoforma do gene *KIF1A* não pode ser utilizado para diferenciação das isoformas conhecidas para este gene. Isto se deve pelo fato do peptídeo “AFVYLSNLLYPVPLVHR” ser compartilhado entre a isoforma hipotética e a isoforma curada NP_004312. Este resultado reflete o desafio de encontrar peptídeos que auxiliem na identificação das isoformas (Nesvizhskii e Aebersold, 2005).

As razões PSM/UP calculadas para ATL e CC indicam que os peptídeos únicos possuem em média de 4 a 5 espectros confirmando-os nos experimentos de MS. Esses valores são pelo menos duas vezes maiores quando comparado com a razão obtida na linhagem de oligodendrócitos. Aplicando-se a razão PSM/UP nas isoformas encontradas em ATL, das 132 isoformas, 52 têm sua razão superior à calculada para as proteínas canônicas. Em CC, o número é reduzido de 109 para 19 isoformas. Ao verificarmos a distribuição delas entre as regiões, foram contabilizadas 58 isoformas distintas, sendo a maioria exclusiva de ATL, 6 exclusivas de CC e 13 compartilhadas. Esses valores mostram que dependendo do número de isoformas encontradas, é possível aplicar este critério de seleção em experimentos de MS que utilizam repositórios personalizados constituídos por peptídeos.

Ao analisarmos os genes que dão origem às proteínas identificadas, foi observado um número significativo de genes que expressavam somente isoformas. Entretanto, em a inspeção manual dessas isoformas mostrou que em outras bases de dados, como UniProt/SwissProt, essas são classificadas como canônicas. Dessa forma, podemos concluir que dependendo do critério de classificação utilizado, o número de isoformas pode variar de acordo com a base de dados utilizada para a construção dos repositórios customizados. De uma forma geral, não existe consenso na literatura indicando como definir, entre os transcritos expressos de um determinado gene, qual deve ser considerado canônico ou quais são derivados de eventos de *splicing* alternativo. A razão para isso se dá pela especificidade celular deste evento, além da geração de outras isoformas derivadas promotores alternativos e poliadenilações alternativas (revisto por de Klerk e 't Hoen, 2016).

Entre os genes que expressavam as proteínas canônicas e as isoformas, apenas dois apresentaram o número de espectros da isoforma foi maior que o da canônica, sendo eles: *NUDT16* e *NEBL*. O gene *NUDT16* é membro da superfamília de hidrolases conhecida como NUDIX e um dos responsáveis por regular a estabilidade e RNAs mensageiros a partir da remoção do *cap* 5' (Song *et al*., 2010). Já o gene *NEBL* codifica proteínas que ligam os filamentos intermediários à linha Z dos sarcômeros (Holmes e Moncman, 2008). Esta análise foi realizada para a comparação com a mesma abordagem na linhagem de oligodendrócitos, onde o gene *SDR39U1* apresentou a mesma diferença entre espectros. Tratando-se de um novo experimento, a expectativa era de que o número de genes que apresentasse o mesmo padrão fosse maior. Entretanto, os resultados indicam que esta análise só pode ser aplicada de forma pontual. Ademais, quando comparamos a classificação entre proteínas canônicas e isoformas na base UniProt/SwissProt, a isoforma encontrada em nosso estudo para o gene *NUDT16* é considera como canônica para esta base e vice-versa. Já a classificação para as proteínas do gene *NEBL* coincidiu com o UniProt/SwissProt.

Um dos peptídeos encontrados (VTVAESSSDGR) para a proteína canônica (NP_808878) do gene *SYT12*, também foi encontrado no trabalho original dos dados de ATL e CC de Martins-de-Souza e colaboradores (2014). Apesar do resultado promissor, este mesmo peptídeo é compartilhado pelas demais proteínas expressas por este gene, mostrando a importância da identificação de peptídeos proteotípicos (Kuster *et al*., 2005).

O repositório proteico feito para identificação de isoformas em camundongo foi aplicado em dados de proteômica de corpo caloso provenientes de três réplicas biológicas. O total de proteínas identificadas em cada uma foi aproximadamente quatro vezes maior quando comparado com os resultados obtidos em corpo caloso de homem. Entre outros fatores, esta diferença tem explicação na preparação das amostras no equipamento utilizado. Enquanto para os dados humanos foi empregada uma coluna cromatográfica menor e um espectrômetro de massas de menor resolução (*ion trap*), os dados de camundongo utilizaram uma coluna maior e um espectrômetro de maior resolução (Q Exactive).

Usando o critério de 1% de PEP, foram identificadas 110 isoformas distintas e sua distribuição apontou que aproximadamente 36% delas eram específicas de cada réplica e menos da metade (44) era compartilhada. Isto sugere que mesmo com uma média de 75 isoformas por réplica, houve uma variabilidade significativa de proteínas exclusivas de cada experimento. Quando comparado ao número de isoformas identificadas em corpo caloso de humano, o número inferior de isoformas encontradas tem possível explicação na composição do repositório proteico gerado. Baseado também na classificação de proteínas canônicas e isoformas do sítio da UCSC, apenas 3.150 proteínas foram classificadas como isoformas, enquanto para humano foram classificadas 9.821.

O gene *Slc4a10* apresentou um isoforma hipotética derivada exclusivamente do programa Trinity e que foi comparada com uma isoforma curada do mesmo gene. O gene *Slc4a10*, é descrito como responsável por regular o pH intracelular de neurônios e astrócitos em camundongos (Liu *et al*., 2010). Alterações nesse gene acometem a diminuição de ventrículos cerebrais e reduzem a excitabilidade neuronal (Jacobs *et al*., 2008). O alinhamento entre a isoforma hipotética a isoforma curada demonstra que as porções inicial e final são diferentes na composição de aminoácidos. O motivo para tal diferença está provavelmente relacionado com uma falha no programa Trinity nas mesmas regiões do transcrito hipotético de origem. A segunda isoforma identificada exclusivamente pelo programa Trinity pertencia ao gene *Ralgapa1*, relacionado com microcefalia e retardo psicomotor (Schwarzbraun *et al*., 2004). O alinhamento da isoforma hipotética contra uma das isoformas curadas do mesmo gene (NP_001273192) indica o mesmo problema com a montagem na região inicial, porém, a região final da isoforma hipotética foi corretamente montada e traduzida até o éxon 34. O peptídeo encontrado e que está localizado no éxon 20 é compartilhado entre as duas proteínas, impossibilitando

novamente a identificação de qual isoforma foi realmente encontrada no experimento.

A razão PSM/UP calculada para as réplicas apontou que os peptídeos únicos das proteínas canônicas apresentavam em média dois espectros e, que das 110 isoformas distintas encontradas, 35 seriam selecionadas por razão superior a calculada. A distribuição delas mostra que quase metade das isoformas com *cut-off* são exclusivas de cada réplica analisada. Isso demonstra a importância de estudos individualizados na identificação de isoformas em estudos de espectrometria de massas.

Da mesma forma como os resultados em homem, foi observado um número significativo de genes que expressavam somente isoformas. Entre os genes que apresentavam a proteína canônica e a isoforma expressa, três tiveram o número de espectros de suas isoformas maior que os espectros canônicos: *Mical3* (membro da família das monooxigenases associadas aos microtúbulos), *Dnm1l* (relacionado com fissão mitocondrial; Brooks *et al*., 2009) e *Nup98* (integrante do complexo proteico no poro nuclear em humano; Griffis *et al*., 2002). Esses resultados indicam novamente que esta análise só pode ser aplicada de forma pontual além de mostrar novamente a influência da base de dados utilizada para a classificação entre proteínas canônicas e isoformas. Ao compararmos a classificação com o Uniprot/SwissProt, as isoformas identificadas para os genes *Nup98* e *Mical3* são classificadas como canônicas para a referida base de dados. E a proteína canônica do gene *Dnm1l* é atribuída a uma proteína predita computacionalmente pelo projeto RefSeq (prefixo XP).

A análise de todas as proteínas identificadas em corpo caloso de homem e camundongo revelou uma quantidade significativa de genes ortólogos de proteínas canônicas e sete genes ortólogos entre as isoformas. Tais resultados sugerem que os transcritos expressos pelas isoformas dos genes desempenham funções importantes para os tipos celulares que compõem esta região cerebral. Quatro dos sete genes (*TPM3*, *EEF1D*, *PKM* e *SET*) também foram encontrados no estudo com a linhagem celular de oligodendrócitos e discutidos anteriormente em relação às suas isoformas em humano. Os genes *CDC42*, *SEPT7* e *RUFY3* estão envolvidos em diversas funções tanto em homem quanto camundongo e serão discutidos a seguir.

O gene *CDC42* pertence à família de GTPases Rho e está relacionado com o ciclo celular (Olson *et al*., 1995), reorganização do citoesqueleto (Dutartre *et al*.,

1996) entre outras funções (revisto por Van Aelst e D'Souza-Schorey, 1997). Este gene produz duas isoformas idênticas (RefSeq NP_001782 e NP_001034891) que diferem de uma terceira (RefSeq NP_426359) na região N-terminal. O peptídeo "NVFDEAILAALEPPETQPK" encontrado em nossa análise possibilitou a confirmação da isoforma NP_426359 no experimento por conter cinco aminoácidos que as distingue das demais isoformas. Em camundongo, o gene ortólogo *Cdc42* está relacionado na regulação do citoesqueleto (revisto por Hall, 1998) e com a produção de mielina em oligodendrócitos juntamente com o gene *Rac1* (Thurnherr *et al*., 2006). Este gene produz duas isoformas (RefSeq NP_001230698 e NP_033991) que diferem na região N-terminal da mesma maneira como no gene humano. Foi possível confirmar a presença da isoforma NP_001230698 no experimento através da arginina encontrada no peptídeo "YVECSALTQR" e nos cinco últimos aminoácidos do peptídeo "NVFDEAILAALEPPETQPK", o mesmo encontrado em corpo caloso de humano.

As septinas pertencem a uma família de genes relacionada com a organização do citoesqueleto através da formação de filamentos (revisto por Kinoshita, 2003; Zent *et al*., 2011). Estudos já demonstraram a relação entre a superexpressão deste gene com inibição de glioma em humanos (Jia *et al*., 2010; Xu *et al*., 2010) e sua baixa expressão em pacientes com esquizofrenia juntamente com o aumento da expressão de *CDC42* (Ide e Lewis, 2010). Kinoshita e colaboradores (2000) analisaram a presença das septinas em diferentes regiões do cérebro de camundongo onde a septina 7 (também conhecida por CDC10) foi localizada no cortéx cerebral, no hipocampo e no corpo caloso. Em humano, o gene *SEPT7* expressa três isoformas, uma menor (RefSeq NP_001229885) e duas maiores (RefSeq NP_001011553 e NP_001779). Estas duas variantes de *splicing* de sequência com alta similaridade diferem em apenas um aminoácido (valina) na porção inicial. Nos experimentos analisados, foram encontrados seis peptídeos atribuídos às duas isoformas maiores deste gene e que não possibilitaram a distinção entre elas. Em camundongo, o gene *Sept7* expressa apenas duas isoformas que se distinguem pela presença de uma valina na porção inicial da isoforma NP_033989. Em nossas análises, apenas um peptídeo foi encontrado "SVNCGTMAQPK" e atribuído a isoforma NP_001192296, sendo este suficiente para distingui-la da isoforma NP_033989.

O gene *RUFY3* pertence a uma família de proteínas que está relacionada com a regulação da polaridade neuronal em ratos (Mori *et al*., 2007) e no tráfego de

membrana (revisto Kitagishi e Matsuda, 2013). Wang e colaboradores (2015) demonstraram que *RUFY3* é regulado positivamente pelo gene *PAK1*, apresentando relação com a migração e invasão de tumores gástricos. Em camundongo, Wei e colaboradores (2014) relacionaram Rufy3 com a morfogênese dos axônios. No experimento em humano, três peptídeos foram identificados e atribuídos a duas isoformas do gene *RUFY3* (RefSeq NP_055776 e NP_001124181). Porém, estes peptídeos não foram suficientes para confirmar qual das duas isoformas estava presente no experimento. No experimento em camundongo, apenas o peptídeo "INSLQLEVEALTR" foi atribuído à isoforma NP_001276703 do gene *Rufy3*. Quando comparado com as demais proteínas deste gene, o peptídeo encontrado poderia ser utilizado para distingui-la das demais.

Assim, acreditamos que há muito ainda a ser aprimorado no campo da proteogenômica, mas que os estudos realizados no âmbito deste doutorado propiciaram a resolução de problemas inerentes à construção de repositórios de sequências proteicas personalizados, e a sua aplicação na análise de dados de MS no que concerne a identificação de variantes por *splicing* alternativo. Como exposto, esta tese buscou reunir o aprimoramento da análise de variantes de *splicing* por uma abordagem de proteogenômica. O desenvolvimento de repositórios personalizados foi determinante para a identificação das isoformas em diferentes linhagens e amostras teciduais. A estratégia de montagem de transcriptoma com genoma de referência e *de novo* conseguiu reconstruir os transcritos a partir de corridas de RNA-Seq, apesar das etapas que reduzirem o número de *reads* destinados a montagem com o programa Trinity. A identificação das variantes de *splicing*, a tradução *in silico* e a detecção destas em experimentos de espectrometria de massas obteve êxito apesar de o programa Trinity ter mostrado não ser tão eficiente em relação à contribuição de variantes de *splicing* inéditas ausentes em outros bancos de dados de sequências expressas.

Desta forma, esperamos ter contribuído com o desenvolvimento de uma abordagem computacional para a tradução computacional de dados de transcriptoma e a descrição de variantes de *splicing* identificadas em dados de MS de uma linhagem celular de olidendrócitos humanos, bem como de duas regiões do cérebro humano e de camundongo.

6. CONCLUSÃO

A integração de dados de proteômica e transcriptômica caracteriza a área da proteogenômica, que vem ganhando destaque ao longo dos últimos anos. As tecnologias de alta vazão como RNA-Seq e a espectrometria de massas ofertam uma grande quantidade de dados e, ao mesmo tempo, desafiam na identificação de novas variantes de *splicing* e principalmente sua função em diferentes tipos celulares e amostras. Esta tese apresentou análises de diferentes variantes de *splicing em* dados de transcriptoma e proteoma de homem e camundongo. A partir de dados de ESTs e RNA-Seq que foram mapeados nos genomas dos organismos estudados, foi possível identificar as variantes de *splicing*, traduzi-las computacionalmente e criar repositórios com sua tradução hipotética. Assim, esses repositórios puderam ser utilizados em diferentes dados de espectrometria de massas para a identificação e análise das variantes identificadas. Portanto, concluímos que a integração de dados de diferentes plataformas pode ampliar a descoberta de novas variantes por *splicing* alternativo.

7. REFERÊNCIAS

Anczuków O, Rosenberg AZ, Akerman M, Das S, Zhan L, Karni R, *et al*. The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. Nat Struct Mol Biol. 2012 Jan 15;19(2):220-8.

Aschoff M, Hotz-Wagenblatt A, Glatting KH, Fischer M, Eils R, König R. SplicingCompass: differential splicing detection using RNA-seq data. Bioinformatics. 2013 May 1;29(9):1141-8.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.

Aston FW. A positive ray spectrograph. Phils Mag. 1919;38:707-714.

Bantscheff M, Lemeer S, Savitski MM, Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal Bioanal Chem. 2012 Sep;404(4):939-65.

Bauer S, Gagneur J, Robinson PN. GOing Bayesian: model-based gene set analysis of genome-scale data. Nucleic Acids Res. 2010 Jun;38(11):3523-32.

Black DL. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell. 2000 Oct;103(3):367-70.

Blank T, Prinz M. NF-κB signaling regulates myelination in the CNS. Front Mol Neurosci. 2014 May 26;7:47.

Blencowe BJ. Alternative splicing: new insights from global analyses. Cell. 2006 Jul;126(1):37-47.

Blencowe BJ, Ahmad S, Lee LJ. Sequencing : deepening insights into mammalian transcriptomes. Genes Dev. 2009 Jun;23(12):1379-86.

Boguski MS, Lowe TM, Tolstoshev CM. dbEST--database for "expressed sequence tags". Nat Genet. 1993 Aug;4(4):332-3.

Breathnach R, Chambon P. Organization and expression of eucaryotic split genes coding for proteins. Annu Rev Biochem. 1981;50:349-83.

Brinkman BM. Splice variants as cancer biomarkers. Clin Biochem. 2004 Jul;37(7):584-94.

Brooks C, Wei Q, Cho SG, Dong Z. Regulation of mitochondrial dynamics in acute kidney injury in cell culture and rodent models. J Clin Invest. 2009 May;119(5):1275-85.

Burd CG, Swanson MS, Görlach M, Dreyfuss G. Primary structures of the heterogeneous nuclear ribonucleoprotein A2, B1, and C2 proteins: a diversity of RNA binding proteins is generated by small peptide inserts. Proc Natl Acad Sci U S A. 1989 Dec;86(24):9788-92.

Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 2000 Nov;28(21):4364-75.

Campos-Sandoval JA, Martín-Rufián M, Cardona C, Lobo C, Peñalver A, Márquez J. Glutaminases in brain: Multiple isoforms for many purposes. Neurochem Int. 2015 Sep;88:1-5.

Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113-20.

Cao Y, Portela M, Janikiewicz J, Doig J, Abbott CM Characterisation of translation elongation factor eEF1B subunit expression in mammalian cells and tissues and co-localisation with eEF1A2. PLoS One. 2014 Dec 1;9(12):e114117.

Chan DC. Mitochondrial fusion and fission in mammals. Annu Rev Cell Dev Biol. 2006;22:79-99.

Chen F, Lin L, Zhang J, He Z, Uchiyama K, Lin JM. Single-Cell Analysis Using Drop-on-Demand Inkjet Printing and Probe Electrospray Ionization Mass Spectrometry. Anal Chem. 2016 Apr 19;88(8):4354-60.

Christian KJ, Lang MA, Raffalli-Mathieu F. Interaction of heterogeneous nuclear ribonucleoprotein C1/C2 with a novel cis-regulatory element within p53 mRNA as a response to cytostatic drug treatment. Mol Pharmacol. 2008 May;73(5):1558-67.

Christofk HR, Vander Heiden MG, Harris MH, Ramanathan A, Gerszten RE, Wei R, *et al*. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. Nature. 2008 Mar 13;452(7184):230-3.

Chu HT, Hsiao WW, Chen JC, Yeh TJ, Tsai MH, Lin H, *et al*. EBARDenovo: highly accurate de novo assembly of RNA-Seq with efficient chimera-detection. Bioinformatics. 2013 Apr 15;29(8):1004-10.

Clarke K, Yang Y, Marsh R, Xie L, Zhang KK. Comparative analysis of de novo transcriptome assembly. Sci China Life Sci. 2013 Feb;56(2):156-62.

Clower CV, Chatterjee D, Wang Z, Cantley LC, Vander Heiden MG, Krainer AR. The alternative splicing repressors hnRNP A1/A2 and PTB influence pyruvate kinase isoform expression and cell metabolism. Proc Natl Acad Sci U S A. 2010 Feb 2;107(5):1894-9.

