Explorando a Complexidade do Transcriptoma Humano

José Eduardo Kroll

Tese apresentada
ao
Programa Interunidades em Bioinformática
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Área: Bioinformática

Orientador: Prof. Dr. Sandro José de Souza

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, Dezembro de 2013

Explorando a Complexidade do Transcriptoma Humano

Esta versão da tese contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 12/12/2013. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Sandro José de Souza (orientador) - ICe-UFRN
- Prof^{a}. Dr^{a}. Ariane Machado Lima - EACH-USP
- Prof^{a}. Dr^{a}. Bettina Malnic - IQ-USP
- Prof. Dr. Georgio Joannis Pappas Júnior - UnB
- Prof^{a}. Dr^{a}. Helena Paula Brentani - FM-USP

Resumo

KROLL, J. E. **Explorando a Complexidade do Transcriptoma Humano**. 2013. 60 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2013.

O *splicing* alternativo é um processo no qual moléculas idênticas de pré-mRNA são processadas de diferentes formas. Ele é fundamental em organismos complexos, pois é responsável por criar uma ampla diversidade de proteínas a partir de um número relativamente pequeno de genes. Contudo, poucas proteínas advindas do *splicing* alternativo já foram identificadas, visto que a maioria dos espectros de espectrometria de massa em *tandem* (MS/MS) não encontra sequências correspondentes nos diversos bancos de dados de proteínas disponíveis. Entre diversos fatores, isso ocorre porque um número reduzido de eventos de *splicing* alternativo (ASEs) são conhecidos e devidamente estudados. Nesse trabalho, o espectro de eventos observáveis foi ampliado por meio da análise de eventos complexos de *splicing* alternativo (CASEs), que consideram múltiplos ASEs em um ou diferentes transcritos. Foi desenvolvido um novo método de análise utilizando expressões regulares (*regexes*) associada a uma sintaxe baseada em caracteres intuitivos. O método de análise e a sintaxe foram implementados em uma ferramenta *web* denominada de SPLOOCE (http://www.bioinformatics-brazil.org/splooce) que também apresenta ferramentas extras de análise. Adicionalmente, os subestimados eventos do tipo retenção de íntron (IR) foram explorados em busca de evidências funcionais por meio de análises de MS/MS. Como resultado, eventos bastante incomuns foram observados no proteoma humano, sugerindo que muito pouco ainda é conhecido sobre a complexidade transcriptômica e proteômica humana. Portanto, com base nesses dados, esse trabalho representa um grande avanço no estudo de fenômenos de *splicing* alternativo ainda pouco explorados.

Palavras-chave: *splicing* alternativo, SPLOOCE, transcriptoma, proteoma, retenção de íntron, espectrometria de massa.

Abstract

KROLL, J. E. **Exploring the Complexity of Human Transcriptome**. 2013. 60 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2013.

Alternative splicing is defined, basically, as a process in which identical pre-mRNA molecules are processed in different ways in terms of usage of exon/introns borders. It is a fundamental process in complex organisms, and is responsible for creating a large diversity of proteins from a relatively small number of genes. However, just a few proteins resulted from alternative splicing were already identified, since only a small part of *tandem* mass spectrometry (MS/MS) spetras match proteins in sequence databases. Among different factors, it occurs because a reduced number of alternative splicing events (ASEs) are known and properly studied. In this work, the landscape of observable events was amplified through the analysis of complex alternative splicing events (CASEs), which consider different ASEs within the same or different transcripts. A method of analysis was developed using regular expressions (*regexes*) associated with a syntax composed of intuitive characters. Those features were implemented in a web tool called SPLOOCE (http://www.bioinformatics-brazil.org/splooce) that also has extra analysis tools. Furthermore, the understudied events known as intron retention (IR) were explored using MS/MS analyses as a strategy to identify functional roles. As result, very uncommon events were observed in human proteome, suggesting that little is currently known about the complexity of the human proteome and transcriptome. Based on those data, it can be concluded that this work represents a significant advance in the study of uncommon and understudied alternative splicing events.

Keywords: Alternative splicing, SPLOOCE, transcriptome, proteome, intron retention, mass spectrometry.

Sumário

Lista de Abreviaturas

AS	*Alternative Splicing* (*Splicing* Alternativo)
ASE	*Alternative Splicing Event* (Evento de *Splicing* Alternativo)
CASE	*Complex Alternative Splicing Event* (Evento Complexo de *Splicing* Alternativo)
Regex	*Regular Expression* (Expressão Regular)
MS	*Mass Espectrometry* (Espectrometria de Massa)
MS/MS	*Tandem Mass Espectrometry* (Espectrometria de Massa em *Tandem*)
SPLOOCE	$Spl + {}_{G}OO_{gle} + ce$
IR	*Intron Retention* (Retenção de Íntron)
ES	*Exon Skipping* (Éxon *Skipping*)
DSS	*Dual-specific Splice Site* (Sítio de *Splice* Duplamente Específico)
A3'SS	*Alternative 3' Splice Site* (Sítio de *Splice* 3' Alternativo)
A5'SS	*Alternative 5' Splice Site* (Sítio de *Splice* 5' Alternativo)
RefSeq	*Reference Sequence* (Sequência de Referência)
EST	*Expressed Sequenced Tag*
NGS	*Next-generation of Sequencing* (Sequenciamento de Próxima Geração)
RNASeq	*RNA Sequencing* (Sequenciamento de RNA/Transcriptoma)
DNA	*Deoxyribonucleic Acid* (Ácido Desoxirribonucleico)
cDNA	*Complementary DNA* (DNA Complementar)
RNA	*Ribonucleic Acid* (Ácido Ribonucléico)
mRNA	*Messenger RNA* (RNA Mensageiro)
snRNA	*Small Nuclear RNA* (Pequeno RNA nuclear)
RNP	*Ribonucleoprotein* (Ribonucleoproteína)
hnRNP	*Heterogeneous Nuclear RNP* (RNP Heterogênea)
BS	*Branch Site* (Sítio Ramificado)
PPT	*Polypyrimidine Tract* (Trato de Polipirimidina)
PTB	*Polypyrimidine Tract-binding Protein* (Proteína que se liga à PPT)
SRE	*Splicing Regulatory Element* (Elemento regulatório de *Splicing*)
ESE	*Exonic Splicing Enhancer* (Estimulador Exônico de *Splicing*)
ISE	*Intronic Splicing Enhancer* (Estimulador Intrônico de *Splicing*)
ESS	*Exonic Splicing Silencer* (Silenciador Exônico de *Splicing*)
ISS	*Intronic Splicing Silencer* (Silenciador Intrônico de *Splicing*)
SR	*Serine/Arginine-Rich Protein* (Proteína Rica em Serina/Arginina)

SNP	*Single Nucleotide Polymorphism* (Polimorfirsmo de um Único Nucleotídeo)
FGF	*Fibroblast Growth Factor* (Fator de Crescimento de Fibroblastos)
FGFR	*Fibroblast Growth Factor Receptor* (Receptor de FGF)
pb	Par de Base
DB	*Database* (Banco de Dados)
CG	Citosina + Guanina
CDS	*Coding Sequence* (Sequência Codificadora)
UTR	*Untranslated Region* (Região Não Codificadora)
FxF	*Full-insert versus Full-insert Sequences*
FxA	*Full-insert versus All Sequences*
ORF	*Open Reading Frame*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Apresentação

Esse trabalho é um grande esforço no sentido de melhor entender os eventos de *splicing* alternativo. Aliás, o que é *splicing* alternativo e por que estudá-los? Para melhor entender esse fenômeno e sua importância no desenvolvimento de doenças como o câncer, uma revisão geral sobre o assunto é apresentada (capítulo 2).

Posteriormente, o SPLOOCE é descrito (capítulo 3). Ele é um banco de dados híbrido e inovador que tem com objetivo a análise de eventos complexos de *splicing* alternativo por meio do uso de uma sintaxe intuitiva. Embora atualmente exista pelo menos uma dezena de bancos de dados sobre *splicing* alternativo publicamente disponíveis, cada um deles possui diferentes características. Novas abordagens sobre esse fenômeno são benéficas e podem trazer avanços importantes para o melhor entendimento sobre a complexidade celular humana.

Em seguida, com base em análises feitas no ano de 2004 (Galante *et al.*, 2004), os eventos conhecidos como "retenção de íntron" foram reexplorados utilizando dados recentes e algumas tecnologias previamente implementadas no SPLOOCE (capítulo 4). Esses eventos são pouco frequentes e funcionalmente bastante subestimados, umaz vez que podem ter como causa o processamento imcompleto de sequências de pré-mRNA. Análises buscaram evidências funcionais por meio de dados transcriptômicos e proteômicos.

No último capítulo (5) uma conclusão global e sugestões futuras são apresentadas.

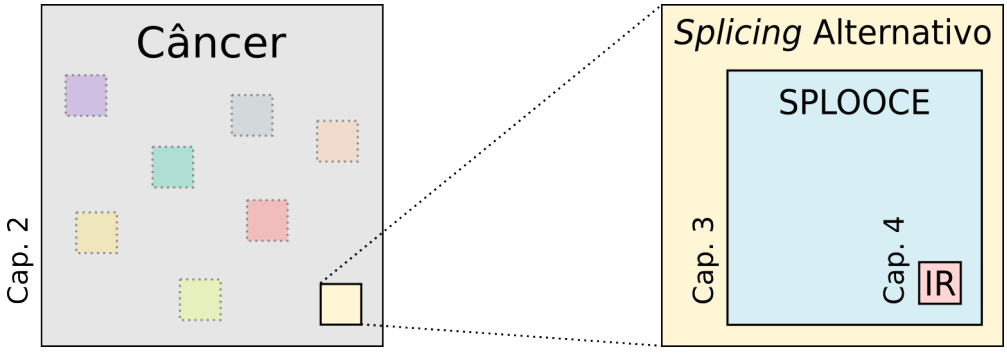

Figura 1.1: *Organização e localização contextual dos assuntos abordados nesse trabalho.*

Resumidamente, esse trabalho contém 2 projetos de pesquisa distintos (capítulos 3 e 4) e está organizado conforme apresentado na figura 1.1. Os capítulos possuem um formato independente e seguem a mesma estrutura normalmente encontrada em artigos científicos.

Capítulo 2

O *Splicing* Alternativo no Câncer

2.1 Introdução

O sequenciamento de próxima geração (*Next-generation of sequencing*) tem permitido a exploração da complexidade do transcriptoma humano por meio de tecnologias de RNASeq (Martin e Wang , 2011; Mutz *et al.*, 2012). O que era uma sonho há 10 anos atrás, 20 milhões de *reads* para o transcriptoma de uma célula ou um tecido, por exemplo, pode ser atualmente considerado algo comum. Esse avanço tecnológico tem trazido diferentes desafios relacionados com a bioinformática, uma vez que os *reads* sequenciados são menores que os ESTs convencionais, o que torna a identificação de variantes de *splicing* problemática. Apesar dos desafios, o RNASeq tem sido amplamente utilizado em uma variedade de modelos e condições experimentais.

O tema "*splicing* alternativo" vem sendo bastante explorado, principalmente no estudo de doenças. No câncer, o impacto do *splicing* alternativo é bastante claro e reconhecido. Em um futuro próximo, acredita-se que pesquisas sobre esse assunto terão uma profunda consequência sobre práticas clínicas ao redor do mundo.

Nesse capítulo, uma ampla revisão sobre o *splicing* alternativo é apresentada com uma ênfase especial sobre o seu impacto na pesquisa do câncer. Esse trabalho será publicado em breve como um capítulo do livro "Book on Genomics and Drug Discovery" (*River Publishers*, Dinamarca).

2.2 O *Splicing*

A natureza "fragmentada" da ampla maioria dos genes eucarióticos, os forçam a sofrer um processo pós-transcricional, denominado de *splicing*, no qual íntrons são removidos e éxons são unidos em uma única sequência (Black, 2003). Esse processo é conduzido pelo spliceossomo, um complexo molecular composto por ribonucleoproteínas (RNPs), pequenos RNAs nucleares (snRNA: U1, U2,U4, U5 e U6), e por mais de 150 proteínas (Deckert *et al.*, 2006; Hartmuth *et al.*, 2002; Jurica e Moore, 2003; Zhou *et al.*, 2002).

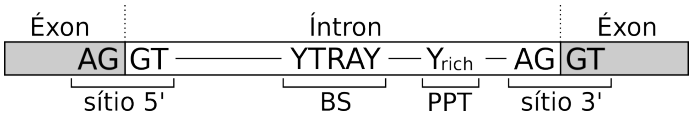

Figura 2.1: *Sinais de* splicing *mais importantes.*

O spliceossomo reconhece éxons e íntrons por meio de sinais presentes em *cis* ao longo da molécula de RNA. Os sinais de *splicing* mais importantes são os sítios 5' (doador) e 3' (aceptor), que são conservados em mais de 95% de todas as bordas éxon-íntron (Deckert *et al.*, 2006; Hartmuth *et al.*, 2002); o sítio ramificado (*branch site*, BS); e o trato de polipirimidina (PPT), localizado *upstream* ao sítio 3' de *splice* (Black, 2003; Graveley, 2001)(figura 2.1). Em metazoários, esses sinais não são suficientes para promover o processo de *splicing* e podem representar menos da metade da informação necessária para tal (Lim e Burge, 2001).

O processo de *splicing* pode ser regulado por elementos regulatórios *cis-acting*, os quais apresentam sequências curtas (entre 4 e 18 nucleotídeos) que podem interagir com fatores *trans-acting*,

determinando, consequentemente, se um sítio de *splice* será utilizado ou não. Esses elementos regulatórios de *splicing* (SRE), são encontrados em sequências intrônicas e exônicas, normalmente próximos à bordas de *splicing*, e podem ser classificadas como estimuladores (*enhancers*; ESE e ISE) ou silenciadores (*silencers*; ESS e ISS). Eles são necessários para o *splicing* constitutivo, assim como para a regulação do *splicing* alternativo (Blencowe, 2000; Cáceres e Kornblihtt, 2002; Cartegni *et al.*, 2002; Fairbrother *et al.*, 2002; Graveley, 2000; Woodley e Valcárcel, 2002). Os reguladores mais conhecidos são as proteínas ricas em serina/arginina (*serine/arginine-rich*, SR) e as ribonucleoproteínas heterogêneas (*heterogeneous nuclear ribonucleoproteins*, hnRNP)(Sanford *et al.* , 2005; Singh e Valcárcel, 2005). Proteínas SR são conhecidas por se ligarem aos ESEs e promoverem o *splicing*, enquanto as hnRNPs são conhecidas por se ligarem aos ESSs e ISSs, reprimindo o reconhecimento de sítios de *splice* adjacentes (Cartegni *et al.*, 2002).

2.3 O *Splicing* Alternativo

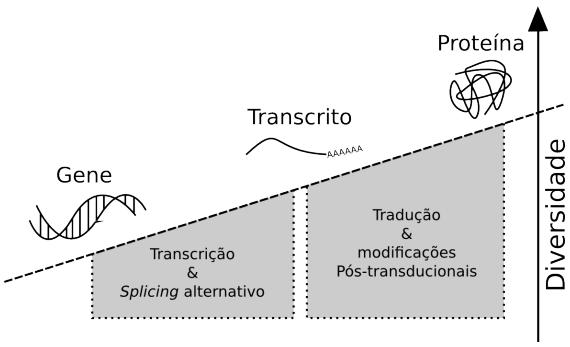

Figura 2.2: *A diversidade molecular aumenta conforme a informação passa por diferentes formas de processamento.*

O *splicing* alternativo é definido, basicamente, como um processo no qual moléculas idênticas de pré-mRNA são processadas de diferentes formas. Ele é um processo fundamental em organismos complexos (Black, 2003; Maniatis e Tasic, 2002), pois é responsável por criar uma ampla diversidade de proteínas a partir de um número relativamente pequeno de genes (Cork *et al.*, 2012)(figura 2.2) e, também, por afetar a transcrição em sua eficiência e estabilidade.

A ampla maioria dos genes humanos, assim como anotados pelo Consórcio Internacional de Sequênciamento do Genoma Humano (Collins *et al.*, 2004), são capazes de transcrever mais de um mRNA a partir do *splicing* alternativo (Johnson *et al.*, 2003). Diferentes eventos resultantes desse processo podem ser observados no desenvolvimento humano (Black e Grabowski, 2003; Venables, 2002), no qual aproximadamente 18% dos eventos parecem ser tecido-específicos (Markovic e Grammatopoulos, 2009).

Os eventos de *splicing* alternativo são capazes de modular a atividade de diversos processos celulares (Hsu e Hertel, 2009), e anormalidades nesse processo podem resultar em diferentes doenças, como o câncer (Kirschbaum-Slager *et al.*, 2005; Venables, 2004; Wang e Cooper, 2007), isquemia (Daoud *et al.*, 2002) e outros tipos de desordens humanas (Faustino e Cooper, 2003; Garcia-Blanco *et al.*, 2004; Pagani e Baralle, 2004).

2.3.1 Tipos de Eventos de *Splicing* Alternativo

Cinco tipos simples de *splicing* alternativo são atualmente conhecidos: éxon *skipping*, sítios alternativos 5' e 3' de *splice*, retenção de íntron, e sítio duplamente específico de *splice* (*dual-specific splice sites*, DSSs), este último identificado e caracterizado apenas recentemente (Zhang *et al.*, 2007). O éxon *skipping* é identificado quando um éxon é removido de um transcrito e os sítios alternativos 5' e 3' são identificados quando diferentes sítios 5' ou 3' são utilizados por um éxon, respectivamente. O evento de retenção de íntron é identificado quando um íntron não é removido e continua presente no mRNA maduro (Kim *et al.*, 2007; Sugnet *et al.*, 2004). Por fim, o sítio duplamente específico de *splice*, o tipo mais incomum de evento, é caracterizado quando um sítio de *splice* pode eventualmente atuar como 5' ou 3' (Zhang *et al.*, 2007)(figura 2.3)

Eventos complexos de *splicing* alternativo ocorrem quando dois ou mais eventos são observados em um mesmo mRNA ou em diferentes mRNAs (éxons mutualmente exclusivos, por exemplo)

(Kroll *et al.*, 2012). Esses eventos existem devido à complexa regulação do *splicing* e à alta frequência de eventos simples de *splicing* alternativo. O caso mais radical de evento complexo, por exemplo, ocorre no gene *Dscam* em Drosophila. Esse gene contém um grupo de 48 éxons mutualmente excludentes que, em princípio, pode gerar milhares de variantes de *splicing* (Schmucker *et al.*, 2000).

Padrões anormais e complexos de *splicing* alternativo tem sido diretamente relacionados com o câncer (Venables, 2006). Por exemplo, aproximadamente 46% e 12% de todas as sequências variantes expressas em tecidos normais e patológicos humanos apresentam mais de 1 e 2 eventos, respectivamente (Kroll *et al.*, 2012). Um dos eventos complexos mais conhecidos são os éxons mutualmente excludentes, no qual a presença de um éxon promove o *skipping* de um éxon adjacente, e vice-versa (figura 2.3). O evento complexo mais frequente, porém, é o *skipping* de múltiplos éxons adjancentes. Promotores alternativos e poliadenilação alternativa também podem ser considerados como eventos complexos (Beaudoing *et al.*, 2000; Black, 2003; Breitbart *et al.*, 1987; Letunic *et al.*, 2002).

2.3.2 Identificação de Variantes de *Splicing*

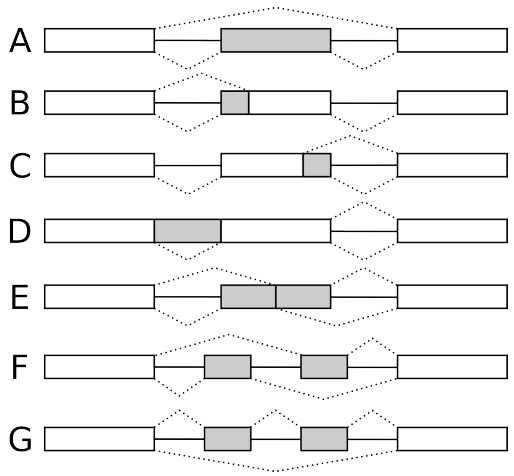

Figura 2.3: *Eventos de* splicing *alternativo: (A): éxon* skipping*; (B) borda alternativa 3'; (C) borda alternativa 5'; (D) retenção de íntron; (E) sítio duplamente específico; (F) éxons mutuamente exclusivos; (G) éxon* skipping *de dois éxons adjacentes.*

A identificação de eventos de *splicing* alternativo não é trivial. Recentemente, análises genômicas baseadas em ESTs (*expressed sequence tags*), *microarrays* e RNASeq se tornaram métodos padrão no estudo desses eventos. Banco de dados de ESTs e *microarrays*, por exemplo, já foram extensivamente explorados por alguns estudos (Brett *et al.*, 2000, 2002; Modrek e Lee, 2002). As primeiras iniciativas utilizando grandes quantidades de dados foram feitas a partir de ESTs, encontradas em banco de dados como o dbEST que atualmente possui mais de 20 millhões de sequências humanas (Benson *et al.*, 2012).