De Bortoli M, Castellino RC, Lu XY, Deyo J, Sturla LM, Adesina AM, *et al*. Medulloblastoma outcome is adversely associated with overexpression of EEF1D,

RPL30, and RPS20 on the long arm of chromosome 8. BMC Cancer. 2006 Sep 12;6:223.

de Klerk E, 't Hoen PA. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends Genet. 2015 Mar;31(3):128-39.

Dempster AJ. A New Method of Positive Ray Analysis. Phys Rev. 1918;11:316.

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, *et al*. The PeptideAtlas project. Nucleic Acids Res. 2006 Jan 1;34:D655-8.

Domon B, Aebersold R. Mass spectrometry and protein analysis. Science. 2006 Apr 14;312(5771):212-7.

Domoto-Reilly K, Sapolsky D, Brickhouse M, Dickerson BC; Alzheimer's Disease Neuroimaging Initiative. Naming impairment in Alzheimer's disease is associated with left anterior temporal lobe atrophy. Neuroimage. 2012 Oct 15;63(1):348-55.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, *et al*. An integrated encyclopedia of DNA elements in the human genome. ENCODE Project Consortium. Nature. 2012 Sep 6;489(7414):57-74.

Dutartre H, Davoust J, Gorvel JP, Chavrier P. Cytokinesis arrest and redistribution of actin-cytoskeleton regulatory components in cells expressing the Rho GTPase CDC42Hs. J Cell Sci. 1996 Feb;109 (Pt 2):367-77.

Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. Cell. 1980 Jun;20(2):313-9.

Ekblom R, Wolf JB. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 2014 Nov;7(9):1026-42.

Elgadi KM, Meguid RA, Qian M, Souba WW, Abcouwer SF. Cloning and analysis of unique human glutaminase isoforms generated by tissue-specific alternative splicing. Physiol Genomics. 1999 Aug 31;1(2):51-62.

Elias AP, Dias S. Microenvironment changes (in pH) affect VEGF alternative splicing. Cancer Microenviron. 2008 Dec;1(1):131-9.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74.

Fanayan S, Smith JT, Lee LY, Yan F, Snyder M, Hancock WS, *et al*. Proteogenomic analysis of human colon carcinoma cell lines LIM1215, LIM1899, and LIM2405. J Proteome Res. 2013 Apr 5;12(4):1732-42.

Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev. 2003 Feb;17(4):419-37.

Florea L. Bioinformatics of alternative splicing and its regulation. Brief Bioinform. 2006 Mar;7(1):55-69.

Gandre-Babbe S, van der Bliek AM. The novel tail-anchored membrane protein Mff controls mitochondrial and peroxisomal fission in mammalian cells. Mol Biol Cell. 2008 Jun;19(6):2402-12.

Gilbert W. Why genes in pieces? Nature. 1978 Feb;271(5645):501.

Goodison S, Urquidi V, Tarin D. CD44 cell adhesion molecules. Mol Pathol. 1999 Aug;52(4):189-96.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011 May 15;29(7):644-52.

Griffis ER, Altan N, Lippincott-Schwartz J, Powers MA. Nup98 is a mobile nucleoporin with transcription-dependent dynamics. Mol Biol Cell. 2002 Apr;13(4):1282-97.

Gundry RL, White MY, Murray CI, Kane LA, Fu Q, Stanley BA, *et al*. Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow. Curr Protoc Mol Biol. 2009 Oct;Chapter 10:Unit10.25.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, *et al*., Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010 May;28(5):503-10.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol. 1999 Oct;17(10):994-9.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003 Oct 1;31(19):5654-66.

Hall A. Rho GTPases and the actin cytoskeleton. Science. 1998 Jan 23;279(5350):509-14.

Hallegger M, Llorian M, Smith CW. Alternative splicing global insights. FEBS J. 2010 Feb;277(4):856-66.

Han J, Xiong J, Wang D, Fu XD. Pre-mRNA splicing: where and when in the nucleus. Trends Cell Biol. 2011 Jun;21(6):336-43.

Hanke S, Besir H, Oesterhelt D, Mann M. Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level. J Proteome Res. 2008 Mar;7(3):1118-30.

Harper SJ, Bates DO. VEGF-A splicing: the key to anti-angiogenic therapeutics? Nat. Rev Cancer. 2008 Nov;8(11):880-7.

Herzel L, Neugebauer KM. Quantification of co-transcriptional splicing from RNA-Seq data. Methods. 2015 Sep 1;85:36-43.

Higareda-Almaraz JC, Valtierra-Gutiérrez IA, Hernandez-Ortiz M, Contreras S, Hernandez E, Encarnación-Guevara S. Analysis and prediction of pathways in HeLa cells by integrating biological levels of organization with systems-biology approaches. PLoS One. 2013 Jun 10;8(6):e65433.

Holmes WB, Moncman CL. Nebulette interacts with filamin C. Cell Motil Cytoskeleton. 2008 Feb;65(2):130-42.

Hong Y, Kim WJ, Bang CY, Lee JC, Oh YM. Identification of Alternative Splicing and Fusion Transcripts in Non-Small Cell Lung Cancer by RNA Sequencing. Tuberc Respir Dis (Seoul). 2016 Apr;79(2):85-90.

Hu B, Luo W, Hu RT, Zhou Y, Qin SY, Jiang HX. Meta-Analysis of Prognostic and Clinical Significance of CD44v6 in Esophageal Cancer. Medicine (Baltimore). 2015 Aug;94(31):e1238.

Ide M, Lewis DA. Altered cortical CDC42 signaling pathways in schizophrenia: implications for dendritic spine deficits. Biol Psychiatry. 2010 Jul 1;68(1):25-32.

Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, *et al*. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics. 2005 Sep;4(9):1265-72.

Iwata K, Matsuzaki H, Manabe T, Mori N. Altering the expression balance of hnRNP C1 and C2 changes the expression of myelination-related genes. Psychiatry Res. 2011 Dec 30;190(2-3):364-6.

Iwata K, Café-Mendes CC, Schmitt A, Steiner J, Manabe T, Matsuzaki H, *et al*. Proteomics. 2013 Dec;13(23-24):3548-53.

Jacobs S, Ruusuvuori E, Sipilä ST, Haapanen A, Damkier HH, Kurth I, *et al*. Mice with targeted Slc4a10 gene disruption have small brain ventricles and show reduced neuronal excitability. Proc Natl Acad Sci U S A. 2008 Jan 8;105(1):311-6.

Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics. 2004 Jan;4(1):59-77.

Jeong SK, Hancock WS, Paik YK. GenomewidePDB 2.0: A Newly Upgraded Versatile Proteogenomic Database for the Chromosome-Centric Human Proteome Project. J Proteome Res. 2015 Sep 4;14(9):3710-9.

Jia ZF, Huang Q, Kang CS, Yang WD, Wang GX, Yu SZ, *et al*. Overexpression of septin 7 suppresses glioma cell growth. J Neurooncol. 2010 Jul;98(3):329-40.

Kaitsuka T, Matsushita M. Regulation of translation factor EEF1D gene function by alternative splicing. Int J Mol Sci. 2015 Feb 12;16(2):3970-9.

Käll L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res. 2008 Jan;7(1):40-4.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, *et al*. The UCSC Genome Browser Database. Nucleic Acids Res. 2003 Jan;31(1):51-4.

Keerthikumar S, Gangoda L, Liem M, Fonseka P, Atukorala I, Ozcitti C, *et al*. Proteogenomic analysis reveals exosomes are more oncogenic than ectosomes. Oncotarget. 2015 Jun 20;6(17):15375-96.

Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64.

Kim N, Lee C. Bioinformatics detection of alternative splicing. Methods Mol Biol. 2008;452:179-97.

Kinoshita A, Noda M, Kinoshita M. Differential localization of septins in the mouse brain. J Comp Neurol. 2000 Dec 11;428(2):223-39.

Kinoshita M. The septins. Genome Biol. 2003;4(11):236.

Kircher M, Kelso J. High-throughput DNA sequencing--concepts and limitations. Bioessays. 2010 Jun;32(6):524-36.

Kitagishi Y, Matsuda S. RUFY, Rab and Rap Family Proteins Involved in a Regulation of Cell Polarity and Membrane Trafficking. Int J Mol Sci. 2013 Mar 21;14(3):6487-98.

Krainer AR. The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. Nat Struct Mol Biol. 2012 Jan;19(2):220-8.

Krasnov GS, Dmitriev AA, Kudryavtseva AV, Shargunov AV, Karpov DS, Uroshlev LA, *et al*. PPLine: An Automated Pipeline for SNP, SAP, and Splice Variant Detection in the Context of Proteogenomics. J Proteome Res. 2015 Sep 4;14(9):3729-37.

Krug K, Popic S, Carpy A, Taumer C, Macek B. Construction and assessment of individualized proteogenomic databases for large-scale analysis of nonsynonymous single nucleotide variants. Proteomics. 2014 Dec;14(23-24):2699-708.

Kuster B, Schirle M, Mallick P, Aebersold R. Scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol. 2005 Jul;6(7):577-83.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al*. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921. Erratum in: Nature 2001 Aug 2;412(6846):565. Nature. 2001 Jun;411(6838):720.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. The Sequence Alignment/Map format and SAMtools.Bioinformatics. 2009 Aug 15;25(16):2078-9.

Lin JJ, Eppinga RD, Warren KS, McCrae KR. Human tropomyosin isoforms in the regulation of cytoskeleton functions. Adv Exp Med Biol. 2008;644:201-22.

Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, *et al*. Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A. 2014 Dec 2;111(48):17224-9.

Liu H, Sadygov RG, Yates JR 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem. 2004 Jul 15;76(14):4193-201.

Liu Y, Xu K, Chen LM, Sun X, Parker MD, Kelly ML, *et al*. Distribution of NBCn2 (SLC4A10) splice variants in mouse brain. Neuroscience. 2010 Sep 1;169(3):951-64.

Love JE, Hayden EJ, Rohn TT. Alternative Splicing in Alzheimer's Disease. J Parkinsons Dis Alzheimers Dis. 2015 Aug;2(2). pii: 6.

Lu B, Zeng Z, Shi T. Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. Sci China Life Sci. 2013 Feb;56(2):143-55.

Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, *et al*. Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol. 2010 Dec 21;6:450.

Márquez J, de la Oliva AR, Matés JM, Segura JA, Alonso FJ. Glutaminase: a multifaceted protein not only involved in generating glutamate. Neurochem Int. 2006 May-Jun;48(6-7):465-71.

Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011 Sep 7;12(10):671-82.

Martins-de-Souza D, Carvalho PC, Schmitt A, Junqueira M, Nogueira FC, Turck CW, *et al*. Deciphering the human brain proteome: characterization of the anterior temporal lobe and corpus callosum as part of the Chromosome 15-centric Human Proteome Project. J Proteome Res. 2014 Jan 3;13(1):147-57.

Martins-de-Souza D, Gattaz WF, Schmitt A, Novello JC, Marangoni S, Turck CW, *et al*. Proteome analysis of schizophrenia patients Wernicke's area reveals an energy metabolism dysregulation. BMC Psychiatry. 2009 Apr 30;9:17.

McFarlane S, Coulter JA, Tibbits P, O'Grady A, McFarlane C, Montgomery N, *et al*. CD44 increases the efficiency of distant metastasis of breast cancer. Oncotarget. 2015 May 10;6(13):11465-76.

McGrath JP, Capon DJ, Smith DH, Chen EY, Seeburg PH, Goeddel DV, *et al*. Structure and organization of the human Ki-ras proto-oncogene and a related processed pseudogene. Nature. 1983 Aug 11-17;304(5926):501-6.

Menon R, Im H, Zhang EY, Wu SL, Chen R, Snyder M, *et al*. Distinct splice variants and pathway enrichment in the cell-line models of aggressive human breast cancer subtypes. J Proteome Res. 2014 Jan 3;13(1):212-27.

Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, *et al*. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016 May 25;534(7605):55-62.

Metzker ML Sequencing technologies – the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46.

Meurer EC. Tecnicas Modernas em Espectrometria de Massas: Aplicações analíticas e no estudo de reações íon/molécula na fase gasosa. São Paulo. Tese [Doutorado em Química] – UNICAMP; 2003.

Modrek B, Lee C. A genomic view of alternative splicing. Nat Genet. 2002 Jan;30(1):13-9.

Mori T, Wada T, Suzuki T, Kubota Y, Inagaki N. Singar1, a novel RUN domain-containing protein, suppresses formation of surplus axons for neuronal polarity. J Biol Chem. 2007 Jul 6;282(27):19884-93.

Mortazavi A, Schwarz EM, Williams B, Schaeffer L, Antoshechkin I, Wold BJ, *et al*., Scaffolding a Caenorhabditis nematode genome with RNA-seq. Genome Res. 2010 Dec;20(12):1740-7.

Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C. Splicing signals in Drosophila: intron size, information content, and consensus sequences. Nucleic Acids Res. 1992 Aug;20(16):4255-62.

Mount SM. A catalogue of splice junction sequences. Nucleic Acids Res. 1982 Jan;10(2):459-72.

Munson MSB, Field FH. Chemical Ionization Mass Spectrometry. I. General Introduction. J. Am. Chem. Soc. 1966;88:2621-2630.

Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, *et al*. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol. 2011 Nov 8;7:548.

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, *et al*. Genome sequence of Aedes aegypti, a major arbovirus vector. Science. 2007 Jun 22;316(5832):1718-23.

Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics. 2005 Oct;4(10):1419-40.

Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods. 2014 Nov;11(11):1114-25.

Niikura Y, Kitagawa K. Identification of a novel splice variant: human SGT1B (SUGT1B). DNA Seq. 2003 Dec;14(6):436-41.

Ning K, Nesvizhskii AI. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. BMC Bioinformatics. 2010 Dec 14;11 Suppl 11:S14.

Ohsawa N, Koebis M, Suo S, Nishino I, Ishiura S. Alternative splicing of PDLIM3/ALP, for α-actinin-associated LIM protein 3, is aberrant in persons with myotonic dystrophy. Biochem Biophys Res Commun. 2011 May 27;409(1):64-9.

Okamoto K, Shaw JM. Mitochondrial morphology and dynamics in yeast and multicellular eukaryotes. Annu Rev Genet. 2005;39:503-36.

Olson MF, Ashworth A, Hall A. An essential role for Rho, Rac, and Cdc42 GTPases in cell cycle progression through G1. Science. 1995 Sep 1;269(5228):1270-2.

Orengo JP, Cooper TA. Alternative splicing in disease. Adv Exp Med Biol. 2007;623:212-23.

Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, Cho SY, *et al*. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotechnol. 2012 Mar 7;30(3):221-3.

Park YM, Hwang SJ, Masuda K, Choi KM, Jeong MR, Nam DH, *et al*. Heterogeneous nuclear ribonucleoprotein C1/C2 controls the metastatic potential of glioblastoma by regulating PDCD4. Mol Cell Biol. 2012 Oct;32(20):4237-44.

Paul LK, Brown WS, Adolphs R, Tyszka JM, Richards LJ, Mukherjee P, *et al*. Agenesis of the corpus callosum: genetic, developmental and functional aspects of connectivity. Nat Rev Neurosci. 2007 Apr;8(4):287-99.

Peng Y, Leung HC, Yiu SM, Lv MJ, Zhu XG, Chin FY. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics. 2013 Jul 1;29(13):i326-34.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 2009 Jan;37:D32-6.

Reimers M, Carey VJ. Bioconductor: an open source framework for bioinformatics and computational biology. Methods Enzymol. 2006;411:119-34.

Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000 Jun;16(6):276-7.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, *et al*. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010 Nov;7(11):909-12.

Ruggles KV, Tang Z, Wang X, Grover H, Askenazi M, Teubl J, *et al*. An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. Mol. Cell Proteomics 15, 1060–1071 (2015).

Sacomoto GA, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot MF, *et al*. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinformatics. 2012 Apr 19;13 Suppl 6:S5.

Santos M, Domingues SC, Costa P, Muller T, Galozzi S, Marcus K, *et al*. Identification of a novel human LAP1 isoform that is regulated by protein phosphorylation. PLoS One. 2014 Dec 2;9(12):e113732.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, *et al*., Insights into hominid evolution from the gorilla genome sequence. Nature. 2012 Mar 7;483(7388):169-75.

Schmidt A, Kellermann J, Lottspeich F. A novel strategy for quantitative proteomics using isotope-coded protein labels. Proteomics. 2005 Jan;5(1):4-15.

Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012 Apr 15;28(8):1086-92.

Schwarzbraun T, Vincent JB, Schumacher A, Geschwind DH, Oliveira J, Windpassinger C, *et al*. Cloning, genomic structure, and expression profiles of TULIP1 (GARNL1), a brain-expressed candidate gene for 14q13-linked neurological phenotypes, and its murine homologue. Genomics. 2004 Sep;84(3):577-86.

Schwerk C, Schulze-Osthoff K. Regulation of apoptosis by alternative pre-mRNA splicing. Mol Cell. 2005 Jul;19(1):1-13.

Senior A, Gerace L. Integral membrane proteins specific to the inner nuclear membrane and associated with the nuclear lamina. J Cell Biol. 1988 Dec;107(6 Pt 1):2029-36.

Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. 1987 Sep;15(17):7155-74.

Sharma K, Schmitt S, Bergner CG, Tyanova S, Kannaiyan N, Manrique-Hoyos N. Cell type- and brain region-resolved mouse brain proteome. Nat Neurosci. 2015 Dec;18(12):1819-31.

Sharp PA. Split genes and RNA splicing. Cell. 1994 Jun 17;77(6):805-15.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, *et al*. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.

Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM. Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. Annu Rev Anal Chem (Palo Alto Calif). 2016 Mar 30.

Sheynkman GM, Shortreed MR, Frey BL, Scalf M, Smith LM. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. J Proteome Res. 2014 Jan 3;13(1):228-40.

Sheynkman GM, Shortreed MR, Frey BL, Smith LM. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. Mol Cell Proteomics. 2013 Aug;12(8):2341-53.

Song MG1, Li Y, Kiledjian M. Multiple mRNA decapping enzymes in mammalian cells. Mol Cell. 2010 Nov 12;40(3):423-32.

Staal JA, Lau LS, Zhang H, Ingram WJ, Hallahan AR, Northcott PA. *et al*. Proteomic profiling of high risk medulloblastoma reveals functional biology. Oncotarget. 2015 Jun 10;6(16):14584-95.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006 Jul;34:W435-9.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005 Oct 25;102(43):15545-50.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, *et al*. An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct 1;526(7571):75-81.

Szeliga M, Matyja E, Obara M, Grajkowska W, Czernicki T, Albrecht J. Relative expression of mRNAS coding for glutaminase isoforms in CNS tissues and CNS tumors. Neurochem Res. 2008 May;33(5):808-13.

Tavares R, de Miranda Scherer N, Pauletti BA, Araújo E, Folador EL, Espindola G. *et al*. SpliceProt: a protein sequence repository of predicted human splice variants. Proteomics. 2014 Feb;14(2-3):181-5.

Tavares R, Renaud G, Oliveira PS, Ferreira CG, Dias-Neto E, Passetti F. Identical sequence patterns in the ends of exons and introns of human protein-coding genes. Comput Biol Chem. 2012 Feb;36:55-61.

Tavares R, Scherer NM, Ferreira CG, Costa FF, Passetti F. Splice variants in the proteome: a promising and challenging field to targeted drug discovery. Drug Discov Today. 2015 Mar;20(3):353-60.