Sequências expressas são consideradas uma rica fonte para a identificação de eventos de *splicing* alternativo. Através de um simples alinhamento de uma EST contra um genoma de referência é possível detectar qualquer tipo de evento, incluindo diversos casos complexos. Diferentes algoritmos já foram desenvolvidos para o alinhamento de sequências expressas, como o SIM4 (Florea *et al.*, 1998) e o BLAT (Kent, 2002a). Esses algoritmos normalmente levam em consideração bordas de sítios de *splice*, necessárias para diferenciar *indels* (inserções/deleções) de regiões intrônicas (Feng *et al.*, 2012).

Uma alternativa bastante comum para ESTs são os *microarrays*. Grande parte dos *arrays* são baseados em técnicas de hibridização diferencial, que consideram sequências exônicas e junções éxon-éxon (Blencowe *et al.*, 2006) e permitem a identificação de padrões de *splicing* com uma precisão considerável (Li *et al.*, 2006; Zhang *et al.*, 2012, 2013). Porém, *microarrays* convencionais são limitados porque suas sondas são desenvolvidas para se ligar a sequências complementares. Isso significa que apenas eventos de *splicing* alternativo conhecidos podem ser detectados (Malone e Oliver , 2011). O método de hibridização *tiling array*, por outro lado, tem demonstrado ser uma melhor escolha uma vez que ele permite a descoberta de novos transcritos (Bertone *et al.*, 2004). Essa tecnologia tem sido explorada por diversos estudos para identificar padrões de expressão entre amostras, bem como para identificar eventos de *splicing* e também para descobrir novos biomarcadores (Rajan *et al.*, 2009; Wang *et al.*, 2008).

Finalmente, plataformas de sequenciamento de próxima geração (*next-generation of sequencing*, NGS) apresentam um alto rendimento na análise de transcriptomas (RNASeq)(Feng *et al.*, 2012;

Ozsolak e Milos, 2010). O RNASeq consiste, basicamente, em uma amostra de RNA purificada, que é convertida para uma biblioteca de cDNA e, posteriormente, sequenciada. Essas metodologias tem demonstrado ilimitadas possibilidades de análises: expressão de genes, mutações (SNVs), detecção de fusão de genes, quantificação absoluta e identificação de variantes de *splicing* (Chen *et al.*, 2011; Levin *et al.*, 2009; Sultan *et al.*, 2008). Atualmente, as plataformas Illumina, *Applied Biosystems* SOLiD e Roche 454 *Life Sciences* tem sido utilizadas para esse propósito (Feng *et al.*, 2012). Cada uma delas possui peculiaridades próprias, associadas com diferentes estratégias e ferramentas de análise.

Um vasto conhecimento em bioinformática é requerido em análises de RNASeq, e diversas ferramentas já foram desenvolvidas para esse propósito, como o Tophat, MMES, Split-Seek e SpliceMap (Ameur *et al.*, 2010; Au *et al.*, 2010; Trapnell *et al.*, 2009; Wang *et al.*, 2010). Primeiramente, esses algoritmos alinham as sequências contra um genoma referência. Os dados não mapeados são então alinhados no genoma baseados em junções éxon-éxon (figura 2.4). Por fim, o transcriptoma é reconstruído utilizando algoritmos como o Scripture (Guttman *et al.*, 2010) e Cufflinks (Trapnell *et al.* , 2010).

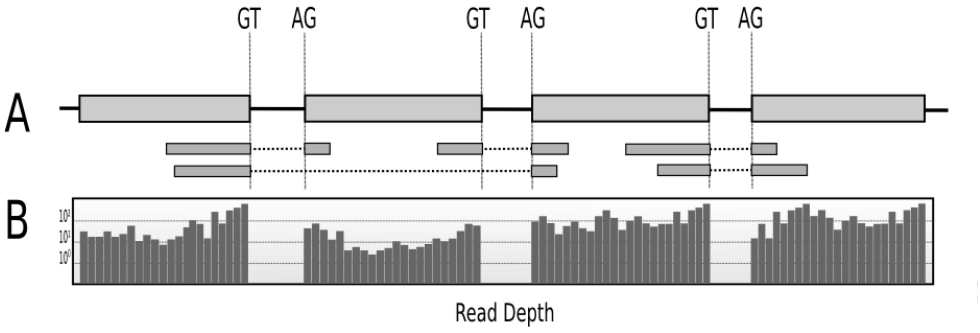

Figura 2.4: *Alinhamento de* reads *de RNASeq no genoma. (A) Identificação de bordas de* splicing*; (B) Cobertura vertical do sequenciamento.*

2.4 O *Splicing* Alternativo no Câncer

No câncer, um grande número de eventos de *splicing* alternativo já foram descritos (Srebrow e Kornblihtt, 2006; Venables, 2004), e muitos deles apresentam relação com mutações em sinais de *splicing* e elementos do complexo spliceossomal (Pajares *et al.*, 2007). Estudos mostraram, além do mais, que alguns eventos também podem ser o resultado da expressão diferencial de proteínas SR e hnRNPs (Jensen *et al.*, 2009), assim como o resultado de mecanismos epigenéticos que afetam a eficiência do reconhecimento dos sítios de *splice* (de la Mata *et al.*, 2003).

Atualmente, é difícil determinar se um evento de *splicing* alternativo é apenas um simples erro ou um evento que possui algum significado funcional. Muitos eventos introduzem *stop* códons prematuros nos transcritos, os quais podem ser rapidamente destruídos por meio de diferentes mecanismos (Frischmeyer *et al.*, 2002; McGlincy e Smith, 2008; Passos *et al.*, 2009). Apesar disso, é indiscutível a participação do *splicing* alternativo na modulação da atividade de diferentes tipos de proteínas (Srebrow e Kornblihtt, 2006; Wang e Cooper, 2007), como fatores de transcrição, transdutores de sinais celulares e componentes da matriz extracelular (Venables, 2004, 2006).

2.4.1 Mutações que Afetam Sinais e Elementos Regulatórios de *Splicing*

Diversas mutações que afetam sítios 5' e 3' de *splice* estão relacionadas com diferentes doenças e tipos de câncer, e são uma importante forma de promover o *splicing* alternativo. Os sítios 5' e 3' de *splice* são bastante conservados e sensíveis a quaisquer variações. Eles normalmente apresentam os dinucleotídeos GT e AG na região intrônica da sequência, respectivamente (Krawczak *et al.*, 1992, 2007), e mutações nessas regiões podem causar diferentes eventos de *splicing* alternativo, como éxon *skipping*, ativação de sítios crípticos de *splicing* e retenção de íntron (Ward e Cooper, 2010).

Um exemplo clássico é o gene *TP53*, o qual possui atividade de supressão tumoral e está, portanto, relacionado com o câncer. Dezenas de diferentes mutações em sítios de *splice* do gene *TP53* já foram reportadas para diferentes tipos de câncer. Curiosamente, algumas dessas mutações mostram ser "neutras" porque não alteram o aminoácido, porém, são capazes de afetar o *splicing* (Holmila *et al.*, 2003). O dado mais recente disponível para o gene *TP53* mostrou uma nova mutação em sítio de *splice* relacionada com o desenvolvimento de osteossarcoma. Análises observaram uma mutação no sítio 5' do íntron 6 (transição de G para A), que criou uma inserção de 6 aminoácidos (Sakurai *et al.*, 2013). Outros exemplos incluem os genes *SMARCB1*, *MLH1*, *ATM*, *BRCA1*, *NF2*, os quais foram identificados por diferentes estudos (Broeks *et al.*, 2003; De Klein *et al.*, 1998; Kurahashi *et al.*, 1995; Tanko *et al.*, 2002; Taylor *et al.*, 2000).

Atualmente, devido aos avanços tecnológicos, análises *genome-wide* tem sido possíveis, permitindo a descoberta de diversos genes relacionados com o câncer e de novos eventos de *splicing* alternativo. Um recente estudo, por exemplo, fez o sequenciamento genômico e transcriptômico de 19 amostras de câncer de pulmão e de três pares de amostras normais/tumorais do mesmo tecido. No total, 106 mutações em sítios de *splice* foram associadas com o *splicing* alternativo em diversos genes relacionados com o câncer (Liu *et al.*, 2012).

Algumas mutações também podem afetar o sinal PPT, alvo da PTB (*polypyrimidine tract-binding protein*), que possui diversos papéis no processamento do RNA, como no *splicing* alternativo, localização do mRNA e tradução (Shibayama *et al.*, 2009). O *MLH1* em câncer coloretal, por exemplo, apresentou um sinal PPT mutante, causando um éxon *skipping* relacionado com a indução do desenvolvimento do câncer (Clarke *et al.*, 2000).

Embora mutações em reguladores de *splicing* não apresentem uma influência direta sobre eventos de *splicing* alternativo, os mesmos podem eventualmente acontecer. Um exemplo é o gene *KLF6*, um fator de transcrição conhecido por suprimir tumores. Um SNP (*single nucleotide polymorphism*) criou um novo sítio de ligação para a proteína SR, *SRp40*, o que aumentou a expressão da isoforma *KLF6-SV1* (Narla *et al.*, 2005). Essa isoforma não apresenta um domínio *zinc finger* (DiFeo *et al.*, 2009), antagonizando, consequentemente, a isoforma *wild-type* da *KLF6*. Portanto, a superexpressão da *KLF6-SV1* acelera a progressão do câncer e da metástase, e o polimorfismo associado com o câncer de próstata sugere um maior risco de câncer (Ward e Cooper, 2010).

2.4.2 Fatores de *Splicing* Afetados no Câncer

Proteínas SR são um dos fatores mais importantes de *splicing*, e elas podem atuar como *enhancers* ou *silencers* por interferirem na construção do complexo spliceossomal (Maas *et al.*, 2001; Tacke e Manley, 1999). Pequenas alterações nos níveis de expressão das proteínas SR podem desregular os eventos de *splicing* alternativo, afetando de maneira geral o comportamento celular (Ghigna *et al.*, 1998; Mukherji *et al.*, 2006; Pind e Watson, 2003; Stickeler *et al.*, 1999). Análises mostraram que os níveis de expressão das proteínas SR são menores durante o desenvolvimento tumoral (Ghigna *et al.*, 1998).

Em adenocarcinoma de mama de camundongos, a expressão diferenciada das proteínas SR alteram o padrão de *splicing* do gene *CD44*, o qual é induzido a expressar mRNAs variantes (Naor *et al.*, 2002; Stickeler *et al.*, 1999). Algumas proteínas SRs, como a *PTBP*, mostrou níveis elevados durante o desenvolvimento tumoral em ovário, enquanto outros fatores, como o *SRSF1* e *U2AF65*, não mostraram variações nos respectivos níveis de expressão (He *et al.*, 2007). Existe uma clara associação entre níveis de expressão das proteínas SR específicas e o câncer, porém, os mecanismos por trás dessa regulação ainda precisam ser melhor elucidados (Khan *et al.*, 2012).

2.4.3 Famílias de Proteínas Afetadas pelo *Splicing* Alternativo no Câncer

O *splicing* alternativo presente em algumas famílias de proteínas tem um profundo impacto na fisiologia de células e tecidos. Certas famílias de proteínas estão relacionadas com processos de regulação. Por exemplo, o gene *DNMT3b*, que codifica uma DNA metiltransferase, não apresenta um éxon. Esse evento resulta em uma proteína truncada, relacionada com a hipometilação de regiões

pericentroméricas do genoma em tumores de fígado (Saito *et al.*, 2002). Outro exemplo é o gene *PASG*, homólogo do gene *SNF2* e responsável por codificar um proteína remodeladora da cromatina. Um evento específico de borda alternativa no sítio 5' mostrou uma alta prevalência em amostras de leucemia aguda (Lee *et al.*, 2000). A disfunção dessas proteínas afetam de forma epigenética a eficiência do reconhecimento dos sítios de *splice* pelo complexo spliceossomal (de la Mata *et al.*, 2003).

Fatores de transcrição são alvos frequentes do *splicing* alternativo. Variantes de *splicing* do gene *NRSF*, um fator de silenciamento de *splicing*, são conhecidos por codificar proteínas truncadas em câncer de pulmão (Coulson *et al.*, 2000). Outros exemplos estão relacionados com hormônios, como o observado no gene *AIB1*, que codifica um hormônio coativador de receptor nuclear. Em amostras tumorais, uma grande quantidade de variantes do gene *AIB1* que não apresentam o éxon 3 pode ser observada. Esses variantes são conhecidos por promoverem a transcrição mediada por receptores de estrógeno em um processo que está relacionado com o desenvolvimento do câncer de mama (Reiter *et al.*, 2001). Formas variantes dos receptores de andrógeno (AR) também já foram observadas no câncer de mama (Zhu *et al.*, 1997).

Muitos eventos de *splicing* alternativo podem ser observados para proteínas de superfície celular, as quais são expressas por genes como o *CD44* e por genes da família dos *FGFRs* (*fibroblast growth factor receptors*). Entre outras famílias de proteínas, temos o gene *SVH*, superexpresso em câncer de fígado e apresenta um papel importante no crescimento e sobrevivência celular. Entre alguns de seus variantes, apenas o *SVH-B* é observado em tumores, e a sua inibição pode causar apoptose (Huang *et al.*, 2003). Algumas proteínas da superfície possuem função de adesão. Um exemplo é o gene *MUC1*, que apresenta uma função essencial na formação de barreiras mucosas em superfícies epiteliais e está envolvido no processo de metástase. Em câncer de tiróide, o gene *MUC1* pode apresentar um novo éxon críptico (Weiss *et al.*, 1996).

Uma outra classe importante de proteínas está relacionada com a sinalização celular, especialmente as proteínas solúveis, as quais podem transmitir sinais oncogênicos. O gene *NF1*, por exemplo, é um supressor de tumor que interfere na cascata de sinalização do sinal *Ras*. Em meduloblastomas e tumores neuroectodermais primitivos, variantes mais fracos do supressor *NF1* são predominantes, diferentemente do observado em tecidos normais de cérebro (Scheurlen e Senf, 1995). Outro exemplo é o gene *SYK*, codificador de uma tirosina quinase envolvida na supressão da metástase. Em câncer de mama, um variante específico do gene *SYK* mostrou ser incapaz de prevenir a metástase (Wang *et al.*, 2003).

Finalmente, as proteínas extracelulares são normalmente secretadas na matriz extracelular e possuem um papel essencial no desenvolvimento da metástase. O *uPG* (*urokinase-type plasminogen activator*), por exemplo, é uma proteína secretada, e alguns de seus variantes estão envolvidos na degradação de proteínas da matriz extracelular (Luther *et al.*, 2003). Assim como o *uPG*, o gene *WISP1* pode apresentar um variante capaz de causar a invasão celular em carcinoma gástrico (Tanaka *et al.*, 2001).

2.4.4 Variantes de *Splicing* como Marcadores de Câncer

Biomarcadores baseados em variantes de *splicing* alternativo câncer-específicos são de grande importância, uma vez que podem ser utilizados com objetivos diagnósticos e/ou terapêuticos. Anticorpos monoclonais para alvos câncer-específicos tem sido utilizados desde os anos 70 contra tumores sólidos (Halin *et al.*, 2001). Em 2001, aproximadamente 700 anticorpos monoclonais contra o câncer se encontravam em testes clínicos, patrocinados por mais de 200 laboratórios de biotecnologia (Walsh, 2010). Atualmente, devido ao rápido desenvolvimento científico e tecnológico, esses números devem superar dezenas de milhares.

Os genes do projeto Surfaceoma (Da Cunha *et al.*, 2009) codificam proteínas que se encontram na superfície celular, onde elas apresentam um importante papel na comunicação intercelular e podem ser facilmente identificadas por anticorpos. Muitas dessas proteínas estão relacionadas com o desenvolvimento do câncer, no qual muitos casos específicos de *splicing* alternativo já foram observados. Entre os genes de superfície celular mais comumente afetados, temos o famoso *CD44* e

os genes da família *FGFR*.

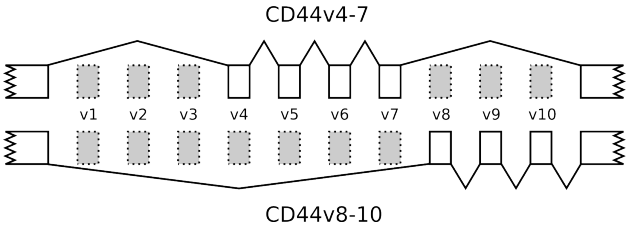

Figura 2.5: *Variantes de* splicing *alternativo do gene CD44 relacionados com o câncer.*

O gene *CD44* possui os eventos de *splicing* alternativo mais estudados em câncer. Ele codifica mais de 20 isoformas de proteínas conhecidas, resultantes da incorporação variada de 10 éxons que se encontram próximos à domínios extracelulares. Suas formas normais e prevalentes não apresentam eventos alternativos, e as isoformas associadas ao câncer normalmente são caracterizadas pela presença dos éxons v4-7 ou v8-10 (Venables, 2004)(figura 2.5). Muitas dessas isoformas são expressas em diferentes tipos de câncer e em diferentes espécies animais, e são, consequentemente, importantes alvos terapêuticos. Anticorpos contra isoformas do gene *CD44* já mostraram ser capazes de reduzirem a habilidade do tumor em desenvolver metástase e resistência contra diferentes moléculas terapêuticas (Kerbel, 2000), embora limitações contra tumores sólidos já foram observadas (Halin *et al.*, 2001). Clinicamente, por meio da presença de algumas variantes de *CD44* no sangue, é possível identificar e diferenciar formas tumorais. Por exemplo, a baixa expressão do variante *CD44*v6 está relacionada com a malignidade tumoral e um prognóstico favorável em câncer de próstata. Por outro lado, tumores benignos de próstata superexpressam o variante *CD44*v7-9 (Aaltomaa *et al.*, 2001; Iczkowski *et al.*, 2003).

Proteínas da família *FGFR*, por sua vez, tem sido detectadas em células normais e malignas, e possuem um papel crucial na diferenciação e desenvolvimento. Todos os genes desse grupo sofrem eventos de *splicing* alternativo, os quais são necessários para regular a especificidade de ligação entre as *FGFRs* e os *FGFs* (*fibroblast growth factors*) (Yeh *et al.*, 2003). Uma característica interessante é que *FGFRs* solúveis, que não apresentam domínios transmembrana devido a eventos de *splicing* alternativo incomuns, tem sido identificadas em amostras de sangue em pacientes com câncer de mama (Jang, 2002). Alguns exemplos, como os genes *FGFR1* (isoforma beta), *FGFR3* e *FGFR2*, mostraram um importante papel no desenvolvimento do câncer (Vickers *et al.*, 2002), tumorigênese coloretal (Jang *et al.*, 2001) e câncer de próstata (Kwabi-Addo *et al.*, 2001), respectivamente.

Por fim, outra família interessante de proteínas é a *GPCR*, uma superfamília de receptores de transmembrana, que possuem uma larga distribuição e capacidade de identificar um grande número de ligantes. Consequentemente, a ativação e inibição dos sinais das *GPCRs* pode afetar diversos processos patofisiológicos (Markovic e Challiss, 2009). Sabe-se que aproximadamente 50% dos genes da família *GPCR* não possuem íntrons em regiões codificadoras, e os demais genes podem sofrer *splicing* alternativo e gerar diferentes isoformas de proteínas, diferindo em suas propriedades de sinalização e regulação (Markovic e Challiss, 2009). Esses eventos tem sido considerados de grande importância, uma vez que já foram identificados em diferentes tipos de câncer, incluindo melanoma, câncer de mama, ovário, próstata, fígado e câncer gastrointestinal (Hellmich *et al.*, 2000; Lee *et al.*, 2008; Srebrow e Kornblihtt, 2006).

2.5 Perspectivas

A disponibilidade do genoma e sequências do transcriptoma de milhares de tumores humanos permitirá a identificação de diversas alterações genéticas envolvidas no câncer. Juntamente com dados clínicos, essas informações representarão uma excelente fonte para o desenvolvimento de novas estratégias diagnósticas e terapêuticas. Em um futuro próximo, variantes de *splicing* surgirão como importante alvos para tais estratégias.

Capítulo 3

SPLOOCE

3.1 Introdução

Figura 3.1: *Logotipo do SPLOOCE.*

Eventos de *splicing* alternativo (ASEs) estão presentes em quase todos os genes humanos que apresentam mais de um éxon (Pan *et al.*, 2008; Wang *et al.*, 2008), e são considerados como um dos componentes mais significativos por trás da complexidade de organismos multicelulares (Galante *et al.*, 2004; Modrek e Lee, 2002; Wang *et al.*, 2008). Além do mais, os ASEs estão envolvidos na etiologia de uma ampla variedade de doenças humanas (Garcia-Blanco *et al.*, 2004), e a regulação do *splicing* constitutivo e alternativo é feita por uma complexa rede de elementos celulares, da qual fazem parte fatores *trans-acting* e sequências *cis-acting* encontradas no RNA primário (Alló *et al.*, 2009; Batsché *et al.*, 2005; Ip *et al.*, 2011; Luco *et al.*, 2010; Muñoz *et al.* , 2009; Saint-André *et al.*, 2011; Sakabe e de Souza, 2007; Schor *et al.*, 2009).