Tembe WD, Pond SJ, Legendre C, Chuang HY, Liang WS, Kim NE, *et al*. Open-access synthetic spike-in mRNA-seq data for cancer gene fusions. BMC Genomics. 2014 Sep 30;15:824.

The Cancer Genome Atlas Pan-Cancer analysis project. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, *et al*. Nat Genet. 2013 Oct;45(10):1113-20.

The UniProt Consortium, Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, *et al*. UniProt: a hub for protein information. Nucleic Acids Res. 2015 Jan;43(Database issue):D204-12.

Thiede B, Höhenwarter W, Krah A, Mattow J, Schmid M, Schmidt F, *et al*. Peptide mass fingerprinting. Methods. 2005 Mar;35(3):237-47.

Thomson, JJ. Cathode rays. Phil Mag.1897 Oct;44:293.

Thurnherr T, Benninger Y, Wu X, Chrostek A, Krause SM, Nave KA, *et al*. Cdc42 and Rac1 signaling are both required for and act synergistically in the correct formation of myelin sheaths in the CNS. J Neurosci. 2006 Oct 4;26(40):10110-9.

Tjhay F, Motohara T, Tayama S, Narantuya D, Fujimoto K, Guo J, *et al*. CD44 variant 6 is correlated with peritoneal dissemination and poor prognosis in patients with advanced epithelial ovarian cancer. Cancer Sci. 2015 Oct;106(10):1421-8.

Tokumoto YM, Tang DG, Raff MC. Two molecularly distinct intracellular pathways to oligodendrocyte differentiation: role of a p53 family protein. EMBO J. 2001 Sep 17;20(18):5261-8.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5.

Van Aelst L, D'Souza-Schorey C. Rho GTPases and signaling networks. Genes Dev. 1997 Sep 15;11(18):2295-322.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014 Sep;30(9):418-26.

Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, *et al*. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 2014 Mar;32(3):223-6.

Wajnberg G, Passetti F. Using high-throughput sequencing transcriptome data for INDEL detection: challenges for cancer drug discovery. Expert Opin Drug Discov. 2016 Mar;11(3):257-68.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative Isoform regulation in human tissue transcriptomes. Nature. 2008 Nov;456(7221):470-6.

Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet. 2007 Oct;8(10):749-61.

Wang G, Zhang Q, Song Y, Wang X, Guo Q, Zhang J, *et al*. PAK1 regulates RUFY3-mediated gastric cancer cell migration and invasion. Cell Death Dis. 2015 Mar 12;6:e1682.

Ward AJ, Cooper TA. The pathobiology of splicing. J Pathol. 2010; Jan;220(2):152-63.

Wei Z, Sun M, Liu X, Zhang J, Jin Y. Rufy3, a protein specifically expressed in neurons, interacts with actin-bundling protein Fascin to control the growth of axons. J Neurochem. 2014 Sep;130(5):678-92.

Welch JD, Hu Y, Prins JF. Robust detection of alternative splicing in a population of single cells. Nucleic Acids Res. 2016 May 5;44(8):e73.

Welman A, Burger MM, Hagmann J. Structure and function of the C-terminal hypervariable region of K-Ras4B in plasma membrane targetting and transformation. Oncogene. 2000 Sep 21;19(40):4582-91.

Wery M, Kwapisz M, Morillon A. Noncoding RNAs in gene regulation. Wiley Interdiscip Rev Syst Biol Med. 2011 Dec;3(6):728-38.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 2003 Jan 1;31(1):28-33.

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, *et al*. Mass-spectrometry-based draft of the human proteome. Nature. 2014 May 29;509(7502):582-7.

Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics. 2016 Feb 25;17:103.

Xu S, Jia ZF, Kang C, Huang Q, Wang G, Liu X, *et al*. Upregulation of SEPT7 gene inhibits invasion of human glioma cells. Cancer Invest. 2010 Mar;28(3):248-58.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, *et al*. Ensembl 2016. Nucleic Acids Res. 2016 Jan 4;44(D1):D710-6.

Yates JR 3rd. Mass spectrometry and the age of the proteome. J Mass Spectrom. 1998 Jan;33(1):1-19.

Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng. 2009;11:49-79.

Zent E, Vetter I, Wittinghofer A. Structural and biochemical properties of Sept7, a unique septin required for filament formation. Biol Chem. 2011 Aug;392(8-9):791-7.

Zhang J, Yuan L, Zhang X, Hamblin MH, Zhu T, Meng F, *et al*. Altered long non-coding RNA transcriptomic profiles in brain microvascular endothelium after cerebral ischemia. Exp Neurol. 2016 Mar;277:162-70.

Zhang YV, Rockwood A. Impact of automation on mass spectrometry. Clin Chim Acta. 2015 Oct 23;450:298-303.

Zhao W, Hoadley KA, Parker JS, Perou CM, Identification of mRNA isoform switching in breast cancer. BMC Genomics. 2016 Mar 3;17(1):181.

8. ANEXOS

Anexo 1 - SpliceProt: a protein sequence repository of predicted human splice variants.

Proteomics 2014, *14*, 181–185 DOI 10.1002/pmic.201300078

TECHNICAL BRIEF

SpliceProt: A protein sequence repository of predicted human splice variants

Raphael Tavares[1], Nicole de Miranda Scherer[1], Bianca Alves Pauletti[4], Elói Araújo[5], Edson Luiz Folador[1], Gabriel Espindola[1], Carlos Gil Ferreira[2], Adriana Franco Paes Leme[4], Paulo Sergio Lopes de Oliveira[3] and Fabio Passetti[1]

[1] Bioinformatics Unit, Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rio de Janeiro, Brazil
[2] Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rio de Janeiro, Brazil
[3] Laboratório Nacional de Biociências (LNBio), CNPEM, Campinas, Brazil
[4] Laboratório de Espectrometria de Massas, Laboratório Nacional de Biociências (LNBio), CNPEM, Campinas, Brazil
[5] Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Campo Grande, Brazil

The mechanism of alternative splicing in the transcriptome may increase the proteome diversity in eukaryotes. In proteomics, several studies aim to use protein sequence repositories to annotate MS experiments or to detect differentially expressed proteins. However, the available protein sequence repositories are not designed to fully detect protein isoforms derived from mRNA splice variants. To foster knowledge for the field, here we introduce SpliceProt, a new protein sequence repository of transcriptome experimental data used to investigate for putative splice variants in human proteomes. Current version of SpliceProt contains 159 719 non-redundant putative polypeptide sequences. The assessment of the potential of SpliceProt in detecting new protein isoforms resulting from alternative splicing was performed by using publicly available proteomics data. We detected 173 peptides hypothetically derived from splice variants, which 54 of them are not present in UniprotKB/TrEMBL sequence repository. In comparison to other protein sequence repositories, SpliceProt contains a greater number of unique peptides and is able to detect more splice variants. Therefore, SpliceProt provides a solution for the annotation of proteomics experiments regarding splice isofoms. The repository files containing the translated sequences of the predicted splice variants and a visualization tool are freely available at http://lbbc.inca.gov.br/spliceprot.

Received: February 22, 2013
Revised: October 3, 2013
Accepted: November 6, 2013

Keywords:
Alternative splicing / Bioinformatics

Additional supporting information may be found in the online version of this article at the publisher's web-site

Alternative splicing is a molecular mechanism capable of significantly enriching the proteomic repertoire of an organism through the generation of different transcripts from a single gene [1]. Moreover, it has been the goal of numerous studies due to its biological relevance in the diversity of organisms [2] and its association with many diseases [3]. More than 90% of human genes are affected by alternative splicing events [4], indicating a promising field to be explored in proteomics, because it is speculated that this mechanism influences at most the construction of polypeptide chains [5].

In addition, high-throughput proteomics experiments have shown protein sequences that have not been

Correspondence: Dr. Fabio Passetti, Bioinformatics Unit, Clinical Research Coordination, Instituto Nacional de Câncer (INCA), RJ, Brazil
E-mail: passetti@inca.gov.br

Abbreviations: Glu-C_bicarb, Glu-C in ammonium bicarbonate buffer; **Glu-C_phosph**, Glu-C in phosphate buffer; **HTS**, high-throughput sequencing; **FDR**, false discovery rate

Colour Online: See the article online to view Fig. 1 in colour.

characterized yet [6]. It is hypothesized that these sequences may be originated by alternative splicing events. In this context, some studies used different approaches to solve this problem, as specific databases [7–11], algorithms for searching spectra against genomic annotation databases [6, 12–15] or even the concentration of protein variants originated by alternative splicing events for the diagnostics of patients with cancer [16].

Furthermore, the advent of high-throughput sequencing (HTS) of genomes and transcriptomes [17] provided a large amount of information to be investigated. As these technologies can be applied for the study of alternative splicing [18], they became an important strategy associated with high-throughput proteomics experiments to discover new splice forms [10]. To address this issue, an effort of the chromosome-centric human proteome project provided new insights of splice forms in the human chromosome 17 using high throughput proteomic and transcriptomic data [19].

In this study, we present SpliceProt, a new sequence repository composed by predicted alternative spliced variants translated in silico and based on experimental human transcriptomic data. Dataset preparation, software utilized, and proof of concept are detailed in Supporting Information.

SpliceProt is based on ESTs, RefSeq transcripts, and HTS data obtained in the Sequence Read Archive (SRA) repository hosted at NCBI website (Figure 1). The ternary matrices methodology was applied to the collected transcripts to identify splice variants (a detailed explanation is provided in Supporting Information).

A total of 267 632 variants were identified and their sequences were reconstructed based on the genomic sequence. The translation of all variants was performed by TRANSEQ [20] and resulted in 161 915 polypeptide sequences and provided 159 719 (98.64%) unique sequences. We compared the SpliceProt sequences to those provided by RefSeq [21], ENSEMBL Gene v.69 [22], and UniProtKB/Swiss-Prot [23]. To perform this analysis, we excluded the redundant sequences in each repository and counted the number of identical sequences found in SpliceProt. Only sequences with at least 24 amino acid residues in length were considered in this comparison because this is the size of the smaller polypeptide sequence in RefSeq. The RefSeq dataset was chosen as the reference repository because it is composed of manually curated mRNA sequences and their encoded proteins. SpliceProt matched 23 770 sequences (91.98%) of the RefSeq protein dataset, while ENSEMBL Gene recovered 35 260 sequences (43.05%) and UniProtKB/Swiss-Prot matched 16 660 sequences (82.49%). Therefore, our repository recovered most of RefSeq sequences and its coverage is similar to the ENSEMBL Gene repository.

Comparing SpliceProt to ENSEMBL Gene, they shared 35 260 sequences, which represented 22.07 and 43.05% of each repository, respectively (Table 1). This difference may be explained by the processing of the transcripts prior to their translation. SpliceProt applies the ternary matrix methodology to reconstruct the transcripts according to their splicing patterns. In addition, the reading frames of all translated sequences in SpliceProt are computationally compared to manually curated sequence proteins of the corresponding gene provided by RefSeq.

The analysis of SpliceProt and UniProtKB/Swiss-Prot redundancy provided that they share 16 660 sequences, which represented 10.43 and 82.49% of each repository, respectively. However, 3535 sequences from UniProtKB/Swiss-Prot are not represented in SpliceProt, probably because UniProtKB/Swiss-Prot is composed by experimentally defined protein sequences, while SpliceProt contains the result of predicted translated transcripts (Table 1).

In order to evaluate the contribution of SpliceProt for MS experiments, the sequences were submitted to a computational digestion simulation of four enzymes: trypsin, lysine, Glu-C in ammonium bicarbonate buffer (Glu-C_bicarb), and Glu-C in phosphate buffer (Glu-C_phosph). Our analysis suggests that trypsin is the enzyme that permits the production of more unique peptides, followed by Glu-C in phosphate buffer, Glu-C in ammonium bicarbonate buffer, and lysine (Supporting Information). Based on these findings, we compared the computational tryptic digestion of all sequence repositories. A total of 700 492 nonredundant polypeptides ranging from 6 to 24 amino acids in length were generated from our repository. The same simulation was performed with RefSeq, ENSEMBL Gene, and UniProtKB/Swiss-Prot repositories and 486 532, 578 139 and 506 318 of nonredundant polypeptides were, respectively, generated (Table 1). Figure 2 presents the Venn diagram of redundant and exclusive peptides after the tryptic computational digestion for every sequence repository (for lysine, Glu-C_bicarb, and Glu-C_phosph see Supporting Information).

In comparison to RefSeq, the in silico trypsin digestion of our repository generated 480 878 peptides in common and added 219 614 unique peptides, which can be used to identify new potential protein isoforms. Comparing SpliceProt to ENSEMBL Gene, they had 513 678 fragments in common, while 186 814 peptides were exclusive from the former dataset and 64 461 from the latter. Therefore, our innovative method provided 2.8 times more unique trypsin digested polypeptides than the ENSEMBL Gene repository. When compared to UniProtKB/Swiss-Prot, 481 766 peptides were identical, 218 726 fragments were only in our dataset, and 24 552 only in the UniProtKB/Swiss-Prot results. At this time, our repository offers 8.9 more unique peptides than UniProtKB/Swiss-Prot.

One important step to show that SpliceProt will contribute to the scientific community is to use raw MS proteomics data to detect splice variants not present in other sequence repositories. We appended the nonredundant peptides of tryptic computational digestion of our sequence repository to UniProtKB/Swiss-Prot. Using this merged UniProtKB/Swiss-Prot/SpliceProt sequences, we analyzed as a case study the publicly available proteomics data from a human T lymphocyte cell line [10]. We detected 6,726 UniProtKB/Swiss-Prot protein sequences passing 5% false

Figure 1. Schema for the selection of splicing variants and their translation in silico.

discovery rate (FDR). To search for splice variants, we selected peptides passing both 1 and 5% FDR based on *q*-value and 1 and 5% FDR based on Posterior Error Probability , as previously described [10]. Using this approach, we detected 225 peptides from splice variants exclusively present in SpliceProt. As Fifty-three of 225 peptides were suspected to be sequence artifacts, they were discarded after manual inspection. Therefore, we could detect 173 peptides with high confidence and 54 of them were not found in the human set of UniProtKB/TrEMBL repository. In comparison to the study published by Sheynkman et al. [10], out of the 57 splice variant peptides detected by them, we found 10 in our analysis. Therefore, we conclude that our and Sheynkman's approach are complementary based on the fact that both methodologies to translate the transcriptome, use different computational assumptions. The complete set of results can be found in Supporting Information.

SpliceProt also offers a visualization tool to assist students and researchers to manually inspect the human genome regarding the computationally predicted splice pattern of a given gene and its translation (Supporting Information). Our methodology provides an alternative for researchers working in human proteomics that aims to investigate published and unpublished data regarding splice variants. Our effort aimed to provide a reliable human predicted splice variant protein repository to assist in the discovery of new molecular markers candidates to eventually improve diagnosis and prognosis, because alternative splicing of transcripts is not the focus of the main protein sequence repositories. To this end, we have used in our comparisons the RefSeq project, which is a manually curated sequence repository of transcripts and their translation. Therefore, we demonstrated that our methodology was able to reconstruct the transcriptome sequence based on the mapped transcripts onto the human genome. This approach did not alter their translation sequences when compared to the protein sequences provided by the RefSeq repository. Furthermore, SpliceProt provides a set of nonredundant sequences based on all types of splicing patterns, which enables a straightforward search of alternative spliced variants, increasing the chances of detection of splice variants in proteomic experiments analysis. Furthermore, the number of unique peptides generated by in silico trypsin digestion of all protein sequences contained in our repository allows more opportunities to distinguish splice isoforms.

Due to the very close association of alternative splicing and diseases [24], different strategies to discover new biomarkers and splicing variants have been performed [25, 26] but most of them do not use databases or repositories exclusively designed to take into account all alternative splicing types: exon skipping, intron retention, alternative 5′ splice site, and alternative 3′ splice site. We believe that the union of HTS projects with high-throughput proteomics experiments is a promising strategy for the identification of new splice isoforms. One of the current challenges in bioinformatics is to integrate the knowledge from these two worlds. In this direction, the chromosome-centric human proteome project studied the human chromosome 17 and could detect several not previously described spliced forms [19].

Table 1. Comparison between the protein sequences and computational enzymatic digestion of SpliceProt, RefSeq, ENSEMBL Gene, and UniProtKB/Swiss-Prot

Comparison	Protein number	Peptides produced by computational enzymatic digestion			
		Trypsin	Lysine	Glu-C_bicarb	Glu-C_phosph
SpliceProt vs. RefSeq					
Identical sequences	23 770	480 878	255 091	270 986	502 846
Unique SpliceProt	135 949	219 614	107 804	106 122	173 993
Unique RefSeq	2071	5654	2910	3371	5784
SpliceProt vs. ENSEMBL Gene					
Identical sequences	35 260	513 678	271 712	288 629	531 885
Unique SpliceProt	124 459	186 814	91 183	88 479	144 954
Unique ENSEMBL Gene	46 640	64 461	37 155	39 755	58 928
SpliceProt vs. UniProtKB/Swiss-Prot					
Identical sequences	16 660	481 766	255 064	271 037	504 709
Unique SpliceProt	143 059	218 726	107 831	106 071	172 130
Unique UniProtKB/Swiss-Prot	3535	24 552	12 562	13 501	24 861

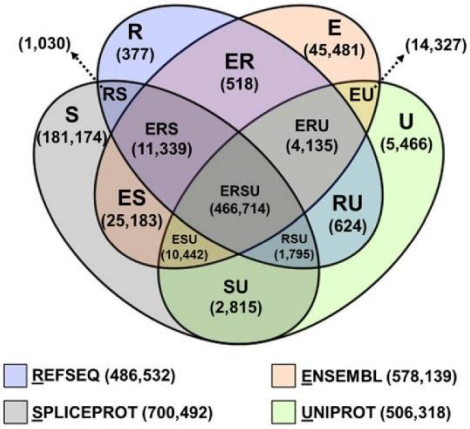

Figure 2. Venn diagram of the nonredundant set of peptides obtained from the trypsin computational digestion of SpliceProt, RefSeq, ENSEMBL Gene, and UniProtKB/Swiss-Prot.

Hereby, we offer a new repository of alternative splicing variants predicted in silico. We wish SpliceProt might contribute to the discovering of new biomarkers and new tools for diagnosis of cancer and other pathologies.

RT is supported by Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES) and Vice-Presidência de Ensino, Informação e Comunicação/Pró-Reitoria–IOC/FIOCRUZ. ELF and GE are supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). FP acknowledges the support of Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Ministério da Ciência e Tecnologia/Fundo Setorial de Saúde (MCT/CT-Saude), Departamento de Ciência e Tecnologia/Secretaria de Ciência, Tecnologia e Insumos Estratégicos/Ministério da Saúde (DECIT/SCTIE/MS) and CNPq. FP, CGF and the Bioinformatics Unit acknowledge the support of the Swiss Bridge Foundation, Fundação do Câncer and CAPES. The authors acknowledge the support of Natasha Andressa Nogueira Jorge and Gabriel Wajnberg.

Potential conflict of interest: FP, PSLO, CGF, and EA are authors or inventors of the patent number 699132 granted in Switzerland regarding the ternary matrix methodology used in the development of this work. However, the results achieved in this work using the patent's methodology have merely scientific interest.

References

[1] Hallegger, M., Llorian, M., Smith, C. W., Alternative splicing global insights. *FEBS J.* 2010, *277*, 856–866.