A complexa regulação do *splicing* e a alta frequência de ASEs explicam o surgimento de Eventos Complexos de *Splicing* Alternativo (CASEs), que apresentam uma combinação regulada de dois ou mais ASEs em transcritos de um gene, ou até no mesmo transcrito. O exemplo mais dramático de CASE ocorre no gene *Dscam* em *Drosophila*. Esse gene contém um grupo de 48 éxons mutualmente excludentes que, em princípio, pode gerar milhares de variantes de *splicing* (Schmucker *et al.*, 2000).

Em humanos, alguns ASEs e CASEs que ocorrem em oncogenes e supressores de tumor já foram associados com o câncer (Galiana-Arnoux *et al.*, 2005; Hayes *et al.*, 2004; Tanko *et al.*, 2002; Venables, 2004, 2006). Por exemplo, o gene *NTRK1* (*nerve growth factor*) possui uma sequência variante, *TrkAIII*, que é comum em determinados tumores e não apresenta três éxons que codificam um domínio regulatório do tipo *immunoglobulin-like*, importante por interações proteína-proteína e proteína-ligante(Tacconelli *et al.*, 2004). Entre outros exemplos de CASEs, o gene *CD44* é um conhecido marcador de malignidade e invasibilidade, e possui aproximadamente 10 ASEs que podem ocorrer em diferentes combinações em uma região responsável por codificar um domínio extracelular (Galiana-Arnoux *et al.*, 2005; Hayes *et al.*, 2004; Venables, 2006).

Apesar de esforços anteriores (Malko *et al.*, 2006; Nagasaki *et al.*, 2006; Sammeth *et al.*, 2008), uma simples e eficiente nomenclatura para levar em consideração todas as variações geradas pelo *splicing* alternativo ainda não existe. Nesse tabalho, uma nova ferramenta (*web-tool & database*), denominada de SPLOOCE[1], foi desenvolvida baseada em expressões regulares (*regexes*) associada à uma sintaxe. O SPLOOCE fornece uma série de ferramentas que permitem ao usuário identificar e analisar variantes de *splicing* (simples e complexas) e seus impactos funcionais.

[1] http://www.bioinformatics-brazil.org/splooce/

3.2 Objetivos

O objetivo geral desse capítulo é desenvolver um conjunto de estratégias que permita a exploração da complexidade apresentada pelo transcriptoma humano ou por qualquer outro organismo complexo. Para isso, os seguintes objetivos secundários foram definidos:

- Desenvolver um algoritmo de análise de *splicing* alternativo capaz de identificar ASEs e CASEs. Possibilitar também a identificação de sítios duplamente específicos, os quais são poucos estudados e, portanto, raramente abordados em outros banco de dados;
- Criar um banco de dados completo de *splicing* alternativo para sequências expressas humanas, baseadas em sequências clássicas, como RefSeqs, mRNAs e ESTs. Adicionalmente, utilizar dados obtidos de sequenciamentos de *next-generation* (NGS) devido ao grande impacto dessas tecnologias na ciência atual;
- Desenvolver uma sintaxe capaz de representar todos os eventos de *splicing* alternativo e suas possíveis combinações. Com o rápido desenvolvimento das tecnologia de NGS e a grande quantidade de dados sendo progressivamente sequenciada, padrões de *splicing* cada vez mais complexos deverão ser observados ao longo do tempo. Por essa razão, formas alternativas e práticas de representações serão necessárias;
- Desenvolver uma ferramenta *web* (SPLOOCE) implementando todos os itens citados anteriormente para facilitar a disponibilização da tecnologia criada.
- Analisar de forma geral e exploratória os dados gerados pelo SPLOOCE.

3.2.1 Justificativa

Eventos de *splicing* alternativo podem afetar a estrutura de proteínas e possuem um importante papel na modulação do fenótipo celular. Milhares de eventos são atualmente conhecidos, sem considerar os eventos complexos (CASEs), os quais são praticamente imensuráveis. Esse trabalho representa um grande avanço no estudo da complexidade do transcriptoma humano, visto que pouco tem sido feito em relação aos eventos complexos. Além do mais, atualmente, pouco se sabe sobre a função da ampla maioria dos eventos de *splicing* alternativo. Portanto, novas ferramentas de análise são extremamente necessárias para explorar e desvendar o "universo" molecular presente nas células humanas.

3.3 Notas

Publicação: O SPLOOCE foi publicado no final do ano de 2012 (Kroll *et al.*, 2012) na *RNA Biology* (IF 4.8). Tang *et al.* (2013) citou o SPLOOCE como sendo o único servidor de análise de *splicing* alternativo, enquanto os demais "serviços", publicados entre 2001–2013, foram considerados como bancos de dados. Mais recentemente, Pohl *et al.* (2013) destacou o SPLOOCE e sua sintaxe como importantes ferramentas no estudo dos éxons mutualmente exclusivos. O artigo oficial do SPLOOCE está anexado no apêndice A.

3.4 Materias & Métodos

Visto que grande parte desse capítulo está relacionada com o desenvolvimento de metodologias, essa seção mostra, portanto, as etapas essenciais para a análise do *splicing* alternativo, no que diz respeito à obtenção e ao pré-tratamento de sequências expressas.

Dados públicos

O genoma humano de referência (NCBI36/hg18) foi gravado do portal da UCSC *Genome Bioinformatics*[2]. As sequências RefSeq foram gravadas do banco de dados de Sequências de Referência, versão 25[3]. Sequências do tipo mRNA foram gravadas do portal da UCSC *Genome Bioinformatics* (*Homo sapiens*), e as sequências do tipo ESTs foram gravadas do dbEST[4]. Dados de RNAseq foram gravados do portal SRA[5]; Foram utilizados os IDs SRX003935, SRX003934, SRX003933, SRX003932, SRX003931, SRX003930, SRX003929, SRX003928, SRX003927 e SRX003926. Esses experimentos de NGS totalizam 63 corridas de sequenciamento (Illumina Genome Analyzer) e 211.006.871 *reads* (36 pb cada).

Alinhamento das sequências

Todas sequências dos tipos RefSeq, mRNA e EST foram alinhadas contra o genoma humano usando um protocolo já descrito anteriormente (Galante *et al.*, 2007). Resumidamente, todas as sequências foram mapeadas contra o genoma humano usando o alinhador BLAT (Kent , 2002b). Em seguida, os transcritos apresentando identidade maior que 95% e com uma cobertura maior que 90% foram remapeadas utilizando o SIM4 (Florea *et al.*, 1998). Os dados de RNASeq, por outro lado, foram mapeados contra o genoma humano usando o pipeline Tophat (Trapnell *et al.*, 2009), e todos os *reads* devidamente mapeados foram submetidos para o programa Cufflinks (Trapnell *et al.*, 2010) (os transcritos preditos foram denominados de NGS). Em ambos os programas optou-se por usar parâmetros padrão.

Agrupamento de sequências

Todas as sequências mapeadas que compartilharam a mesma região genômica foram agrupadas utilizando uma estratégia orientada ao gene. Primeiramente, todas as sequências RefSeq foram anotadas conforme os nomes oficiais dos genes correspondentes, assim como presente no NCBI-Gene[6]. Em seguida, foram agrupados para o mesmo gene todos os transcritos não-RefSeq (mRNAs, ESTs e RNAseq) que apresentaram múltiplos éxons e que compartilharam uma ou mais junções de *splicing* com pelo menos uma sequência RefSeq. Por fim, foram agrupados os transcritos não-RefSeq que apresentaram apenas um éxon e uma sobreposição maior que 30 nucleotídeos com um transcrito RefSeq. As informações sobre o agrupamento de sequências foram armazenadas em um banco de dados relacional.

Splicing Alternativo

As análises de ASEs e CASEs foram feitas utilizando expressões regulares, assim como detalhado ao longo desse trabalho (capítulo 3). Na análise de ASEs, apenas grupos de cDNA contendo RefSeqs e/ou mRNAs ou mais de 10 ESTs/RNASeqs foram utilizados. A redundância de ASEs, como éxons *skipping* e retenção de íntrons, foram eliminadas comparando todos os eventos similares de um gene. Posteriormente, todos os eventos que apresentaram sobreposição de posições foram agrupados e contados como apenas um evento. Os demais eventos, como os sítios 3'/5' alternativos e os sítios duplamente específicos de *splicing*, foram contados apenas verificando a posição genômica do sítio. O número de genes afetados por CASEs foi definido por meio da análise da lista completa de sequências "Comparativas" (descritas na seção 3.5), não considerando o número de sequências de suporte.

[2] http://genome.ucsc.edu
[3] http://www.ncbi.nlm.nih.gov/RefSeq/
[4] http://www.ncbi.nlm.nih.gov/dbEST/
[5] http://www.ncbi.nlm.nih.gov/sra
[6] http://www.ncbi.nlm.nih.gov/gene/

3.5 Identificando Eventos de *Splicing* Alternativo

3.5.1 Algoritmo

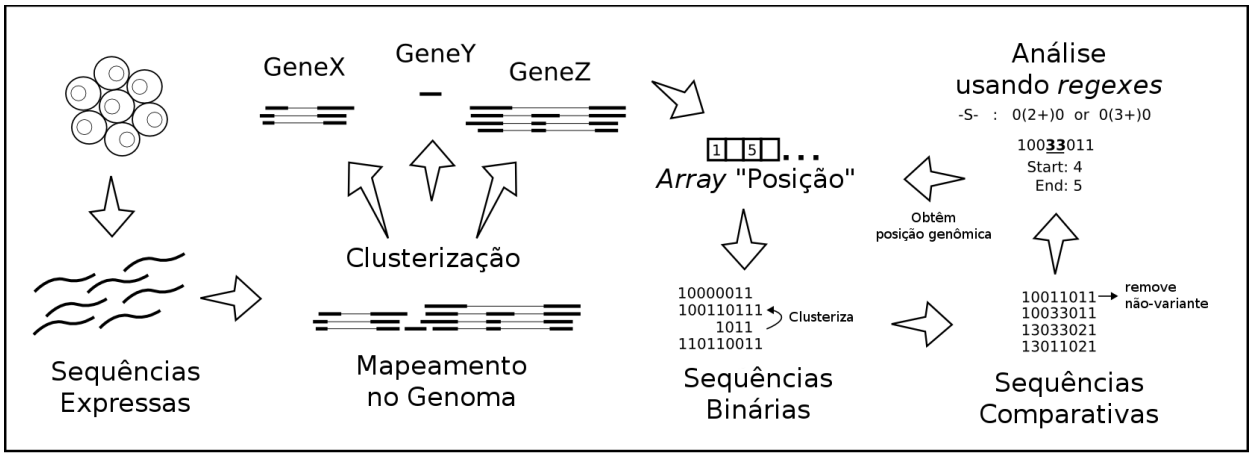

Figura 3.2: *Estratégia global para processar e analisar ASEs e CASEs a partir um grupo de sequências expressas.*

Para identificar e explorar ASEs e CASEs, foi primeiramente construído um banco de dados de genes e transcritos, assim como já descrito anteriormente (Galante *et al.*, 2004; Kirschbaum-Slager *et al.*, 2005). O banco de dados apresentou 19,558 genes anotados e, 7.348.127 sequências expressas, das quais, 25.684 (0,35%) são RefSeqs, 203.649 (2,77%) são mRNAs, 5.946.053 (80,92%) são ESTs e 1.172.761 (15,96%) são NGS (já devidamente processadas pelo Cufflinks). A média de transcritos por gene é 376. Uma vez que o método de análise de *splicing* alternativo é normalmente baseado em comparações par a par, pelo menos 1,4 bilhões de comparações eram esperadas para todos os genes. Porém, o tempo computacional foi drasticamente reduzido através do uso de sequências binárias que podem ser facilmente agrupadas, limitando, portanto, o número de sequências analisadas sem a perda de informações importantes.

Para o processo de agrupamento, todos os transcritos foram primeiramente convertidos para sequências binárias. O passo inicial foi inserir as posições genômicas iniciais e finais dos éxons em *arrays*, de acordo com seus respectivos genes (um *array* para cada gene). Para simplificar, esses *arrays* foram chamados de "Posição", os quais foram, então, numericamente organizados e as posições genômicas redundantes foram removidas. Entre todas as casas do *array* "Posição", novas casas com valores nulos foram adicionados, exceto quando a diferença entre o valor presente em um casa para a próxima casa era de apenas 1 (figura 3.4A).

Uma sequência binária refere-se à versão binária do transcrito mapeado, o qual, por sua vez, é representado por posições genômicas. Sequências binárias podem ser facilmente lidas por algoritmos, uma vez que éxons são simplesmente representados pelo número 1 e íntrons pelo número 0. As sequências binárias foram criadas comparando as posições genômi-

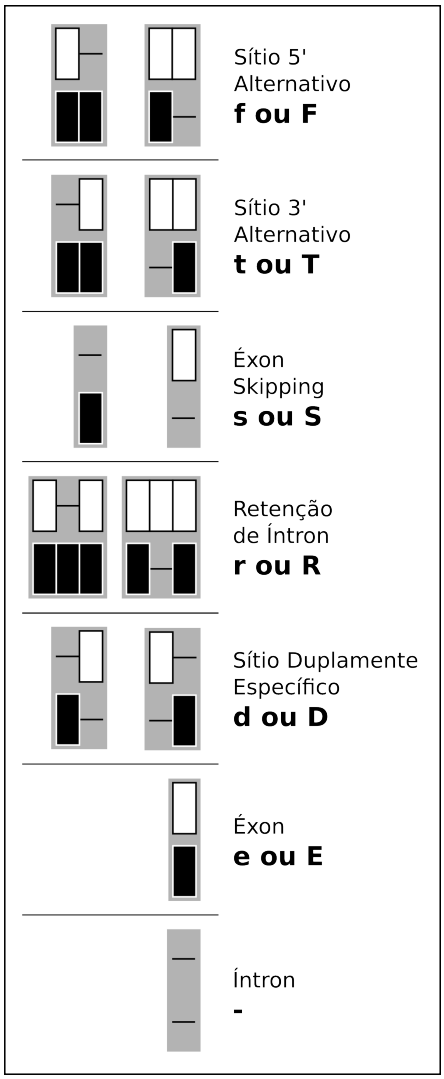

Figura 3.3: *Eventos de* splicing *alternativo na forma de "blocos" e suas respectivas sintaxes.*

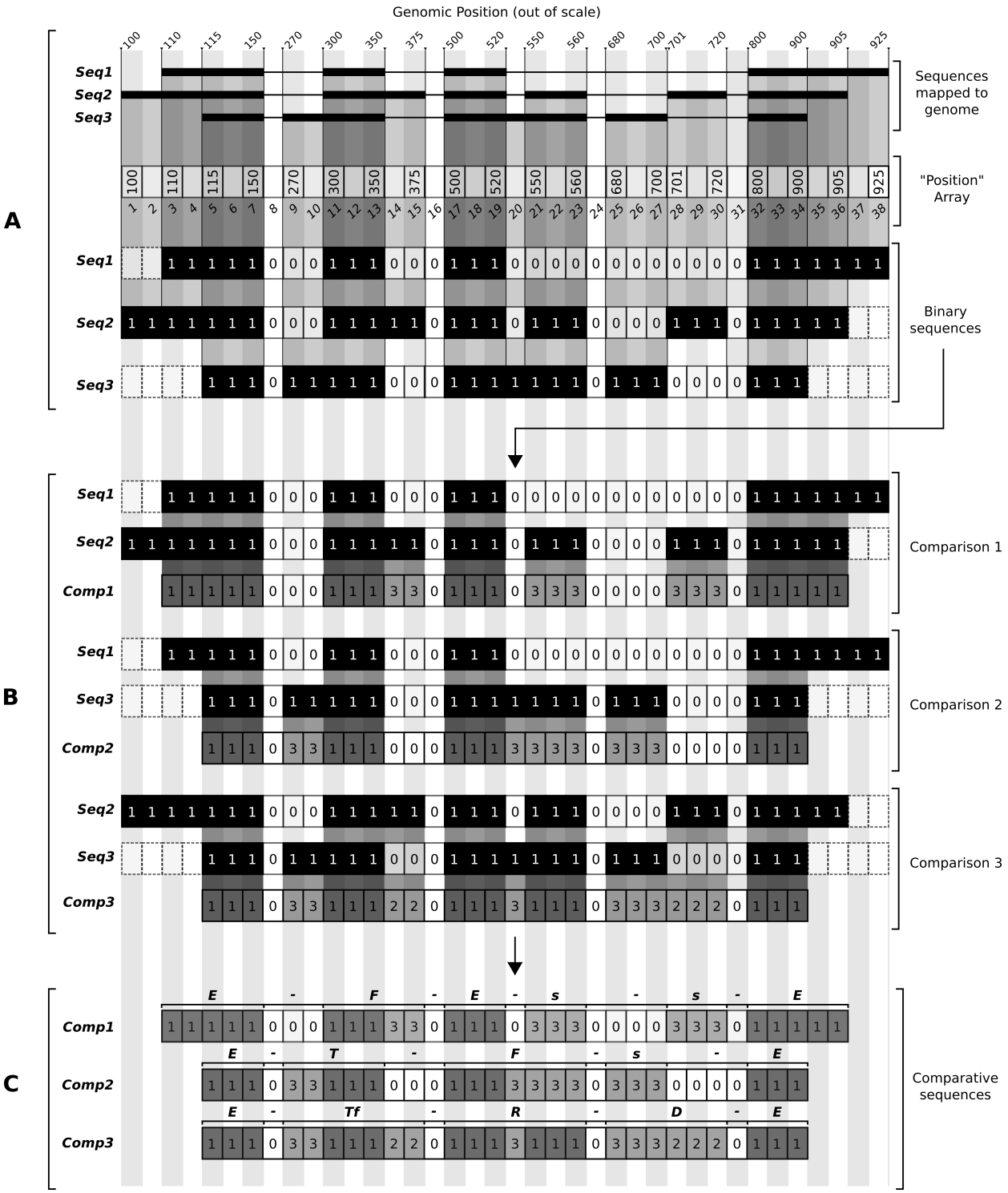

Figura 3.4: *Esquematização do método proposto. A: criação do* array *"Posição" e das sequências binárias a partir dos dados mapeados. B: comparação de duas sequências binárias para criar uma sequência "Comparativa". C: grupo de sequências "Comparativas" (e suas respectivas sintaxes), prontas para serem analisadas através de expressões regulares.*

cas de ínicio e fim dos éxons de cada transcrito contra o *array* "Posição" respectivo do gene. Por exemplo, um transcrito hipotético possui um éxon com posição genômica inicial e final de 10.000 e 20.000, respectivamente. Essas posições são encontradas nas casas de número 8 e 10 do *array* "Posição". Nesse caso, todas as casas entre 8 e 10 na sequência binária devem ser preenchidas com 1, representando um éxon, e todos os espaçoes encontrados entre os éxons (se houver mais de um) devem ser preenchido com 0, o qual é o carácter utilizado para representar íntrons. Esses passos

Evento de Splicing Alternativo	*Expressão Regular*	*Sintaxe*
Sítio 5' Alternativo	1(2+)0 ou 1(3+)0	"F"ou "f"
Sítio 3' Alternativo	0(2+)1 ou 0(3+)1	"T"ou "t"
Retenção de íntron	1(2+)1 ou 1(3+)1	"R"ou "r"
Éxon skipping	0(2+)0 ou 0(3+)0	"S"ou "s"
Sítio duplamente específico	0(2+3+)0 ou 0(3+2+)0	"D"ou "d"
Éxon compartilhado	0(1+)0	"E"
Íntron compartilhado	(0+)	"-"
Qualquer coisa	.*	*

Tabela 3.1: *Eventos de* splicing *alternativo e suas respectivas expressões regulares e sintaxes.*

devem ser repetidos para todos os transcritos e seus respectivos éxons (figura 3.4A). Essa estratégia de criação de sequências binárias é, até certo ponto, similar ao método "*bit matrix*" proposto por (Nagasaki *et al.*, 2006). Finalmente, o agrupamento foi feito comparando todas as sequências binárias, excluindo as que se encaixavam em sequências maiores, reduzindo, assim, o número de sequências no processo de comparação par a par.