[2] Brett, D., Pospisil, H., Valcárcel, J., Reich, J., et al., Alternative splicing and genome complexity. *Nat. Genet.* 2002, *30*, 29–30.

[3] Wang, G. S., Cooper, T. A., Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* 2007, *8*, 749–761.

[4] Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., et al., Alternative Isoform regulation in human tissue transcriptomes. *Nature* 2008, *456*, 470–476.

[5] Black, D. L., Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* 2000, *103*, 367–370.

[6] Tanner, S., Shen, Z., Ng, J., Florea, L., et al., Improving gene annotation using peptide mass spectrometry. *Genome Res.* 2007, *17*, 231–239.

[7] Power, K. A., McRedmond, J. P., de Stefani, A., Gallagher, W. M., et al., High-throughput proteomics detection of novel splice isoforms in human platelets. *PloS One* 2009, *4*, e5001.

[8] Mo, F., Hong, X., Gao, F., Du, L., et al., A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. *BMC Bioinformatics* 2008, *9*, 537.

[9] Li, J., Su, Z., Ma, Z., Slebos, R. J., et al., A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Mol. Cell Proteomics* 2011, *10*, 1–11.

[10] Sheynkman, G. M., Shortreed, M. R., Frey, B. L., Smith, L. M., Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol. Cell Proteomics* 2013, *12*, 2314–2353.

[11] Wang, X., Slebos, R. J., Wang, D., Halvey, P. J., et al., Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* 2012, *11*, 1009–1017.

[12] Lee, S. W., Choi, J. P., Kim, H. J., Hong, J. M., et al., ASPMF: a new approach for identifying alternative splicing isoforms using peptide mass fingerprinting. *Biochem. Biophys. Res. Commun.* 2008, *377*, 253–256.

[13] Blakeley, P., Siepen, J. A., Lawless, C., Hubbard, S. J., Investigating protein isoforms via proteomics: a feasibility study. *Proteomics* 2010, *10*, 1127–1140.

[14] Chang, K. Y., Georgianna, D. R., Heber, S., Payne, G. A., et al., Detection of alternative splice variants at the proteome level in Aspergillus flavus. *J. Proteome Res.* 2010, *9*, 1209–1217.

[15] Hatakeyama, K., Ohshima, K., Fukuda, Y., Ogura, S., et al., Identification of a novel protein isoform derived from cancer-related splicing variants using combined analysis of transcriptome and proteome. *Proteomics* 2011, *11*, 2275–2282.

[16] Baron, A. T., Cora, E. M., Lafky, J. M., Boardman, C. H., et al., Soluble epidermal growth factor receptor (sEGFR/sErbB1) as a potential risk, screening, and diagnostic serum biomarker of epithelial ovarian cancer. *Cancer Epidemiol. Biomarkers Prev.* 2003, *12*, 103–113.

[17] Kircher, M., Kelso, J., High-throughput DNA sequencing—concepts and limitations. *Bioessays.* 2010, *32*, 524–536.

[18] Ning, K., Nesvizhskii, A. I., The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC Bioinformatics*. 2010, *11*, S14.

[19] Liu, S., Im, H., Bairoch, A., Cristofanilli, M., et al., A chromosome-centric human proteome project (C-HPP) to characterize the sets of proteins encoded in chromosome 17. *J. Proteome Res*. 2013, *12*, 45–57.

[20] Rice, P., Longden, I., Bleasby, A., EMBOSS: the european molecular biology open software suite. *Trends Genet* 2000, *16*, 276–277.

[21] Pruitt, K. D., Tatusova, T., Klimke, W., Maglott, D. R., NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res*. 2009, *37*, D32–D36.

[22] Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., et al., The Ensembl automatic gene annotation system. *Genome Res*. 2004, *14*, 942–950.

[23] The UniProt consortium, ongoing and future developments at the universal protein resource. *Nucleic Acids Res*. 2011, *39*, D214–D219.

[24] Brinkman, B. M., Splice variants as cancer biomarkers. *Clin. Biochem*. 2004, *37*, 584–594.

[25] Pampalakis, G., Scorilas, A., Sotiropoulou, G., Novel splice variants of prostate-specific antigen and applications in diagnosis of prostate cancer. *Clin. Biochem*. 2008, *41*, 591–597.

[26] Tress, M. L., Bodenmiller, B., Aebersold, R., Valencia, A., Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol*. 2008, *9*, R162.

Anexo 2 - Splice variants in the proteome: a promising and challenging field to targeted drug discovery.

Drug Discovery Today • Volume 00, Number 00 • November 2014 REVIEWS

ELSEVIER

Splice variants in the proteome: a promising and challenging field to targeted drug discovery

Reviews • INFORMATICS

Raphael Tavares[1], Nicole M. Scherer[1], Carlos G. Ferreira[2], Fabricio F. Costa[3,4] and Fabio Passetti[1]

[1] Bioinformatics Unit, Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rio de Janeiro, RJ 20231-050, Brazil
[2] Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rio de Janeiro, RJ 20231-050, Brazil
[3] Cancer Biology and Epigenomics Program, Ann and Robert Lurie Children's Hospital of Chicago Research Center and Department of Pediatrics, Northwestern University's Feinberg School of Medicine, Chicago, IL 60614, USA
[4] Genomic Enterprise, Chicago, IL 60614, USA

The advent and improvement of high-throughput sequencing over the past decade leveraged the study of whole genomes and transcriptomes of different organisms at lower costs. In transcriptomics, RNA-Seq expands our capacity to understand gene expression in different tissues and pathologies, and how alternative splicing might affect the final protein sequence. Here, we discuss the association of using transcriptome and proteome high-throughput data to foster drug discovery. Using this innovative strategy, some research groups have already identified computationally predicted novel peptides derived from putative splice variants in experimental human proteome data. These discoveries provide new opportunities for targeted drug development.

Introduction

Transcriptomics and proteomics have been extensively explored over the past decade as the improvement in RNA-Seq and mass spectrometry (MS) boosted a new high-throughput stage. Although these fields are represented by different molecular elements (RNAs and proteins, respectively), the link between them has been the goal of studies using MS, microarrays and expressed sequence tag (EST) data [1,2]. Recently, a new wave of studies has narrowed these two fields using RNA-Seq and MS data, offering new perspectives to correlate the levels of expression of mRNAs and proteins. For example, Nagaraj *et al.* [3] analyzed the transcriptome and the proteome of the HeLa cervical cancer cell line, and most of the 8609 identified genes were confirmed by both technologies.

The identification of proteins by MS provides a better understanding of coding-gene expression in different cell lines, tissues, and organisms. From sample preparation to protein identification, many issues have been discussed in the literature on the quantification, database usage, and reliability of statistical validation of the results obtained from this technology [4]. The complete proteome of an organism has been the focus of some research groups, and yeast has been used as a model organism for the improvement of proteomics techniques and computational strategies [5,6].

RNA-Seq emerged as a potential technology to be explored in transcriptomics, taking advantage of high-throughput sequencing and, consequently, of the large amount of data produced [7]. For example, RNA-Seq enabled the discovery of new potential splice isoforms [8,9] and the analysis of transcript expression levels [10] in different cell types and conditions. The methods of sequencing and the assembly of reads have been compared [11] and improved over the past few years, promoting new perspectives in transcriptomics [12], including noncoding RNAs (reviewed in [13]).

Focusing on biological challenges, alternative splicing stands out because of its contribution to increasing the proteomic diversity of organisms [14] and the consequences for pharmacogenomics [15], and protein function and structure (reviewed in [16]). In fact, it is known that even good-quality spectra can remain unidentified because their corresponding proteins are

Corresponding authors: Costa, F.F. (fcosta@luriechildrens.org), (fcosta@genomicenteprise.com), Passetti, F. (passetti@inca.gov.br)

1359-6446/06/$ - see front matter © 2014 Published by Elsevier Ltd. http://dx.doi.org/10.1016/j.drudis.2014.11.002

Please cite this article in press as: Tavares, R. et al. Splice variants in the proteome: a promising and challenging field to targeted drug discovery, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.11.002

not represented in the protein sequence database [17]. To address this issue, efforts such as the Chromosome-centric Human Proteome Project (C-HPP) provided new insights into splice forms in chromosomes using high-throughput proteomic and transcriptomic data [18,19].

Bioinformatics has been following the development of transcriptomics and proteomics, promoting new strategies and improvements for better use and analysis of data [20,21]. Considering the importance and advantages of the aforementioned technologies, here we discuss the combination of RNA-Seq and MS data for the identification alternative spliced isoforms. In addition, we discuss how the association of these two techniques could help in the development of drugs in a systems biology context.

Initiatives to identify novel splice variants

MS and RNA-Seq are technologies with different approaches and purposes, although they are complementary for verifying the balance between transcripts and proteins. Regarding alternative splicing, the large amount of data produced offers the opportunity to discover novel splicing variants in different tissues and conditions, and correlate the levels of expression between transcriptomic and proteomic results.

To the best of our knowledge, Ning and Nesvizhskii [22] presented the first study associating RNA-Seq and MS experiments to identify novel alternative splice forms in three different mouse tissues (brainstem, liver, and skeletal muscle). This study highlighted the combination of a canonical protein sequence database with sequences obtained from RNA-Seq data with the objective to create a customized database to be used in the MS experiment. To create this putative sequence database for MS search, all transcriptome data was translated in six-frames and were considered open reading frames (ORFs) with more than 30 amino acids in length. According to the authors, two to three peptides supporting novel alternative splicing junctions were identified per tissue and most of them could also be confirmed by ESTs. Despite the small number of novel peptides discovered, this study demonstrated how RNA-Seq and MS could be used to infer novel alternative splice variants. As a first effort, these results instigated improvements in subsequent studies.

The article by Menon *et al.* [23] summarizes how the discovery of alternative splice variants can be explored using RNA-Seq/MS experiments. As part of C-HPP, the group focused on protein-coding genes of chromosome 17, which include cancer-associated genes, such as breast cancer 1/2, early onset (*BRCA1/2*) and tumor protein 53 (*TP53*). Using three models of human breast cancer subtypes, the authors characterized comprehensively splice variants expressed in aggressive human epidermal growth factor receptor 2 (ERBB2)+ breast cancers. The gene *ERBB2/HER2* is closely related to this type of cancer because it is amplified in 20–30% of patients [24]. Breast cancer subtypes are currently classified based on the pathological markers, such as estrogen receptor (ER), progesterone receptor (PR), and ERBB2 expression [25]. Following this classification, the cell lines used were: SKBR3 [ERBB2+ (overexpression)/ER−/PR−; adenocarcinoma], SUM190 [ERBB2+ (overexpression)/ER−/PR−; inflammatory breast cancer], and SUM149 [ERBB2 (low expression) ER−/PR−; inflammatory breast cancer]. In terms of aggressiveness, SUM190 and SUM149 are the most lethal forms, whereas SKBR3 has better prognosis and has been used as a reference for a preclinical model and ERBB2-based therapies. Using RNA-Seq, a total of 4406 distinct transcripts were obtained from these three cell lines and 1052 (23.88%) splice variants were shared among them. Although a small percentage, shared variants should also be considered with care, because potential new drugs and therapies could be tested for all three types of cancer, avoiding high costs and preventing patients' physical exhaustion during chemotherapy. As well as *ERBB2*, the epidermal growth factor receptor (*ERRB1/EGFR*) gene is involved in breast cancers and splice variants of these two genes have been obtained and analyzed in these three cell types. Some of the ERBB2 and EGFR splice variants were exclusively expressed in certain cell lines. As a final step, quantitative proteomics data were produced from these cell lines and novel peptides were also discovered. According to the authors, to be considered as a novel peptide, a peptide could not be matched to any know protein sequence in their database (based on an adaptation of the ECGene [26]) but should have evidence in the RNA-Seq data.

More studies involving cancer cell lines, RNA-Seq, and MS have been published over the past few years. Fanayan *et al.* [27] analyzed three colorectal cancer (CRC) cell lines and, using these two technologies, identified cancer-associated proteins with differential expression patterns and a cancer biomarker. In a study by Lundberg *et al.* [28], proteins and transcripts of three different cancer cell lines were analyzed: bone osteosarcoma (U-2 OS), epidermoid squamous cell carcinoma (A-431), and brain glioblastoma (U-251 MG). Using MS, RNA-Seq and antibody-based immunofluorescence confocal microscopy (IF), the authors demonstrated that one third of the gene products were found by all the three techniques in all cell lines. The authors also observed differences in the levels of RNAs and proteins between the cell lines, suggesting control mechanisms for RNA, protein half-life, or differences in translational efficiency. The HeLa cell line has also been the target of two studies focused on identifying transcripts and proteins using RNA-Seq and MS, to quantify them and confirm gene expression [3,29].

Another study by Mazin *et al.* [30] investigated the splicing changes in human brain according to development stage and aging. Brain samples from prefrontal cortex (PFC) and cerebellar cortex (CBC) were collected from 35 individuals, with ages ranging from 2 days to 98 years. In a global view, approximately 40% of the genes pass through splicing change over the lifespan. High-throughput sequencing was important to verify that almost 30% of age-related splicing changes affect the protein-coding portion of the transcripts. In this way, MS verified the presence of these changes in proteomic data from the PFC of 12 healthy individuals. The results identified 19 genes showing significant age-related splicing changes and related these to Alzheimer's disease, stimulation of neurite outgrowth, and regulation of central nervous system (CNS) development. Although the transcriptomic and proteomic data are from different sources and only two brain regions were explored, this research has inspired more studies regarding the potential of splice isoform generation in brain tissue.

Except for the approach of Ning and Nesvizhskii [22], most studies used canonical sequences in the MS experiment analysis, reducing the chance of identifying new isoforms derived from alternative splicing. One strategy to overcome this limitation is to use RNA-Seq data to develop customized or sample-specific databases, which are discussed below.

Please cite this article in press as: Tavares, R. et al. Splice variants in the proteome: a promising and challenging field to targeted drug discovery, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.11.002

The abovementioned research studies applied established technologies to produce high-quality short RNA-Seq reads. Using a technology that enables sequencing of single RNA molecules, Sharon *et al.* [31] investigated the human transcriptome of 20 healthy tissues and organs to assess alternative spliced events. More than 10% of the mRNA molecules sequenced had a splicing pattern that was not previously described in the GENCODE database. These findings show that alternative splicing occurs at a low rate in healthy human tissues. However, because it has been described that alternative spliced transcripts are frequent in pathological states, including cancer, it would be interesting to use this technique to investigate the transcriptome diversity in disease states.

Drug discovery is usually focused on the identification of membrane proteins as potential targets. Some efforts have been taken in this field to assess the complete membrane proteome of an organ or organism [32–34]. Therefore, the identification of the complete membrane proteome could result in the discovery of novel splice variants that could be used as drug targets, including these cases of alteration in the expression pattern of splice variants of eukaryotic host cells while infected by an infectious agent. Omasits *et al.* [35] used RNA-Seq and high-throughput proteomics to address the complete proteome of the proteobacterium *Bartonella henselae*, including a fraction of the membrane. In their study, the authors detected the reorganization of the membrane proteome under two experimental conditions, presenting novel insights for the development of drugs targeting membrane proteins expressed in pathological states.

Database design using RNA-Seq to identify novel splice variants

Reference sequence databases, such as ENSEMBL, Uniprot/SwissProt, RefSeq, IPI, and NCBI, are commonly used in MS experiments to associate a given mass spectrum with a protein expressed in a sample. However, depending on the purpose of each study, this option is limited to known proteins and narrows the chances of discovering novel peptides and/or splicing variants [36].

The approach developed by Sheynkman *et al.* [8] associates canonical sequence databases with peptide sequences corresponding to novel splice-junctions derived from RNA-Seq data. Given that the proteomic and transcriptomic data were obtained from the same origin (Jurkat cells), a promising way of matching these two different sources of data was created. After discovering novel splice junctions, the authors translated them into peptides and appended these peptides to a file containing the Uniprot database. The MS experiment revealed 57 splice junction peptides that were not present in the Uniprot-Trembl proteomic database, including different splicing events. A sample-specific database fosters the identification of ordinary proteins and novel peptides, and provides a more realistic comparison between the levels of expression of transcripts and proteins.

Following a similar strategy of concatenating peptide sequences to reference sequence databases, Tavares *et al.* [9] designed a customized database (SpliceProt) comprising a reference database (Uniprot/SwissProt) and peptide sequences from dbEST, RefSeq, and RNA-Seq data. Using the same proteomic data and statistical assignment presented by Sheynkman *et al.* [8], the usage of the SpliceProt sequence repository yielded the identification of 54 novel peptides that were not present in the UniprotKB/TrEMBL data set. However, only ten peptides were identical in both studies. These complementary studies are important because they demonstrate how many peptides are shared between a general and a specific database using the same proteomic data. Despite the relevance of databases based on RNA-Seq data, the dbEST data set is still seen as a useful source.

The strategy for dealing with transcriptomic and proteomic data from the same origin or distinct sources is depicted in Fig. 1 and the current statistical assignments for peptide identification for splice variant detection are presented in Table 1.

Splice variants in the proteome

Constructing new strategies to identify novel splice variants is one of the first steps to explore their potential role in cell behavior. In addition, there is a link between tissue-specific, condition-specific, and alternative splicing that pushes studies to understand significant functional differences from normally expressed isoforms compared with those specifically expressed. If a novel tumor-specific splice variant has an important role in the development of a disease (e.g. cancer), it can be a subject for drug design studies.

The estrogen receptor 1 (*ESR1*) gene expresses estrogen receptor alpha (ERα), which binds to estradiol and is upregulated in most types of breast cancer, resulting in a hormone dependence for tumor growth [37]. On the basis of this relation between receptor and hormone, one of the strategies for treatment is to block their binding using a molecule (e.g., tamoxifen) and then avoiding activation of the receptor. Unfortunately, this endocrine therapy does not work for approximately 40% of patients who are treated with tamoxifen [38]. Although the ERα structure has been experimentally obtained with estradiol [39] or tamoxifen [40], there is a lack of studies directed to elucidate the structure and behavior of transcript variants derived from alternative splicing. Depending on their function and stability, new drugs could be designed to act on these specifically.

HER2/ERBB2 amplification and overexpression is related to poor prognosis in breast cancer because these genes are associated with an aggressive phenotype [41]. Menon *et al.* [42] performed a study to predict the structural and functional consequences of alternative splice variants in mouse Her2/neu-induced breast cancer. In addition to detecting the presence of splice variants in normal and tumor samples, the comparison between them demonstrated structural and functional differences that could be used to characterize new cancer biomarkers. For example, it is known that Annexin 6 is correlated with cancer and its post-translational phosphorylation is associated with cell growth. Among the variants analyzed, the variant anxa6-001 is more prone to phosphorylation compared with the variant anxa6-002. As a result of an alternative splicing event, two amino acids (Thr-535 and Ser-537) are shifted to the inner part of the loop region in anxa6-001. This study shows the importance of inferring the structure of splice variants for more genes associated with cancer, with the aim of designing new drugs for specific variants for aggressive tumors. In addition, this could help to determine whether alternative isoforms interfere in the effectiveness of different treatments.

Cancer maintenance is closely related to apoptosis and this theme is well represented by *BCL2L1*, which has alternative splicing isoforms with antagonist functions. *BCL2L1* can express the isoforms BCL-XL and BCL-XS with anti- and proapoptotic functions, respectively [43]. The second isoform lacks part of exon 2,

Please cite this article in press as: Tavares, R. et al. Splice variants in the proteome: a promising and challenging field to targeted drug discovery, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.11.002

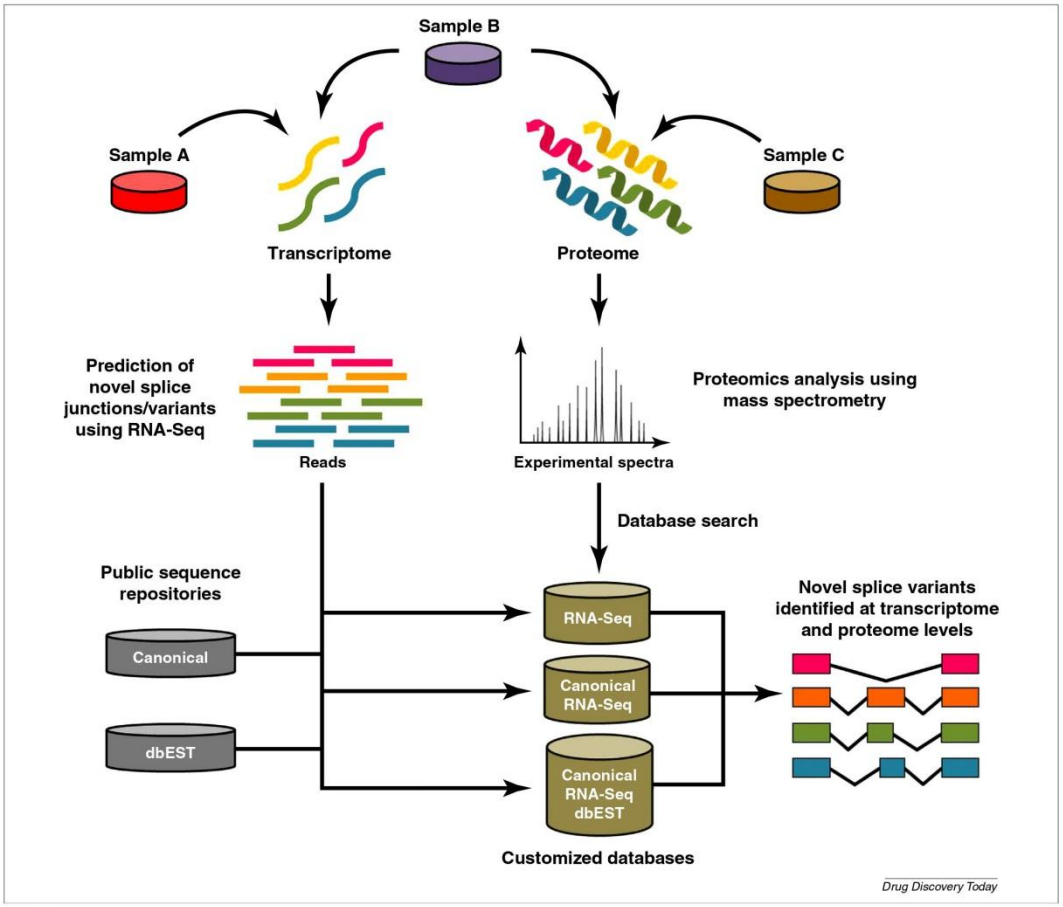

FIGURE 1

Representative scheme of RNA-Seq and mass spectrometry (MS) studies: transcriptomic and proteomic samples can be used from the same origin (sample B) or from different sources (samples A and C). Such a scheme can result in a combination of customized databases using RNA-Seq data and public proteomic repositories.

TABLE 1

List of selected studies using RNA-Seq and MS to discover novel alternative splice variants

Organism	Transcriptomic source	Proteomic source	Database design	Statistical assignment for peptide identification[a]	Refs
Homo sapiens	Jurkat cell line (TIB-152)	Jurkat cells	Customized	FDR and PEP	[8]
	Public databases (dbEST and SRA)	Jurkat cells	Customized	FDR and PEP	[9]
	Breast cancer cell lines (SKBR3, SUM190, and SUM149)	Breast cancer cell lines (SKBR3, SUM190 and SUM149)	Customized	FDR	[23]
	Samples from PFC and CC	Samples from PFC	Customized	FDR	[30]
Mus musculus	Mitochondrial proteome from brainstem and liver	Mitochondrial proteome from brainstem and liver	Canonical (IPI)	FDR	[36]
	Mitochondrial proteome from brainstem, liver, and skeletal muscle	Mitochondrial proteome from brainstem, liver, and skeletal muscle	Customized	FDR	[22]

[a] FDR, false discovery rate; PEP, posterior error probability.

Please cite this article in press as: Tavares, R. et al. Splice variants in the proteome: a promising and challenging field to targeted drug discovery, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.11.002

which contains specific domains to avoid programed cell death. Recent docking studies have evaluated the behavior between BCL-XL and other molecules. *Panax ginseng*, an economically important natural plant used in oriental medicine, produces active compounds known as ginsenosides. In a study by Sathishkumar *et al.* [44], different types of ginsenoside showed binding affinity to the antiapoptotic proteins BCL-2, BCL-XL, and MCL-1, suggesting new agents for use in cancer therapy. Zhou *et al.* [45] used small-molecule inhibitors of BCL-2 and BCL-XL to evaluate apoptosis in a cell lineage of small cell lung cancer. Two synthesized compounds showed affinity to BCL-2/BCL-XL and effectively induced apoptosis in H146 cancer cells.

Cancer growth is also closely related to the vascular endothelial growth factor A (*VEGF-A*) gene, which has a key role in angiogenesis. This gene expresses two families of isoform: proangiogenic and antiangiogenic, classified as VEGFxxx and VEGFxxxb, respectively [46]. Consequently, the balance between these two isoforms interferes in cancer development, such as in neuroblastoma, where the antiangiogenic isoform is upregulated at the levels of mRNA and the resulting protein [47].

Androgen receptor (AR) activation is an important factor for prostate cancer cells because it is responsible for cancer maintenance and survival. This known correlation was demonstrated first by Huggins and Hodges [48] as a result of decreasing androgenic activity through castration or estrogenic injections. Such treatment involves testosterone-lowering hormonal therapy and surgical castration; however, when these strategies do not have effective results and prostate cancer continues to advance, the disease is termed 'castration-resistant prostate cancer' [49]. The reasons for this advance involve not only AR, but also the participation of the glucocorticoid receptor (GR), which is involved with many signaling pathways as well as other nuclear receptors [50]. With the structures of both AR and GR elucidated [51,52], alternative splicing could interfere with the resistance of prostate cancer [53,54], reinforcing the need for more studies correlating the protein structure of splice variants and treatment responsiveness.

Although thousands of splice variants have already been described at the mRNA level, only a few protein structures of these isoforms have been experimentally determined. In 2007, Birzele *et al.* [55] reported that less than ten isoforms were available in the Protein Data Bank (PDB), whereas an article by Omenn *et al.* [56] in 2013 added that only seven full-length pairs were in this set. To collect all the human proteins that have a 3D structure of at least two isoforms in the PDB, Hegyi *et al.* [57] performed an exhaustive search and selected those pairs of matches where the major and a minor isoform of the same protein could be mapped to different PDB entries. They ended up with 15 pairs of PDB entries, from 14 distinct proteins, one of which had three isoforms. Therefore, the studies we describe below were performed with proteins from *Rattus norvegicus*, *Bos taurus*, and *Gallus gallus*, which are orthologs of human genes.

It is well known that several proteins containing laminin/neurexin/SHBG-like (LNS) domains, such as neurexins and agrin, are functionally modulated by alternative splicing [58]. In the brain, alternative splicing in the LNS domain of the presynaptic cell surface adhesion molecule neurexin generates selective preference for postsynaptic neuroligin isoforms. Shen *et al.* [59] solved the crystal structure of rat neurexin 1β (n1β) with and without its natural 30 amino acid alternative splice insert in the presence of Ca^{2+}. This special insert at the site SS#4 in n1β favored binding to neuroligin 2, promoting GABAergic over glutamatergic synaptogenesis, induced major rearrangements to a key portion of the hypervariable surface and also increased Ca^{2+} binding affinity. The authors proposed that the alternative splicing of β neurexins generates distinctly different binding surfaces that discriminate between different neuroligins splice isoforms.

In neuromuscular junctions, alternative splicing in the LNS globular domains of the protein agrin also regulate function. Agrin, a multidomain heparan sulfate proteoglycan, has three LNS globular domains (G1–G3). A splice insert encoding four residues (KSRK) at splice site A (within the G2 domain) is necessary for agrin to bind heparin, whereas splice inserts leading to the incorporation of 8, 11, or 19 residues at the B splice site (G3 domain) are specific to neural cells and are required for the induction of postsynaptic acetylcholine receptors. Stetefeld *et al.* [60] determined the structures of three splice variants in the B splice site of the G3 domains accompanying Ca^{2+} binding. The structures of the Ca^{2+}-bound splice variants containing eight (G3–B8) and 11 (G3–B11) residue inserts were determined by X-ray crystallography, whereas the variant with no insert (G3-B0) was determined by nuclear magnetic resonance (NMR). The authors also solved structures without Ca^{2+}. The authors concluded that the plasticity of the B inserts might enable the agrin-G3 domain to discriminate between binding partners through an induced-fit mechanism.

Another example of alternative splicing affecting synapses has been shown for the neuronal protein Piccolo, which is involved in calcium-dependent signaling events in the presynaptic active zone. Using NMR spectroscopy and biochemical techniques, Garcia *et al.* [61] demonstrated that a nine-residue splice insert in the C_2A calcium-binding domain of Piccolo induced rearrangements in its β-sandwich fold and resulted in a decreased ability to bind calcium ions and phospholipids.

Tools related to RNA-Seq/MS and splice variants

The visualization of transcriptomic and proteomic data is also a challenge in terms of their integration. Nam and colleagues [62] developed a tool that enables the covisualization of RNA-Seq and MS data to identify alternative splicing in mRNA sequences. The possibility of locating an identified peptide in a transcript region enables researchers to formulate hypothesis about the importance of that peptide to the protein structure. Besides visualization, quantification of proteins is another important feature in proteomics experiments and can be performed by SpliceVista [63] and CAPER [64]. Wang and Zhang [65] developed a package in the R project for statistical computing that can be used to generate customized databases using RNA-Seq data for proteomics search. The tool also handles single nucleotide variations (SNVs), short insertion and deletions (INDELS), and novel junctions.

Concluding remarks

The improvement of high-throughput transcriptomics and proteomics has provided new links between gene expression and protein function. When integrating different 'omics' data, systems biology incorporates new perspectives about how a cell

Please cite this article in press as: Tavares, R. et al. Splice variants in the proteome: a promising and challenging field to targeted drug discovery, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.11.002

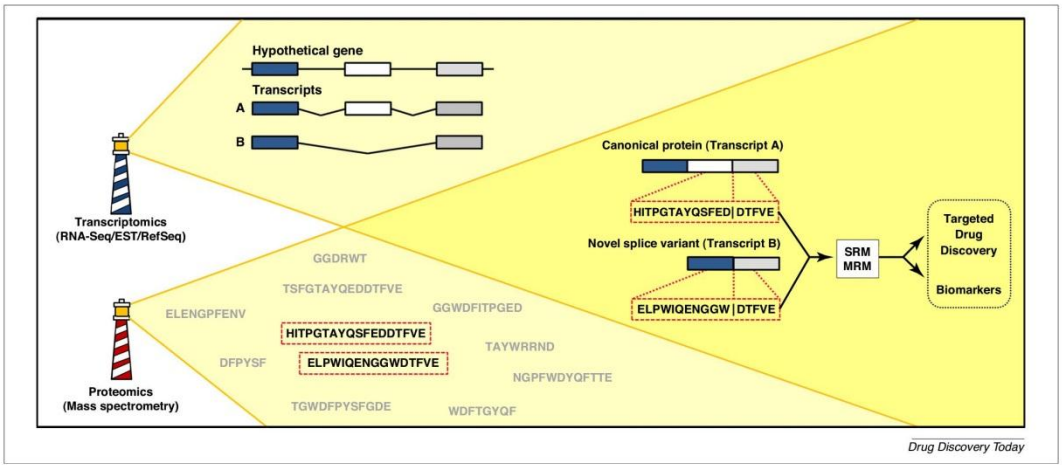

FIGURE 2

How to discover novel peptides shared between splice variants and proteotypic peptides using RNA-Seq and/or mass spectrometry (MS) studies. Given that these peptides correspond to isoforms, they can be applied to targeted drug discovery and as biomarkers in different pathologies.

is orchestrated from its genetic code to complex pathways. However, there are steps that have not yet been done or are still being improved (reviewed in [66]).

The importance of using RNA-Seq and MS experiments to look for novel peptides relies on the huge amount of data produced by both techniques and on the biological relevance of alternative splicing in terms of protein structure and function. The challenges are to confirm the expression of protein-coding genes and to characterize unique and specific peptides (also known as proteotypic peptides [67]) of splice variants in different cell lines, tissues, and organisms. The concept of proteotypic peptides has been expanded to proteome maps, which use the information from transcript variants for the association of proteome expression and post-translational modifications (reviewed in [68]). In this context, selected reaction monitoring (SRM) can be used to search for proteotypic peptides detected in large screening high-throughput proteomics of a group of samples, focusing on the diagnosis and prognosis of many diseases, including cancer. If multiple proteotypic peptides from splice variants are identified and selected for investigation in a group of patient samples, an extrapolation of SRM can be used: multiple reaction monitoring (MRM) (reviewed in [6]).

Regarding drug discovery, we expect that the strategy of using RNA-Seq and MS techniques could focus on membrane proteomes aimed at identifying novel proteins with higher expression levels in eukaryotic host cells infected with pathogens or in neoplasic tissues. Figure 2 depicts the usage of RNA-Seq experiments in conjunction with high-throughput proteomics for the identification of proteotypic peptides focused on targeted drug discovery.

Although there are only a few studies combining RNA-Seq and MS technologies, this association provides a promising strategy to link the transcriptomic and proteomic worlds. As a new approach to discover novel peptides, there are challenges to overcome involving RNA-Seq data treatment, database composition, and statistical assignment of the peptides found.

After many years of transcriptome studies and many described splice variants, recent articles have depicted the importance of detecting these in the proteome. The next step will be to investigate their function and role in current drug therapies. If the novel splice forms have an important role in disease maintenance or progression, targeted drug design could be used to improve treatment, directly impacting patients in the clinic.

Acknowledgments

The authors acknowledge an anonymous reviewer for critically reading the manuscript and suggesting substantial improvements. R.T. is supported by Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES) and Vice-Presidência de Ensino, Informação e Comunicação/Pró-Reitoria – IOC/FIOCRUZ. F.P. acknowledges the support of Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). F.P., C.G.F., N.M.S. and the Bioinformatics Unit acknowledge the support of the Fundação do Câncer and the Ministry of Health of Brazil. F.F.C. acknowledges the support of the Maeve McNicholas Memorial Foundation.

References

1 Edwards, N.J. (2007) Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol. Syst. Biol.* 3, 1–7

2 Wilson, L.O.W. *et al.* (2013) A novel splicing outcome reveals more than 2000 new mammalian protein isoforms. *Bioinformatics* 30, 1–6

Please cite this article in press as: Tavares, R. et al. Splice variants in the proteome: a promising and challenging field to targeted drug discovery, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.11.002