O último passo foi a criação das sequências "Comparativas", as quais podem representar ASEs e CASEs. Elas foram criadas, comparando par a par, todas as sequências binárias de cada gene (figura 3.4B). Quando duas sequências binárias são comparadas (*Seq1* e *Seq2*, por exemplo), uma regra simples deve ser seguida: se ambas as sequências binárias possuem o número 1 ou 0 na mesma posição, esse mesmo valor deve ser mantido na respectiva sequência "comparativa". Por outro lado, quando os valores são diferentes, o número 2 é inserido na sequência "comparativa" se o valor 1 está presente na *Seq1*, ou o número 3 se o valor 1 estiver na *Seq2*. Após esse passo, todas as sequências "Comparativas" de cada gene foram agrupadas para remover possíveis redundâncias, assim como feito anteriormente para as sequências binárias. Para evitar o armazenamento de dados insignificantes, todas as sequências que não reportaram variações de *splicing* foram excluídas. No fim de todo o processo, um total de 4.154.216 sequências "comparativas" foram obtidas. Elas representam menos de 0,16% da quantidade total de sequências "Comparativas" possíveis que poderiam ser criadas a partir de um método não otimizado.

3.5.2 Expressões Regulares

As sequências "Comparativas", criadas anteriormente, foram propriamente analisadas utilizando expressões regulares (*regexes*). Para identificar simples ASEs, os mesmos necessitam ser convertidos para padrões *regex* (tabela 3.1). *Regexes* são capazes de identificar quaisquer ASEs conhecidos e estruturas não variantes, como éxons constitutivos e íntrons. Foi também possível a identificação de um tipo de ASE pouco usual conhecido como sítio duplamente específico (*dual-specific splice site*, DSS) (Zhang *et al.*, 2007).

Como mencionado, regiões exônicas e intrônicas compartilhadas são representadas por 1 e 0, respectivamente. Todas estruturas variantes são representadas por 2 ou 3, dependendo em qual sequência a variação estrutural ocorre. Isso é bastante útil quando éxons mutuamente excludentes são analisados, por exemplo. Para exemplificar, tomemos como exemplo o padrão "1101110220003330110001 1". Nesse caso, o evento mutuamente excludente é caracterizado pela sequência "0220003330" (íntron; éxon skipping A; íntron; éxon skipping B; íntron). O respectivo *regex* para esse evento é a expressão "0(2+)0+(3+)0", que pode capturar, por meio do uso de parêntesis, ambos os eventos de éxon skipping ("22" e "333") e retornar suas posicões na sequência binária. Se as posições genômicas são necessárias para a análise, elas podem ser facilmente obtidas através do *array* "Posição".

Basicamente, a construção de *regexes* para a análise de CASEs pode ser feita por meio da combinação de *regexes* de ASEs individuais (3.1). Por exemplo, para descrever um sítio 5' alternativo (A5'SS) seguido por um sítio 3' alternativo (A3'SS), os respectivos regexes "1(2+)0" e "0(2+)1" devem ser concatenados. O *regex* resultante é "1(2+)0+(2+)1", e não "1(2+)00(2+)1". A quantidade de números 0 (íntron) encontrados entre ambos os *regexes* deve ser indefinido. Portanto, ao invés

de “00”, a representação correta é “0+”, simplesmente porque essa estrutura pode ser formada por um ou vários caracteres. Porém, as estruturas encontradas nas extremidades dos *regexes* podem ser representadas por um número limitado de caracteres apenas se esses dados não forem importantes para a análise. Por exemplo, o *regex* “1(2+)0” pode ser utilizado para procurar eventos A5’SS. Se, além desse evento de *splicing*, toda a estrutura do éxon precisa ser definida (para evitar outros eventos no mesmo éxon, por exemplo), o *regex* deve ser reescrito como “01+(2+)0”.

3.5.3 Sintaxe

Para interpretar ASEs e CASEs de uma forma simples, foi desenvolvida uma representação baseada em caracteres para os diferentes padrões de *splicing* possíveis. Essa representação pode ser rapidamente aprendida, interpretada e visualizada. Ela é útil devido à dificuldade de representar vários CASEs graficamente, visto que existem limitações de espaço e problemas relacionados à complexidade gráfica. Basicamente, os eventos de *splicing* alternativo, assim como éxons e íntrons, são representados por diferentes caracteres (tabela 3.1). Como exemplo, um éxon *skipping* (ES) seguido por um A3’SS pode ser representado por “-s-t-”, onde “-” representa um íntron, “s” representa um evento de ES e “t” um evento A3’SS. Se os eventos estão em diferentes transcritos, a representação deve ser “-s-T-” ou “-S-t-”. Letras minúsculas e maiúsculas são usadas para diferenciar a sequência na qual o evento ocorre (figura 3.5). Essa estratégia permite a identificação de ASEs que ocorrem ou não no mesmo transcrito.

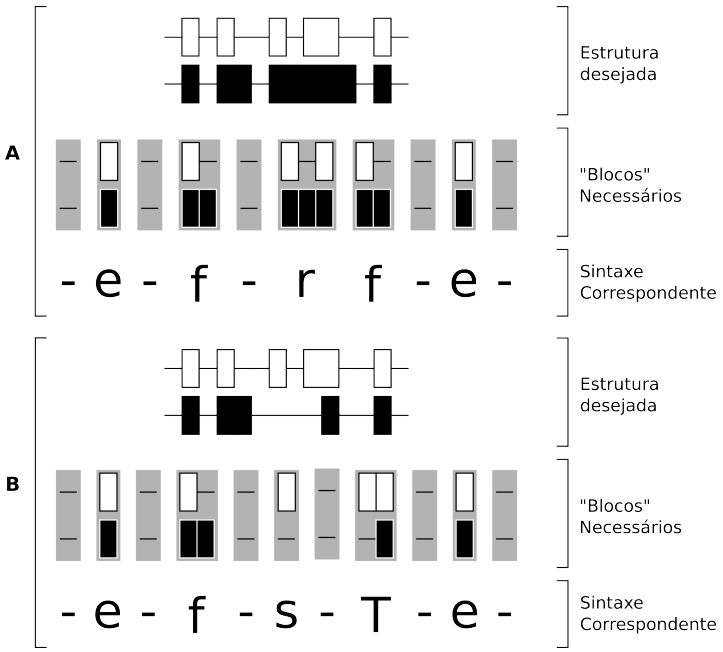

Figura 3.5: *Construção de sintaxes complexas a partir de “blocos” (apresentados anteriormente na figura 3.3).*

Eventos como ES ou retenção de íntron (IR) podem ser facilmente identificados e relacionados com uma sequência variante. Por exemplo, ES ocorre em variantes que não apresentam um éxon, e IR ocorre em variantes que apresentam um íntron retido. Porém, eventos A3’/5’SS não são tão explícitos, uma vez que a única diferença entre os variantes é a utilização de um sítio de *splicing* que altera o tamanho do éxon. Por motivos de conveniência, eventos A3’/5’SS são usualmete relacionados com estrutura que apresenta a maior sequência (por exemplo, o evento “-s-t-” mostra que o variante que possui o ES também possui um A3’SS com a maior porção de sequência). O mesmo problema ocorre com eventos DSS. Nesse caso, foi definido de forma arbitrária que eles ocorrem nos variantes que apresentam o primeiro éxon.

A sintaxe proposta pode facilmente representar padrões de *splicing* complexos. Devido à natureza lógica das sequências expressas, alguns padrões não são permitidos. Por exemplo, o padrão “-ft-” é inválido uma vez que deve haver um íntron entre os eventos “f” e “t”. O carácter “t” sempre deve estar após um íntron (“-t”), e o carácter “f” tem que estar sempre antes de um íntron (“f-”). Existem diversas outra exceções para a representação de CASEs, mas elas podem ser facilmente reconhecidas por meio da análise das estruturas das sequências relacionadas. Por esse motivo, é importante ter cautela em relação à estrutura das sequências para representar ASEs e CASEs por meio da sintaxe.

Evento de Splicing Alternativo	*Genes Afetados*	*Eventos Totais*	*Eventos por Gene*
Éxon skipping	10125 (51,77%)	38060	1,95
Sítio 3' Alternativo	7490 (38,30%)	30172	1,54
Sítio 5' Alternativo	7258 (37,11%)	27585	1,41
Retenção de íntron	6565 (33,57%)	12632	0,65
Sítio duplamente específico	53 (0,27%)	112	0,0057

Tabela 3.2: *Frequêcia dos principais tipos de ASEs.*

3.6 Implementação

Para disponibilizar a implementação do método aqui descrito para a comunidade, uma ferramenta *web* denominada SPLOOCE[7] foi desenvolvida. Para auxiliar os usuários, o SPLOOCE fornece uma caixa de pesquisa com uma tabela de referência explicando a sintaxe e mostrando alguns exemplos. Além de pesquisar pela sintaxe, o SPLOOCE pode também mostrar todos os eventos conhecidos para um gene específico, bastando digitar o nome do gene entre aspas duplas na caixa de pesquisa (figura 3.6A).

O SPLOOCE também fornece opções avançadas para a pesquisa. São fornecidos filtros para cromossomo, fita, nome de gene e tipo de sequência. Os usuários podem também avaliar a especificidade de ASEs e CASEs em relação ao tecido e patologia. Um score para a especificidade é fornecido. Ela corresponde à análise da distribuição χ^2 feita entre as sequências expressas que suportam os variantes correspondentes. A análise é baseada na anotação fornecida pelo eVOC (Kelso *et al.*, 2003) para ESTs e curação manual para sequências RNASeq.

Os resultados fornecidos pelo SPLOOCE podem ser gravados em um arquivo de formato GFF, o qual pode, por exemplo, ser usado como *track* no UCSC *Genome Browser*[8]. Por padrão, os resultados são apresentados no *browser* em uma tabela contendo cromossomo, posições genômicas, nome do gene e uma visualização pictórica do respectivo ASE ou CASE, seguido da respectiva quantidade de sequências de suporte (figura 3.6B). O SPLOOCE também fornece um *link* para o UCSC *Genome Browser* (com *tracks*), e um *link* local para informações adicionais.

Na seção "*Details*", quando uma sequência de referência (RefSeq) está envolvida em um evento, ela é usada como molde para criar uma nova sequência de mRNA contendo o evento alternativo. Para cada uma dessas sequências alternativas, o SPLOOCE prediz o respectivo ORF (*Open Reading Frame*), que é então traduzido para uma proteína. Além do mais, focando inferir alguma importância biológica, o SPLOOCE análisa os domínios das proteínas utilizando o HMMER 3.0 e dados do PFAM (Finn *et al.*, 2010) (figura 3.6C).