3 Nagaraj, N. *et al.* (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* 7, 548
4 Granholm, V. and Käll, L. (2011) Quality assessments of peptide-spectrum matches in shotgun proteomics. *Proteomics* 11, 1086–1093
5 de Godoy, L.M.F. *et al.* (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 455, 1251–1255
6 Picotti, P. (2013) A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* 494, 266–270
7 Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63
8 Sheynkman, G.M. *et al.* (2013) Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol. Cell. Proteomics* 12, 2341–2353
9 Tavares, R. *et al.* (2014) SpliceProt: a protein sequence repository of predicted human splice variants. *Proteomics* 14, 181–185
10 Trapnell, C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515
11 Ruffalo, M. *et al.* (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* 27, 2790–2796
12 McGettigan, P.A. (2013) Transcriptomics in the RNA-seq era. *Curr. Opin. Chem. Biol.* 17, 4–11
13 Jorge, N.A.N. *et al.* (2012) Bioinformatics of cancer ncRNA in high throughput sequencing: present state and challenges. *Front. Genet.* 3, 287
14 Black, D.L. *et al.* (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* 103, 367–370
15 Passetti, F. and Costa, F.F. (2009) The impact of microRNAs and alternative splicing in pharmacogenomics. *Pharmacogenomics J.* 9, 1–13
16 Kelemen, O. *et al.* (2013) Function of alternative splicing. *Gene* 514, 1–30
17 Ning, K. *et al.* (2010) Computational analysis of unassigned high quality MS/MS spectra in proteomic datasets. *Proteomics* 10, 2712–2718
18 Liu, S. *et al.* (2013) A chromosome-centric human proteome project (C-HPP) to characterize the sets of proteins encoded in chromosome 17. *J. Proteome Res.* 12, 45–57
19 Wang, Q. *et al.* (2013) Qualitative and quantitative expression status of the human chromosome 20 genes in cancer tissues and the representative cell lines. *J. Proteome Res.* 12, 151–161
20 Käll, L. and Vitek, O. (2011) Computational mass spectrometry-based proteomics. *PLoS Comput. Biol.* 7, e1002277
21 Costa, F.F. (2014) Big data in biomedicine. *Drug Discov. Today* 19, 433–440
22 Ning, K. and Nesvizhskii, A.I. (2010) The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC Bioinformatics* 11 (Suppl 1), S14
23 Menon, R. *et al.* (2013) Distinct splice variants and pathway enrichment in the cell-line models of aggressive human breast cancer subtypes. *J. Proteome Res.* 13, 212–227
24 Slamon, D. *et al.* (1987) Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 235, 177–182
25 Rakha, E.A. *et al.* (2010) Combinatorial biomarker expression in breast cancer. *Breast Cancer Res. Treat.* 120, 293–308
26 Menon, R. *et al.* (2009) Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. *Cancer Res.* 69, 300–309
27 Fanayan, S. *et al.* (2013) Proteogenomic analysis of human colon carcinoma cell lines LIM1215, LIM1899, and LIM2405. *J. Proteome Res.* 12, 1732–1742
28 Lundberg, E. *et al.* (2010) Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* 6, 450
29 Higareda-Almaraz, J.C. *et al.* (2013) Analysis and prediction of pathways in HeLa cells by integrating biological levels of organization with systems-biology approaches. *PLoS ONE* 8, e65433
30 Mazin, P. *et al.* (2013) Widespread splicing changes in human brain development and aging. *Mol. Syst. Biol.* 9, 633
31 Sharon, D. *et al.* (2013) A single-molecule long-read survey of the human transcriptoma. *Nat. Biotechnol.* 31, 1009–1017
32 Fischer, F. *et al.* (2006) Toward the complete membrane proteome. *Mol. Cell. Proteomics* 5, 444–453
33 Wisniewski, J.R. (2009) Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J. Proteome Res.* 8, 5674–5678
34 Khanna, M.R. *et al.* (2010) Towards a membrane proteome in *Drosophila*: a method for the isolation of plasma membrane. *BMC Genomics* 11, 1471–2164
35 Omasits, U. *et al.* (2013) Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome. *Genome Res.* 23, 1916–1927
36 Ning, K. *et al.* (2013) Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J. Proteome Res.* 11, 2261–2271
37 Platet, N. *et al.* (2004) Estrogens and their receptors in breast cancer progression: a dual role in cancer proliferation and invasion. *Crit. Rev. Oncol. Hematol.* 51, 55–67
38 Normanno, N. *et al.* (2005) Mechanisms of endocrine resistance and novel therapeutic strategies in breast cancer. *Endocr. Relat. Cancer* 12, 721–747
39 Brzozowski, A.M. *et al.* (1997) Molecular basis of agonism and antagonism in the oestrogen receptor target genes. *Nature* 389, 753–758
40 Shiau, A.K. *et al.* (1998) The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* 95, 927–937
41 Mitri, Z. *et al.* (2012) The HER2 receptor in breast cancer: pathophysiology, clinical use, and new advances in therapy. *Chemother. Res. Pract.* 2012, 743193
42 Menon, R. *et al.* (2012) Functional implications of structural predictions for alternative splice proteins expressed in Her2/neu-induced breast cancers. *J. Proteome Res.* 10, 5503–5511
43 Petros, A.M. *et al.* (2001) Solution structure of the antiapoptotic protein bcl-2. *Proc. Natl. Acad. Sci. U. S. A.* 98, 3012–3017
44 Sathishkumar, N. *et al.* (2012) Molecular docking studies of anti-apoptotic BCL-2, BCL-XL, and MCL-1 proteins with ginsenosides from Panax ginseng. *J. Enzyme Inhib. Med. Chem.* 27, 685–692
45 Zhou, H. *et al.* (2013) Structure-based design of potent Bcl-2/Bcl-xL inhibitors with strong *in vivo* antitumor activity. *J. Med. Chem.* 55, 6149–6161
46 Biselli-Chicote, P.M. *et al.* (2012) VEGF gene alternative splicing: pro- and anti-angiogenic isoforms in cancer. *J. Cancer Res. Clin. Oncol.* 138, 363–370
47 Peiris-Pagès, M. *et al.* (2012) Balance of pro-versus anti-angiogenic splice isoforms of vascular endothelial growth factor as a regulator of neuroblastoma growth. *J. Pathol.* 222, 138–147
48 Huggins, C. and Hodges, C.V. (1941) Studies on prostatic cancer. I. The effect of castration, of estrogen and of androgen injection on serum phosphatases in metastatic carcinoma of the prostate. *Cancer Res.* 1, 293–297
49 Egan, A. *et al.* (2014) Castration-resistant prostate cancer: adaptive responses in the androgen axis. *Cancer Treat. Rev.* 40, 426–433
50 Lu, N.Z. *et al.* (2006) International Union of Pharmacology. LXV. The pharmacology and classification of the nuclear receptor superfamily: glucocorticoid, mineralocorticoid, progesterone, and androgen receptors. *Pharmacol. Rev.* 58, 782–797
51 Pereira de Jésus-Tran, K. *et al.* (2006) Comparison of crystal structures of human androgen receptor ligand-binding domain complexed with various agonists reveals molecular determinants responsible for binding affinity. *Protein Sci.* 15, 987–999
52 Kauppi, B. *et al.* (2003) The three-dimensional structures of antagonistic and agonistic forms of the glucocorticoid receptor ligand-binding domain: RU-486 induces a transconformation that leads to active antagonism. *J. Biol. Chem.* 278, 22748–22754
53 Cao, B. *et al.* (2014) Androgen receptor splice variants activating the full-length receptor in mediating resistance to androgen-directed therapy. *Oncotarget* 5, 1646–1656
54 Isikbay, M. *et al.* (2014) Glucocorticoid receptor activity contributes to resistance to androgen-targeted therapy in prostate cancer. *Horm. Cancer* 5, 72–89
55 Birzele, F. *et al.* (2008) Alternative splicing and protein structure evolution. *Nucleic Acids Res.* 36, 550–558
56 Omenn, G.S. *et al.* (2013) Innovations in proteomic profiling of cancers: alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology. *J. Proteomics* 90, 28–37
57 Hegyi, H. *et al.* (2011) Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. *Nucleic Acids Res.* 39, 1208–1219
58 Rudenko, G. *et al.* (1999) The structure of the ligand-binding domain of neurexin Ibeta: regulation of LNS domain function by alternative splicing. *Cell* 99, 93–101
59 Shen, K. *et al.* (2008) Regulation of neurexin 1beta tertiary structure and ligand binding through alternative splicing. *Structure* 16, 422–431
60 Stetefeld, J. *et al.* (2004) Modulation of agrin function by alternative splicing and Ca^{2+} binding. *Structure* 12, 503–515
61 Garcia, J. *et al.* (2004) A conformational switch in the Piccolo C2A domain regulated by alternative splicing. *Nat. Struct. Mol. Biol.* 11, 45–53
62 Nam, C. *et al.* (2014) Tools to covisualize and coanalyse proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. *J. Proteome Res.* 13, 84–98

Please cite this article in press as: Tavares, R. et al. Splice variants in the proteome: a promising and challenging field to targeted drug discovery, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.11.002

63 Zhu, Y. *et al.* (2014) SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. *Mol. Cell. Proteomics* 13, 1552–1562
64 Guo, F. *et al.* (2013) CAPER: a chromosome-assembled human proteome browsER. *J. Proteome Res.* 12, 179–186
65 Wang, X. and Zhang, B. (2013) customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* 29, 3235–3237
66 Berg, E.L. (2014) Systems biology in drug discovery and development. *Drug Discov. Today* 19, 113–125
67 Kuster, B. *et al.* (2005) Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* 6, 577–583
68 Ahrens, C.H. *et al.* (2010) Generating and navigating proteome maps using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* 11, 789–801

Please cite this article in press as: Tavares, R. et al. Splice variants in the proteome: a promising and challenging field to targeted drug discovery, Drug Discov Today (2014), http://dx.doi.org/10.1016/j.drudis.2014.11.002

Anexo 3 - Unveiling alterative splice diversity from human oligodendrocyte proteome data. (*no prelo*)

Journal of Proteomics xxx (2016) xxx-xxx

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com

Unveiling alterative splice diversity from human oligodendrocyte proteome data

Raphael Tavares,[a, b] Gabriel Wajnberg,[a, b] Nicole de Miranda Scherer,[b] Bianca Alves Pauletti,[c] Juliana S. Cassoli,[g] Carlos Gil Ferreira,[d] Adriana Franco Paes Leme,[c] Patricia Savio de Araujo-Souza,[e, f] Daniel Martins-de-Souza,[g] Fabio Passetti [a, b, *]

[a] *Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute, Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, RJ, Brazil*
[b] *Bioinformatics Unit, Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rio de Janeiro, RJ, Brazil*
[c] *Laboratório de Espectrometria de Massas, Laboratório Nacional de Biociências (LNBio), CNPEM, Campinas, SP, Brazil*
[d] *Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rio de Janeiro, RJ, Brazil*
[e] *Department of Immunobiology, Fluminense Federal University (UFF), Niterói, RJ, Brazil*
[f] *Program of Cellular Biology, Instituto Nacional de Câncer (INCA), Rio de Janeiro, RJ, Brazil*
[g] *Laboratory of Neuroproteomics, Department of Biochemistry and Tissue Biology, Institute of Biology, University of Campinas (UNICAMP), Campinas, SP, Brazil*

ARTICLE INFO

Article history:
Received 13 October 2015
Received in revised form 14 May 2016
Accepted 20 May 2016
Available online xxx

Keywords:
Alternative splicing
Bioinformatics
Oligodendrocytes
Proteogenomics
Proteomics

ABSTRACT

Oligodendrocytes produce and maintain the myelin sheath of axons in the central nervous system. Because misassembled myelin sheaths have been associated with brain disorders such as multiple sclerosis and schizophrenia, recent advances have been made towards the description of the oligodendrocyte proteome. The identification of splice variants represented in the proteome is as important as determining the level of oligodendrocyte-associated proteins. Here, we used an oligodendrocyte proteome dataset deposited in ProteomeXchange to search against a customized protein sequence file containing computationally predicted splice variants. Our approach resulted in the identification of 39 splice variants, including one variant from the GTPase *KRAS* gene and another from the human glutaminase gene family. We also detected the mRNA expression of five selected splice variants and demonstrated that a fraction of these have canonical proteins that may participate in direct protein-protein interactions. In conclusion, we believe our findings contribute to the molecular characterization of oligodendrocytes and may encourage other research groups working with central nervous system disorders to investigate the biological significance of these splice variants. The splice variants identified in this study may encode proteins that could be targeted in novel treatment strategies and diagnostic methods.

Significance

Several disorders of the central nervous system (CNS) are associated with misassembled myelin sheaths, which are produced and maintained by oligodendrocytes (OL). Recently, the OL proteome has been explored to identify key proteins and molecular functions associated with CNS disorders. We developed an innovative approach to select, with a higher level of confidence, a relevant list of splice variants from a proteome dataset and detected the mRNA expression of five selected variants: *EEF1D*, *KRAS*, *MFF*, *SDR39U1*, and *SUGT1*. We also described splice variants extracted from OL proteome data. Among the splice variants identified, some are from genes previously linked to CNS and related disorders. Our findings may contribute to oligodendrocyte characterization and encourage other research groups to investigate the biological role of splice variants and to improve current treatments and diagnostic methods for CNS disorders.

1. Introduction

The large amount of mass spectrometry (MS) proteomic datasets available in public repositories encourages the reuse of these experimental data in novel analyses [1]. Scientific journals usually recommend using the PRoteomics IDEntifications (PRIDE) database [2] for proteomics experimental data storage. The analysis of MS proteomics data requires a protein sequence repository to annotate spectra. Current MS search algorithms use sequences of known proteins to computationally generate theoretical mass spectra that match experimental spectra. The Universal Protein Resource (UniProt) [3] and Reference Sequence Database (RefSeq) [4] are examples of protein sequences used to accomplish this task. This analysis is usually followed by functional annotation using several bioinformatic tools of functional network analysis, biological network analysis, and data mining [5]. One drawback of this approach is the restricted proteome coverage available in these protein sequence repositories. Therefore, mass spectra derived from novel protein splice variants are usually not identified when reference sequence databases are used.

Alternative mRNA splicing, by which the same gene may produce distinct protein sequences due to the differential use of exons and splice sites, is one of the major sources of protein sequence diversity [6,7]. Alternative splicing events in the transcriptome can be identified through alignment of transcripts onto a reference genome sequence, followed by identification of exons and splice junctions and

* Corresponding author at: Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute, Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, RJ, Brazil.
Email address: passetti@fiocruz.br (F. Passetti)

http://dx.doi.org/10.1016/j.jprot.2016.05.023
1874-3919/

comparison with canonical sequences. Reconstructed transcript sequences are used to predict putative protein sequences for proteomic analyses. Splice variant identification in proteome data is based on the presence of prototypic peptides that define a specific splice variant in sequence databases. For instance, MS data-based identification of alternatively spliced variants in pancreatic cancer has been successfully performed using the ECgene splice variant database as a reference [8,9,10]. In addition, PeptideProphet [11] and some plugins were used to identify human epidermal growth factor (HER2) splice variants in breast cancer mouse models [10], supporting the screening for splice variants in MS data.

Many studies have explored different strategies to detect splice variants and improve MS-based proteomic data annotation [12]. SpliceProt is one such strategy developed by our research group. SpliceProt uses a protein sequence repository comprised by the *in silico* translation of hypothetical alternative spliced variants detected in transcriptome data [13]. Here, we take advantage of SpliceProt and UniProt/SwissProt repositories to characterize splice variants in a publicly available human oligodendrocyte proteome data generated from the MO3.13 immortalized cell line [14].

Oligodendrocytes (OLs) are glia cells that produce and maintain the myelin sheath around the axons of the central nervous system (CNS). The myelin sheath is an electrical insulating proteolipid layer that transmits nervous impulses to from the cell body through the axon. Damaging or losing this layer may affect axonal function, which is observed in severe brain disorders such as multiple sclerosis, amyotrophic lateral sclerosis (ALS), and schizophrenia [15]. We conducted our study using proteome raw data produced from the MO3.13, a cell line commonly used as a model in OL studies (http://proteomecentral.proteomexchange.org/dataset/PXD000263), and a proteogenomics approach (reviewed by [16]) with focus on the detection of transcriptional splice variants.

2. Material and methods

2.1. Database setup

For proteome annotation, we prepared a customized protein sequence database based on the combination of canonical isoforms from the UniProt/SwissProt repository (version 11_2014) and alternative splice isoforms from UniProt/SwissProt and SpliceProt (version 1.1) repositories. To this end, we selected UniProt/SwissProt and SpliceProt isoforms and performed *in silico* trypsin digestion with a modified version of Digest software (EMBOSS package version 6.3.1). Non-redundant peptide sequences were searched against the canonical sequences and those identical were purged. Peptides not present in any canonical sequence were appended to a FASTA format file together with UniProt/SwissProt canonical protein sequences (Fig. 1). Because most peptides appended in our database can be shared by more than one isoform from the same gene, we were not able to identify all splice variants with proteotypic peptides. However, in most cases, a given peptide was exclusively associated with one or more splice variants from the same gene (Supporting information Fig. S1). The non-redundant customized database had 20,150 canonical sequences, 204,294 peptides, and 224,453 sequences.

2.2. Database search

Proteome Discoverer software platform (version 1.3, Thermo Fisher Scientific) was used to generate the peak lists from human OL mass spectrometry raw data (msf) downloaded from ProteomeXchange, dataset PXD000263. A workflow feature in Proteome Discoverer software using SEQUEST search engine with percolator searched the data against the customized database with 13,825,667

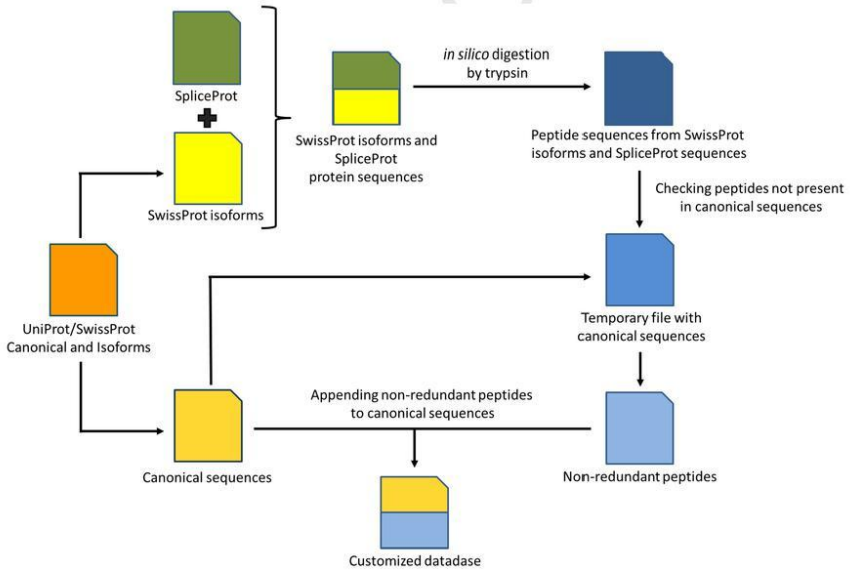

Fig. 1. Protein identification workflow using Proteome Discoverer. The customized database (light orange and light blue) was constructed from the association of SwissProt/Uniprot canonical sequences (light orange) and *in silico* digestion of peptides (dark, intermediate, and light blue) from SwissProt/Uniprot alternative isoforms (yellow) and SpliceProt sequences (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

residues and carbamidomethylation (+ 57.021 Da) as a fixed modification and oxidation of methionine (+ 15.995 Da) as a variable modification. Additional adjustments included two missed trypsin cleavages and tolerances of 10 ppm for precursor and 1 Da for fragment ions. To control the confidence of peptide spectrum matches (PSMs), we set the criteria to either false discovery rate (FDR) lower than 1% based on the q-value or posterior error probability (PEP) lower than 1%, also referred to as local FDR, as previously described [17]. The posterior error probability threshold (PEP < 1%) was applied to search for alternatively spliced variants using the customized sequence database.

2.3. Systems biology in silico

For interpreting functional significance, gene names corresponding to the identified proteins were uploaded into the Ingenuity Pathways Knowledge Base (IPKB) (http://www.ingenuity.com). Core analyses were run to identify potential protein interactions and to determine the most significant biological processes associated with these proteins.

2.4. Gene set enrichment analysis

To detect KEGG (Kyoto Encyclopedia of Genes and Genomes) metabolic pathways [18] overrepresented by the corresponding proteins identified by mass spectrometry, a gene set enrichment analysis (GSEA) was performed using the Model-based Gene Set Analysis (MGSA) (version 1.13) [19] package in Bioconductor's Project (version 2.11) [20], which is based on R statistical software (version 3.1.1) [21]. MGSA calculates the posterior probability of a pathway being overrepresented by the corresponding proteins identified by mass spectrometry. Identified pathways with a probability above 80% were considered overrepresented.

2.5. Quantitative reverse-transcription PCR

Total RNA was isolated from MO3.13 cells using Trizol LS Reagent (Invitrogen, Carlsbad, CA, USA) and reverse transcription was performed using oligo(dT)-primers (Superscript™ II Reverse Transcriptase, Invitrogen). Synthesized cDNA was subjected to qRT-PCR for the detection of *EEF1D*, *KRAS*, *MFF*, *SDR39U1*, and *SUGT1* gene transcripts (splice variants) predicted by our approach. β-actin mRNA amplification was used as an internal control. Amplification was performed with SYBR Green Master Mix (Applied Biosciences) using a 7500 Real-Time PCR System (Applied Biosciences, Foster City, CA, USA). All procedures were performed according to the manufacturer's instructions. Primers for splice variant detection were designed not to align to canonical transcripts (Supporting information Table S1).

3. Results and discussion

The human OL proteome was described by analyzing canonical proteins using standard search algorithms and reference sequence databases [14]. We used the SpliceProt repository [13] to analyze the OL proteome, with at least two unique peptides required for validation of canonical protein sequences. This resulted in 12,990 non-redundant peptides corresponding to 2081 proteins identified at PEP < 1%, supporting the findings of Iwata and colleagues [14] (the output file from Proteome Discoverer is available upon request).

To further explore this OL proteome and dissect the current biological knowledge of all splice variants detected in this dataset, we explored our customized sequence database based on PEP < 1%. We identified 75 peptides, corresponding to 39 isoforms (Table 1) and 38 genes (Supporting information Table S2). The MS/MS spectra of these splice variants is available as Supporting information. Of note, we also detected most of these splice variants using other MS search algorithms (Supporting Information Tables S3 and S4). The cases where the same peptide could be shared by more than one isoform from the same gene, the identification of a specific splice variant was not possible. However, we were able to detect the expression of a peptide that matched protein sequences of a set of splice variants from the same gene but not the canonical protein (Supporting information Fig. S1).

According to IPKB, these 38 genes are significantly involved in biochemical pathways such as glutamine degradation, acute phase response, and calcium signaling. Fig. 2 illustrates how our approach can recognize potential target molecules, which are not identified by traditional MS analysis. Proteins identified in our study are shown in gray, whereas protein hubs that might serve as targets for molecular interventions are shown in white.

Assigning true positives is one of the main challenges in the identification of splice variants in MS data. In addition to using SpliceProt, the alternative approach we propose assigns splice variants with higher confidence based on the ratio between peptide spectrum

Table 1
List of the 39 splice variants identified, their corresponding gene symbols, and the ratio of peptide spectrum matches per unique peptide.

Gene symbol	Uniprot/Refseq ID	≥ Cut off[a]
DBNL	Q9UJU6-2	Yes
DNM2	P50570-3	Yes
EEF1D	P29692-3	Yes
GLS	O94925-3	Yes
HNRNPAB	Q99729-3	Yes
IKBIP	Q70UQ0-4	Yes
KRAS	P01116-2	Yes
PDLIM3	Q53GG5-2	Yes
RAVER1	Q8IY67-2	Yes
SDR39U1	Q9NRG7-2	Yes
SPTAN1	Q13813-3	Yes
SREK1	Q8WXA9-2	Yes
SUGT1	Q9Y2Z0-2	Yes
TPD52L2	O43399-2	Yes
TPM1	P09493-5	Yes
TPM3	P06753-2	Yes
API5	Q9BZZ5-2	No
CAPZB	P47756-2	No
GANAB	Q14697-2	No
GLOD4	Q9HC38-2	No
HNRNPC	P07910-2	No
HNRNPK	P61978-3	No
HNRNPM	XP_005272536.1	No
LOXL3	P58215-3	No
MAP4	P27816-5	No
MFF	Q9GZY8-2	No
MRPL43	Q8N983-4	No
NDUFV3	P56181-2	No
NOL3	O60936-2	No
NOLC1	Q14978-3	No
NUDT4	Q9NZJ9-2	No
PKM	P14618-2	No
PNKD	Q8N490-2	No
RAVER1	NP_597709.2	No
SET	Q01105-2	No
SF1	Q15637-6	No
TOR1AIP1	Q5JTV8-3	No
TPM2	P07951-2	No
ZNF414	Q96IQ9-2	No

[a] ratio of peptide spectrum matches per unique peptides ≥ 1.4. See Methods.

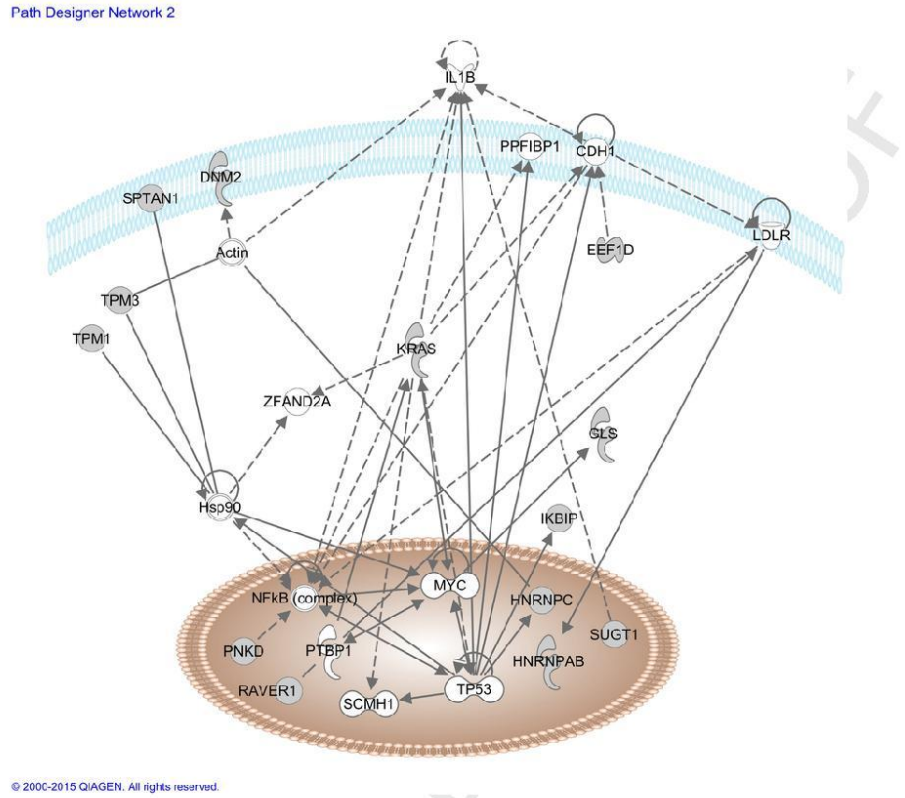

Fig. 2. Association between the 39 alternative isoforms based on known experimental protein-protein interactions. Proteins identified in the study (gray) and protein hubs (white) that might serve as targets for molecular interventions.

matches (PSMs) and unique peptides (UP) from canonical protein sequences. For the OL dataset, we obtained a median of 1.4 PSM/UP for the canonical proteins identified by at least two peptides (Supporting information Fig. S2). Using this criterion, we identified 16 splice forms with higher confidence (Table 1). Because our main objective was to explore alternative splicing diversity in human oligodendrocyte proteome, we present a literature review of all alternative isoforms identified by our approach and focused on brain, CNS, and associated diseases (Table 2). Below, we present an analysis of the splice variants according to the current literature. As we discuss in the next paragraphs, we detected splice variants in genes whose proteins are related to cellular processes associated with an oligodendrocyte cell line such as transcription and translation control, cell proliferation, and neurotransmitter regulation.

KRAS is a key regulator of cellular proliferation. The C-terminal hypervariable domain of the splice variant KRAS4B targets the protein to the plasma membrane, which is necessary for its activity [22]. As shown in Fig. 3A, KRAS4B (Uniprot ID P01116-2) differs from the KRAS4A canonical isoform (Uniprot ID P01116-1) in the C-termini domain by the presence of the coding exon 4 A [23]. The peptide we found matches exactly the splice junction of exon 3 and exon 4B, and is unique to KRAS4B. In addition, two peptides detected by our approach were attributed to the *HRAS* gene, a paralog of *KRAS* with > 80% sequence identity. It should be noted that because these two peptides have identical *N*-terminal sequences they also map to exons 1 and 2 of *KRAS*.

Glutaminase (GLS) catalyzes the hydrolytic deamination of glutamine to glutamate and ammonia and plays a role in the regulation of the excitatory neurotransmitter glutamate in the brain [24]. To date, three splice variants of the K-type mitochondrial glutaminase have been characterized [25]. Here, we identified two of those three splice variants. The *GLS* gene has 19 coding exons and three known splice patterns: the canonical protein KGA (Uniprot ID O94925-1), the GAM splice variant (Uniprot ID O94925-2), and the GAC splice variant (Uniprot ID O94925-3). Fig. 3B shows the distribution of all identified peptides assigned to the respective proteins. Splice variants KGA and GAC have been previously detected in brain and other tissues, whereas high GAC expression has been observed in brain tumors, including oligodendroglioma, astrocytoma, ganglioglioma, and ependymoma [26].

We also identified splice variants of the mitochondrial fission factor (*MFF*) gene. Mitochondrial fission events are regulated by several elements and mechanisms [27,28] that culminate with fragmentation of the organelle. The *MFF* gene is a component of a conserved membrane fission pathway used for constitutive and induced fission of mitochondria and peroxisomes. In addition, this gene codes for nine

Table 2
Splice variants identified with literature support.

Uniprot ID	Gene symbol	Description	≥ Cut off*	Identified in brain tissue	Reference
P29692-3	*EEF1D*	Translation elongation factor	Yes	Yes	[39,40]
O94925-3	*GLS*	Regulation of the neurotransmitter glutamate	Yes	Yes	[24,26]
P07910-2	*HNRNPC*	Binds with HNRNP C2 isoform in response to DNA damage stress	Yes	Yes	[32,34]
P01116-2	*KRAS*	Switch for cellular growth	Yes	No	[22]
Q9GZY8-2	*MFF*	Only exonic description of isoforms	No	No	[29]
Q53GG5-2	*PDLIM3*	Associated with development of heart muscle cells	Yes	Yes	[44]
Q53GG5-3	*PDLIM3*	Associated with development of heart muscle cells	No	No	[44]
P14618-2	*PKM*	Expression influenced by heterogeneous ribonucleoproteins	No	Yes, brain tumor only	[38]
Q9Y2Z0-2	*SUGT1*	Expression described in some tissues such as brain	Yes	Yes	[30]
Q5JTV8-3	*TOR1AIP1*	Associated with the inner nuclear membrane	No	Yes	[43]
P09493-5	*TPM1*	Composition of actin fibers	No	Yes	[46]
P09493-2	*TPM1*	Composition of actin fibers	Yes	No	[46]
P07951-2	*TPM2*	Composition of actin fibers	No	No	[46]
P06753-2	*TPM3*	Composition of actin fibers	Yes	No	[46]
P06753-3	*TPM3*	Composition of actin fibers	Yes	No	[46]
P06753-4	*TPM3*	Composition of actin fibers	Yes	Yes	[46]
P06753-5	*TPM3*	Composition of actin fibers	Yes	No	[46]
P06753-6	*TPM3*	Composition of actin fibers	Yes	No	[46]

* ratio of peptide spectrum matches per unique peptides ≥ 1.4. See Methods.

splice variants that differ by the presence or absence of exon 1 and combinations of exons 5, 6, and 7, which encode the central region of the protein [29]. In our analysis, we detected the isoform Q9GZY8-2, which lacks exon 7, and the canonical protein (Q9GZY8-1) (Fig. 3C).

The human gene *SUGT1* is associated with the kinetochore protein complex and involved in centrosome attachment during the cell cycle. This gene encodes a canonical protein (SUGT1A; Uniprot ID Q9Y2Z0) and an alternative isoform (SUGT1B; Uniprot ID Q9Y2Z0-2) identified by Niikura and Kitagawa (2003) [30]. The two proteins are 91% identical at the protein sequence level, but SGT1B contains an additional 33 amino acids in a region between exons 5 and 6 of *SGT1A* and lacks Ser^{110} of SGT1A. This insertion occurs in a tetratricopeptide repeat (TPR) motif, which is a module that promotes specific protein-protein interactions. Our analysis indicates the presence of both the canonical protein and splice variant in the OL proteome (Fig. 3D).

The heterogeneous nuclear ribonucleoprotein (hnRNP) gene family plays a key role in mRNA metabolism and was previously detected in oligodendrocytes [14]. High levels of the family members HRNPAB and HNRNPC may lead to overexpression of the MYC gene, which is a key transcription factor in cancer and a molecular marker for some medulloblastoma tumors [31]. The *HNRNPC* gene is associated with pre-mRNA processing and transporting [32]. *HNRPNC* encodes two distinct isoforms: *C*1 and C2, with 293 and 306 amino acids in length, respectively [33]. HNRNPC1 was described to bind to p53 mRNA during apoptosis [34]. Here, we identified the expression of both splice variants: canonical (HNRNPC1) and HNRNPC2. The physical interaction of these two protein isoforms seems to control the aggressiveness of glioblastoma cells by downregulating the PDCD4 gene product [32]. hnRNP proteins were also found associated with schizophrenia [35] and may play a pivotal role in the oligodendrocyte dysfunction observed in this disorder [36].

Another canonical protein identified by our approach is pyruvate kinase (PKM). *PKM* encodes a constitutively expressed protein in healthy cells and in medulloblastoma [31]. *PKM* has two mutually exclusive exons (9 and 10), producing two splice variants, M1 (alternative) and M2 (canonical), respectively. We identified expression of splice forms M1 and M2 in the OL proteome. The M2 protein is overexpressed in proliferative and cancer cells, if compared to M1 [37]. Its expression is regulated by hnRNP *A*1/A2 and the polypyrimidine-tract-binding protein (PTB) [38], which were also detected in our study.

We also detected the canonical protein and one known splice variant (Uniprot ID P29692-3) of eukaryotic translation elongation factor 1 delta (EEF1D). This gene is located in the human chromosome 8 and encodes four splice variants expressed in the human brain [39]. All splice variants can be clustered into two groups based in peptide length: the longest isoform 2, with 647 amino acids (Uniprot ID P29692-2) and named eEF1BδL, and three shorter isoforms with similar length. The group of short isoforms includes the canonical protein (eEF1Bδ) (281 amino acids; Uniprot ID P29692-1), isoform number 3 (257 amino acids; Uniprot ID P29692-3), and isoform number 4 (262 amino acids; Uniprot ID P29692-4). The eEF1Bδ and eEF1BδL proteins have distinct cellular localization and biological functions. The former is usually located in the cytoplasm and plays an important role as a translation elongation factor when binding to the other eEF1 complex members eEF1Bα and eEF1Bγ. The latter is located in the nucleus and induces the transcription of *HSPA6*, *CRYAB*, DNAJB1, and *HMOX1*[40]. Using array comparative genomic hybridization (aCGH), de Bartoli et al. (2006) [41] showed that the medulloblastoma outcome is adversely associated with overexpression of *EEF1D* and thus a potential disease biomarker [41].

We also identified the *TOR1AIP1* gene, which encodes the lamina-associated polypeptide 1 (LAP1), an internal protein of the inner nuclear membrane [42]. There are three protein products of *LAP1*: the canonical isoform (Q5JTV8-1), the LAP1B isoform (Q5JTV8-3), and the recently identified LAP1C. The *LAP1* and *LAP1B* isoforms differ by the insertion of a CAG in the 3′ splice site of intron 2 in *LAP1B*, resulting in an additional alanine in the coding sequence when compared to the canonical sequence. Using bioinformatics, mass spectrometry, and molecular biology techniques, Santos et al. (2014) [43] identified LAP1C in a cell line derived from neuroblastoma. In our study, we identified the expression of the canonical LAP1 and the splice variant LAP1B.

The *PDLIM3* gene is associated with myotonic dystrophy (MD) and codes for three splice variants varying in the use of exons 4, 5, and 6 (Uniprot ID Q53GG5-1, Q53GG5-2, and Q53GG5-3). The association of this gene with MD was confirmed by Ohsawa et al. (2011) [44] who demonstrated predominant expression of isoform Q53GG5-2 in patients with the disease. This isoform has an exon 4 insertion and lacks exons 5 and 6, which may influence the binding

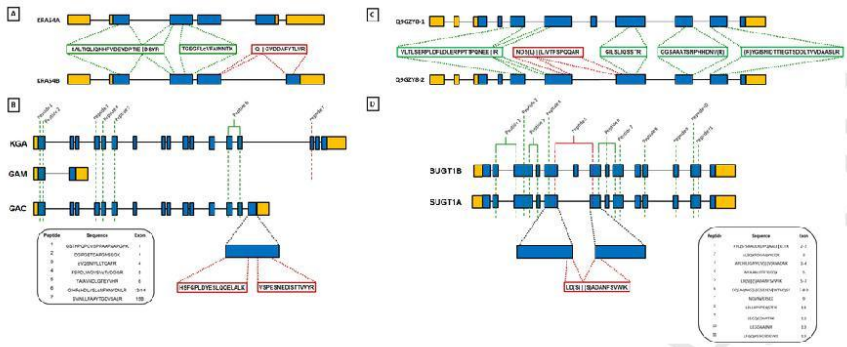

Fig. 3. Schematic representation of splice variants and identified peptides from: A) *KRAS* transcripts, B) *GLS* transcripts, C) *MFF* transcripts, and D) *SUGT1* transcripts. Exons are represented by yellow (UTRs) and blue (coding sequences) boxes and introns are represented by pipes ("|"). Distinct peptides are represented by dashed and dotted red lines and red boxes. Amino acids between parentheses indicate intron phase.

of PDLIM3 and α-actinin 2. The physical interaction between these proteins occurs *via* the PDZ domain located in exon 6 of the *PDLIM3* gene, which is absent in the splice variant Q53GG5-2. All *PDLIM3* splice variants (Uniprot ID Q53GG5-3 and Q53GG5-2) were identified in the OL proteome.

We also developed an approach to determine splice variant expression based on peptide counts mapped onto the sequence of canonical and splice variants. For that, we calculated the difference between PSMs shared by and exclusive to canonical and alternative isoform peptides (Supporting information Table S5). This analysis showed a higher frequency of PKM and HNRNPK canonical proteins compared to their splice variants. Using this approach, we found that Q9NRG7-2 was the only splice variant of the SDR39U1 gene expressed in this dataset, whereas the canonical protein (Q9NRG7) was not expressed. In this case, Q9NRG7 and Q9NRG7-2 share one peptide with three PSMs and the Q9NRG7-2 splice variant has one exclusive peptide with four PSMs (Fig. 4 and Supporting information Table S5). Thus, the three PSMs associated with the peptide shared between the canonical protein and splice variant may reflect only the expression of the splice variant.

To confirm the mRNA expression of some of the splice variants detected at the proteome level in this study, we selected five (*EEF1D*, *KRAS*, *MFF*, *SDR39U1*, and *SUGT1*) for experimental validation (Table 1). All five splice variants were detected at the transcriptome level in the MO3.13 cell line (Supporting Information Fig. S3).

To determine the effect of splice variants on OL, we performed a gene set enrichment analysis (GSEA) from the list of 38 identified genes. This analysis revealed a muscle thin filament tropomyosin (GO: 0005862) with an 80% posterior probability of being overrepresented. Three tropomyosin genes were represented in our 38-gene list: *TPM1*, *TPM2*, and *TPM3*. Tropomyosins are components of actin filaments that are found mainly in muscle cells and in dendrites from neuronal cells [45]. Human tropomyosins are encoded by four genes, known as *TPM1* (chromosome 15), *TPM2* (chromosome 9), *TPM3* (chromosome 1), and *TPM4* (chromosome 19), corresponding to the α, β, γ, and δ mammal genes, respectively [46]. Expression of TmBr2 and TmBr3 rat isoforms (corresponding to human TPM1 isoform 5) was found in astrocytes and oligodendrocytes [47]. Tm1 (corresponding to human *TPM2*) was also identified in rat astrocytes in cell culture [47]. Another study with rat neurons identified the expression of Tm5 isoforms (corresponding to human *TPM3*) in cerebellum tissue culture. In mice, Tm1 (*TPM1*), Tm2 (*TPM2*), and Tm3 (*TPM3*) expression occurs only in a neuroblast-rich zone and expression of the three genes decreases when cells become mature neurons [48]. There are 10 different isoforms of the human *TPM1* gene, three different isoforms of *TPM2*, and seven isoforms of the *TPM3* gene. In human brain, only a few isoforms have been identified to date: isoforms 3 and 5 from *TPM1*, isoform 1 from *TPM2*, and isoform 4 from *TPM3*[49]. Our study identified isoforms 2 and 5 from the *TPM1* gene, isoforms 2 and 3 from the *TPM2* gene, and isoforms 2, 3, 4, 5, and 6 from the *TPM3* gene. Even though isoforms from the *TPM* gene family were detected in our approach, it should be noted that MO3.13 is a hybrid cell line of rhabdomyosarcoma (skeletal muscle tumor) and human adult oligodendrocytes. Thus, we believe that identification of these isoforms may be affected by the biological background of this cell line.

We also investigated known experimental protein-protein interactions from the BioGRID database [50] in the set of 39 splice variants using geneMANIA [51]. Campos-Sandoval et al. (2015) [52] focused on glutamine function and its splice variants, because each isoform regulates *NFKB1*, *NFKB2*, *MYC*, and *TP53* expression in brain tumors: while the p53 protein encoded by the *TP53* gene is responsible for maintaining the genomic stability of oligodendrocyte precursor cells (OPC) [53], MYC and NFκB control another group of genes such as those associated with myelination that also play important roles during differentiation [54]. Fig. 2 shows a similar association, even though MO3.13 is not a malignant cell.

To investigate interaction networks between specific gene isoforms, we used the Human Isoform-level Functional Relationship Network tool (Hisonet) [55]. We looked for the network of eight isoforms from Table 1, but only four were found: isoform 2 of PDLIM3 (P29692-3), isoform 2 of KRAS (P01116-2), isoform 2 of SUGT1 (Q9Y2Z0-2), and isoform 1 of EFF1D (P29692-3). Isoform 1 of the *EEF1D* gene (Fig. 5A) also interacts with isoforms 5 and 7 of the same gene, whereas isoform 2 of *PDLIM3* also interacts with other isoforms from other *PDLIM* family members such as isoform 1 of *PDLIM4*, isoform 4 of *PDLIM5*, and isoforms 1 and 2 of *PDLIM7* (Fig. 5B). As shown in Fig. 5, both KRAS (Fig. 5C and D) and SUGT1 (Fig. 5E and F) splice variants have distinct partners, which may influence the known gene regulatory network based on expressed proteins. These results show the high level of complexity in selecting only few isoforms and the continuous need for studies focusing on splice variants.

Taken together, our findings highlight the importance of exploring full proteome datasets to the complete set of proteins in a healthy tissue or disease state using a proteogenomic approach. We expect our

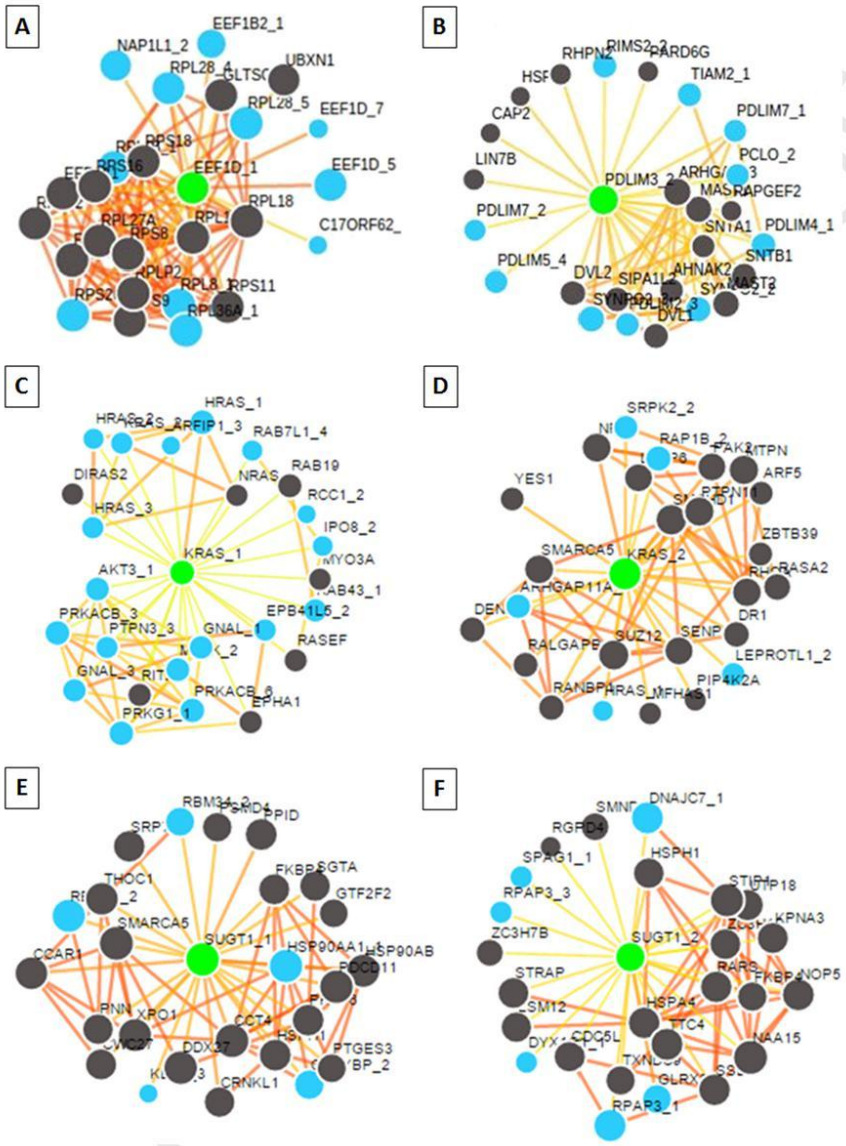

Fig. 5. Splice variant protein-protein network interaction. A) EFF1D (P29692-3), B) PDLIM3 (P29692-3), C) isoform 1 of KRAS (P01116-1), D) isoform 2 of KRAS (P01116-2), E) isoform 1 of SUGT1 (Q9Y2Z0-2), and F) isoform 2 of SUGT1 (Q9Y2Z0-2).

efforts to drive other research groups to investigate the biological significance of the splice variants identified in this study in CNS diseases.

4. Conclusions

While Iwata and colleagues [14] aimed to characterize the human OL proteome, we complemented their findings by focusing on the

identification of splice variants. Additionally, we provide a tool that can be used for any future proteomic investigation. Our customized database composed of unique non-redundant peptides and full canonical protein sequences detected 39 splice variant peptides. We assessed the mRNA expression of five selected splice variants; all isoforms were detected at the transcriptome level. Thus, we provide a potential tool for the identification of splice variants in the MS proteome data based on the identification of prototypic peptides [56]. We expect our results to encourage other research groups to investigate the biological function of these splice variants in oligodendrocyte regulation and development and their potential role in clinical applications using targeted proteomics technologies such as Selected Reaction Monitoring (SRM) and Multiple Reaction Monitoring (MRM).

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2016.05.023.

Conflict of interest statement

FP and CGF are authors or inventors of the patent number 699,132 granted in Switzerland regarding the ternary matrix methodology used in the development of this work. However, the results achieved in this work using the patent's methodology have merely scientific interest.

Transparency document

The Transparency document associated with this article can be found, in the online version.

Acknowledgements

FP, RT and GW are supported by *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) and *Vice-Presidência de Ensino, Informação e Comunicação/Pró-Reitoria* – IOC/FIOCRUZ. FP and PSAS acknowledge the support of *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro* (FAPERJ). FP acknowledges the support of *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq). FP, CGF and the Bioinformatics Unit acknowledge the support of the *Fundação do Câncer*, Ministry of Health of Brazil and CAPES. DMS and JSC are supported by Sao Paulo Research Foundation (FAPESP) grants 13/08711-3 and 14/14881-1. AFPL is supported by FAPESP (grants 2009/54067-3 and 2010/19278-0). All authors acknowledge Veronica Aran for comments on KRAS splice variants and Patricia A. Possik for critical reading. Authors thank Prof. Sabine Bahn (University of Cambrigde) to allow access to Ingenuity Pathways Knowledge Base (IPKB).

References

[1] Y. Perez-Riverol, E. Alpi, R. Wang, H. Hermjakob, J.A. Vizcaíno, Making proteomics data accessible and reusable: current state of proteomics databases and repositories, Proteomics 15 (2015) 930–949.
[2] J.A. Vizcaíno, R.G. Côte, A. Csordas, J.A. Dianes, A. Fabregat, J.M. Foster, et al., The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013, Nucleic Acids Res. 41 (Database issue) (2013) D1063–D1069.
[3] UniProt Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43 (Database issue) (2015) D204–D212.
[4] K.D. Pruitt, G.R. Brown, S.M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, et al., RefSeq: an update on mammalian reference sequences, Nucleic Acids Res. 42 (Database issue) (2014) D756–D763.
[5] T.N. Villavicencio-Diaz, A. Rodriguez-Ulloa, O. Guirola-Cruz, Y. Perez-Riverol, Bioinformatics tools for the functional interpretation of quantitative proteomics results, Curr. Top. Med. Chem. 14 (2014) 435–449.
[6] D.L. Black, Protein diversity from alternative splicing: achallenge for bioinformatics and post-genome biology, Cell 103 (2000) 367–370.
[7] O. Kelemen, P. Convertini, Z. Zhang, Y. Wen, M. Shen, M. Falaleeva, et al., Function of alternative splicing, Gene 514 (2013) 1–30.
[8] Y. Lee, Y. Lee, B. Kim, Y. Shin, S. Nam, P. Kim, et al., ECgene: an alternative splicing database update, Nucleic Acids Res. 35 (2007) D99–103.
[9] R. Menon, Q. Zhang, Y. Zhang, D. Fermin, N. Bardeesy, R.A. DePinho, et al., Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer, Cancer Res. 69 (2009) 300–309.
[10] G.S. Omenn, A.K. Yocum, R. Menon, Alternative splice variants, a new class of protein cancer biomarker candidates: findings in pancreatic cancer and breast cancer with systems biology implications, Dis. Markers 28 (2010) 241–251.
[11] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Anal. Chem. 74 (2002) 5383–5392.
[12] R. Tavares, N.M. Scherer, C.G. Ferreira, F.F. Costa, F. Passetti, Splice variants in the proteome: a promising and challenging field to targeted drug discovery, Drug Discov. Today 20 (2015) 353–360.
[13] R. Tavares, N.M. Scherer, B.A. Pauletti, E. Araújo, E.L. Folador, G. Espindola, et al., SpliceProt: a protein sequence repository of predicted human splice variants, Proteomics 14 (2014) 181–185.
[14] K. Iwata, C.C. Café-Mendes, A. Schmitt, J. Steiner, T. Manabe, H. Matsuzaki, et al., The human oligodendrocyte proteome, Proteomics 13 (2013) 3548–3553.
[15] A. Nonneman, W. Robberecht, L. Van Den Bosch, The role of oligodendroglial dysfunction in amyotrophic lateral sclerosis, Neurodegener Dis Manag. 4 (2014) 223–239.
[16] A.I. Nesvizhskii, Proteogenomics: concepts, applications and computational strategies, Nat. Methods 11 (2014) 1114–1125.
[17] G.M. Sheynkman, M.R. Shortreed, B.L. Frey, L.M. Smith, Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq, Mol. Cell. Proteomics 12 (2013) 2341–2353.
[18] M. Kanehisa, S. Goto, Y. Sato, et al., Data, information, knowledge and principle: back to metabolism in KEGG, Nucleic Acids Res. 42 (2014) D199–D205.
[19] S. Bauer, J. Gagneur, P.N. Robinson, GOing Bayesian: model-based gene set analysis of genome-scale data, Nucleic Acids Res. 38 (2010) 3523–3532.
[20] M. Reimers, V.J. Carey, Bioconductor: an open source framework for bioinformatics and computational biology, Methods Enzymol. 411 (2006) 119–134.
[21] R Core Team, R: A Language and Environment for Statistical Computing, 2014.
[22] A. Welman, M.M. Burger, J. Hagmann, Structure and function of the C-terminal hypervariable region of K-Ras4B in plasma membrane targetting and transformation, Oncogene 19 (2000) 4582–4591.
[23] J.O. McGrath, D.J. Capon, D.H. Smith, E.Y. Chen, P.H. Seeburg, D.V. Goeddel, et al., Structure and organization of the human Ki-ras proto-oncogene and a related processed pseudogene, Nature 304 (1983) 501–506.
[24] J. Márquez, A.R. de la Oliva, J.M. Matés, J.A. Segura, F.J. Alonso, Glutaminase: a multifaceted protein not only involved in generating glutamate, Neurochem. Int. 48 (2006) 465–471.
[25] K.M. Elgadi, R.A. Meguid, M. Qian, W.W. Souba, S.F. Abcouwer, Cloning and analysis of unique human glutaminase isoforms generated by tissue-specific alternative splicing cloning and analysis of unique human glutaminase isoforms generated by tissue-specific alternative splicing, Physiol. Genomics 1 (2014) 51–62.
[26] M. Szeliga, E. Matyja, M. Obara, W. Grajkowska, T. Czernicki, J. Albrecht, Relative expression of mRNAS coding for glutaminase isoforms in CNS tissues and CNS tumors, Neurochem. Res. 33 (2008) 808–813.
[27] K. Okamoto, J.M. Shaw, Mitochondrial morphology and dynamics in yeast and multicellular eukaryotes, Annu. Rev. Genet. 39 (2005) 503–536.
[28] D.C. Chan, Mitochondrial fusion and fission in mammals, Annu. Rev. Cell Dev. Biol. 22 (2006) 79–99.
[29] S. Gandre-Babbe, A.M. van der Bliek, The novel tail-anchored membrane protein Mff controls mitochondrial and peroxisomal fission in mammalian cells, Mol. Biol. Cell 19 (2008) 2402–2412.
[30] Y. Niikura, K. Kitagawa, Identification of a novel splice variant: human SGT1B (SUGT1B), DNA Seq. 14 (2003) 436–441.
[31] J.A. Staal, L.S. Lau, H. Zhang, W.J. Ingram, A.R. Hallahan, P.A. Northcott, et al., Proteomic profiling of high risk medulloblastoma reveals functional biology, Oncotarget 6 (2015) 14584–14595.
[32] Y.M. Park, S.J. Hwang, K. Masuda, K.M. Choi, M.R. Jeong, D.H. Nam, et al., Heterogeneous nuclear ribonucleoprotein C1/C2 controls the metastatic potential of glioblastoma by regulating PDCD4, Mol. Cell. Biol. 32 (2012) 4237–4244.
[33] C.G. Burd, M.S. Swanson, M. Görlach, G. Dreyfuss, Primary structures of the heterogeneous nuclear ribonucleoprotein A2, B1, and C2 proteins: a diversity of RNA binding proteins is generated by small peptide inserts, Proc. Natl. Acad. Sci. U. S. A. 86 (1989) 9788–9792.
[34] K.J. Christian, M.A. Lang, F. Raffalli-Mathieu, Interaction of heterogeneous nuclear ribonucleoprotein C1/C2 with a novel cis-regulatory element within p53 mRNA as a response to cytostatic, Drug Treatment. Molecular Pharmacol. 73 (2008) 1558–1567.

[35] D. Martins-de-Souza, W.F. Gattaz, A. Schmitt, J.C. Novello, S. Marangoni, C.W. Turck, et al., Proteome analysis of schizophrenia patients Wernicke's area reveals an energy metabolism dysregulation, BMC Psychiatry. 30 (2009) 9–17.
[36] K. Iwata, H. Matsuzaki, T. Manabe, N. Mori, Altering the expression balance of hnRNP C1 and C2 changes the expression of myelination-related genes, Psychiatry Res. 190 (2011) 364–366.
[37] H.R. Christofk, M.G. Vander Heiden, M.H. Harris, A. Ramanathan, R.E. Gerszten, R. Wei, et al., The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth, Nature 452 (2008) 230–233.
[38] C.V. Clower, D. Chatterjee, Z. Wang, L.C. Cantley, M.G. Vander Heiden, A.R. Krainer, The alternative splicing repressors hnRNP A1/A2 and PTB influence pyruvate kinase isoform expression and cell metabolism, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 1894–1899.
[39] Y. Cao, M. Portela, J. Janikiewicz, J. Doig, C.M. Abbott, Characterisation of translation elongation factor eEF1B subunit expression in mammalian cells and tissues and Co-localisation with eEF1A2, PLoS One 9 (2014) 1–18.
[40] T. Kaitsuka, M. Matsushita, Regulation of translation factor EEF1D gene function by alternative splicing, Int. J. Mol. Sci. 16 (2015) 3970–3979.
[41] M. de Bartoli, R.C. Castellino, X.Y. Lu, J. Deyo, L.M. Sturla, A.M. Adesina, et al., Medulloblastoma outcome is adversely associated with overexpression of EEF1D, RPL30, and RPS20 on the long arm of chromosome 8, BMC Cancer 6 (2006) 1–13.
[42] A. Senior, L. Gerace, Integral membrane proteins specific to the inner nuclear membrane and associated with the nuclear lamina, J. Cell Biol. 107 (1988) 2029–2036.
[43] M. Santos, S.C. Domingues, P. Costa, T. Muller, S. Galozzi, K. Marcus, et al., Identification of a novel human LAP1 isoform that is regulated by protein phosphorylation, Plos Ones 9 (2014) 1–32.
[44] N. Ohsawa, M. Koebis, S. Suo, I. Nishino, S. Ishirua, Alternative splicing of PDLIM3/ALP, for α-actinin-associated LIM protein 3, is aberrant in persons with myotonic dystrophy, Biochem BiophysRes Commun. 409 (2011) 64–69.
[45] D.A. Fletcher, R.D. Mullins, Cell mechanics and the cytoskeleton, Nature 463 (2010) 485–492.
[46] J.J.C. Lin, R.D. Esppinga, K.S. Warre, K.R. McCrae, Human tropomyosin isoforms in the regulation of cytoskeleton functions, Adv. Exp. Med. Biol. 644 (2008) 201–222.
[47] L. Had, C. Faivre-Sarrailh, C. Legrand, A. Rabié, The expression of tropomyosin genes in pure cultures of rat neurons, astrocytes and oligodendrocytes is highly cell-type specific and strongly regulated during development, Brain Res. Mol. Brain Res. 18 (1993) 77–86.
[48] J.A. Hughes, C.M. Cooke-Yarborough, N.C. Chadwick, G. Schevzov, S.M. Arbuckle, P. Gunning, et al., High-molecular-weight tropomyosins localize to the contractile rings of dividing CNS cells but are absent from malignant pediatric and adult CNS tumors, Glia 42 (2003) 25–35.
[49] J.J.C. Lin, R.D. Esppinga, K.S. Warre, K.R. McCrae, Human tropomyosin isoforms in the regulation of cytoskeleton functions, Adv. Exp. Med. Biol. 644 (2008) 201–222.
[50] B.J. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, M. Livstone, et al., The BioGRID interaction database : 2008 update, Nucleic Acids Res. 36 (2008) 637–640.
[51] D. Warde-Farley, S.L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, et al., The GeneMANIA prediction server : biological network integration for gene prioritization and predicting gene function, Nucleic Acids Res. 38 (2010) 214–220.
[52] J.A. Campos-Sandoval, M. Martín-Rufián, C. Cardona, C. Lobo, A. Peñalver, J. Márquez, Glutaminases in brain: multiple isoforms for many purposes, Neurochem. Int. 15 (2015) 0186–0197.
[53] Y.M. Tokumoto, D.G. Tang, M.C. Raff, Two molecularly distinct intracellular pathways to oligodendrocyte differentiation: role of a p53 family protein, EMBO J. 20 (2001) 5261–5268.
[54] T. Bank, M. Prinz, NF-κB signaling regulates myelination in the CNS, Front. Mol. Neurosci. 47 (2014) 1–6.
[55] H.D. Li, R. Menon, B. Govindarajoo, B. Panwar, Y. Zhang, G.S. Omenn, et al., Functional networks of highest-connected splice isoforms: from the chromosome 17 human proteome project, J. Proteome Res. 14 (9) (2015) 3484.
[56] A.I. Nesvizhskii, R. Aebersold, Interpretation of shotgun proteomic data: the protein inference problem, Mol. Cell. Proteomics 4 (2005) 1419–1440.