3.6.1 Análise

Para ilustrar o uso do SPLOOCE, alguma questões básicas sobre a frequência e tipos de eventos de *splicing* alternativo foram analisadas. A tabela 3.2 mostra a frequência de todos os tipos de ASEs presentes no banco de dados. Como esperado, ES é o tipo de splicing alternativo mais frequente. Um aspecto interessante do método é que ele permite explorar a distribuição de eventos de sítios duplamente específicos. Esse tipo de evento foi encontrado em 53 (0,27%) dos genes humanos codificantes. Em uma análise menos restritiva, sem considerar o número de sequências de suporte, o número de genes mostrando esse tipo de evento subiu para 577. Os eventos DSSs são encontrados frequentemente em genes como *DIABLO*, *IRF3* e *MAG*, e eles podem ocorrer em conjunto com outros eventos (tabela 3.3).

Outro aspecto interessante é que o método permite avaliar a combinação de eventos que ocorrem em uma mesma molécula de mRNA. Cada sequência "Comparativa" criada através do nosso *pipeline* contém em média 1,6 eventos, e 46.87% e 12.21% dessas sequências reportam mais que um ou dois eventos, respectivamente. A frequência da combinação entre alguns eventos pode ser observada na

[7] http://www.bioinformatics-brazil.org/splooce
[8] http://genome.ucsc.edu/

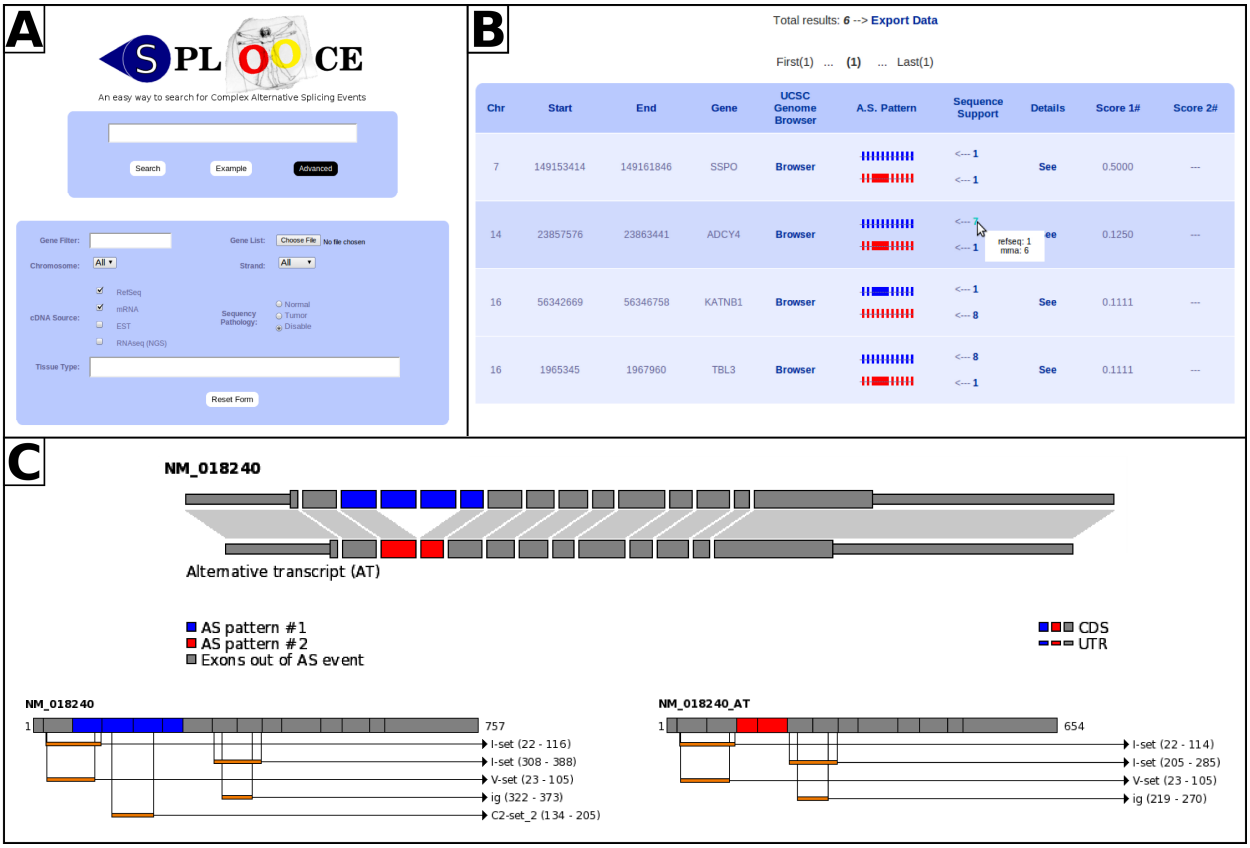

Figura 3.6: *A: Caixa de busca mostrando a aba de parâmetros avançados; B: Exemplo da tabela de resultados; C: Alguns resultados presentes na seção "Details". Nesse exemplo, um éxon* skipping *duplo (-e-s-s-e-) que afeta um domínio proteíco é mostrado.*

tabela 3.4.

Por fim, para melhor entender o que influencia a frequência de ASEs e CASEs, alguns padrões foram explorados. Por exemplo, a combinação de dois eventos de ES adjacentes é significativamente mais frequente quando comparado com outros CASEs (tabela 3.5). Esse excesso é ausente em situações onde ambos os eventos não são adjacentes, como em -s-E-s- e -s-E-E-s-. É possível observar que esses eventos, quando adjacentes, possuem tendência de manter a fase do ORF, por exemplo, 60,78% dos eventos -E-E-s-s-E-E- a mantém. Isso é significativamente maior do que o esperado por chance ($p < 0{,}01$) baseado na manutenção do ORF do primeiro e segundo evento ES (48,60% e 48,49%, respectivamente). Claramente, isso sugere que os eventos de ES adjacentes estão sobre uma forte seleção para manter a fase do ORF (figura 3.7).

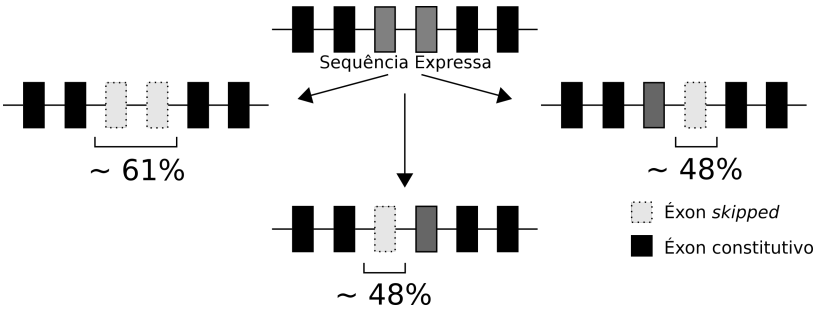

Figura 3.7: *Probabilidade de manter o ORF para o* skipping *de dois éxons adjacentes e para cada um dos respectivos éxons, individualmente.*

Sintaxe	*Frequência (genes)*	*Padrão (simplificado)*
d	577 (2.95%)	23 ou 32
-d-	181 (0.93%)	0230 ou 0320
-d-s-	85 (0.43%)	023030 ou 032020
-Ed-	57 (0.29%)	01230 ou 01320
-d-T	57 (0.29%)	023031 ou 032021
-dE-	56 (0.28%)	02310 ou 03210
f-d-T	26 (0.13%)	12023031 ou 13032021
E-d-T	20 (0.10%)	1023031 ou 1032021
-EdE-	17 (0.087%)	012310 ou 013210
-d-t	11 (0.06%)	023021 ou 032031
-d-S-	8 (0.04%)	023020 ou 032030
-df-	5 (0,025%)	023120 ou 032130
-tD-	3 (0.015%)	021320 ou 031230
f-D-T	2 (0.01%)	12032021 ou 13023031

Tabela 3.3: *Frequência de eventos DSS combinados com outros ASEs.*

	ES	*IR*	*DSS*	*ASS3'*	*ASS5'*
ES	—	6636	257	10992	10884
IR	6636	—	78	6386	6163
DSS	257	78	—	211	217
ASS3'	10992	6386	211	—	11879
ASS5'	10884	6163	217	11879	—

Tabela 3.4: *Frequência de genes apresentando combinações de eventos de* splicing *alternativo.*

Sintaxe	*Frequência (genes)*	*Sintaxe*	*Frequência (genes)*	*Sintaxe*	*Frequência (genes)*
-s-	14461 (73.04%)	-s-F-	3258 (16.66%)	-s-S-s-	597 (3.05%)
-f-	11220 (57.37%)	-s-f-	3052 (15.60%)	-t-E-t-	591 (3.02%)
-s-s-	10020 (51.23%)	-s-E-S-	2718 (13.90%)	-f-E-f-	549 (2.81%)
-s-s-s-	6844 (34.99%)	-s-S-S-	2627 (13.43%)	-f-E-F-	541 (2.77%)
-f-S-	5733 (29.31%)	-t-S-	2563 (13.10%)	-rR-	430 (2.20%)
-s-S-	5532 (28.28%)	-s-s-S-	2168 (11.08%)	-r-r-	272 (1.39%)
-s-T-	5302 (27.11%)	-s-E-E-s-	1627 (8.32%)	-r-R-	228 (1.17%)
-r-	4687 (23.96%)	-s-E-E-S-	1580 (8.08%)	-f-E-t-	226 (1.16%)
-s-t-	3934 (20.11%)	-t-T-	1150 (5.88%)	-rrr-	224 (1.15%)
-f-s-	3710 (18.97%)	-f-F-	1033 (5.28%)	-f-E-T-	218 (1.11%)
-f-T-	2866 (14.65%)	-t-t-	1022 (5.23%)	-s-E-S-E-s-	73 (0.37%)
-f-t-	2800 (14.32%)	-f-f-	1021 (5.22%)	-rRR-	42 (0.21%)
-s-E-s-	2768 (14.15%)	-rr-	928 (4.74%)	-rrR-	37 (0.19%)
-E-r-E-	2222 (11.36%)	-t-E-T-	629 (3.22%)	-rRr-	27 (0.14%)

Tabela 3.5: *Número de genes afetados por diferentes tipos de ASEs e CASEs.*

3.6.2 *Behind the Scenes*

O SPLOOCE é composto por um banco de dados relacional (MySQL) e alguns scripts escritos em Perl, os quais somam aproximadamente 10.000 linhas de código. Essa ferramenta é capaz de explorar todo o transcriptoma humano em busca de qualquer sintaxe em apenas poucos segundos. Essa performance foi obtida por meio do pré-processamento e armazenamento de informações sobre as sintaxes mais simples e frequentes. As demais sintaxes, porém, podem ser processadas *on-the-fly* (em tempo real) utilizando matrizes e diversas outras informações armazenadas em um banco de dados relacional (figura 3.8 e script 3.1). Essas características tornam o SPLOOCE uma ferramenta híbrida de alta eficiência.

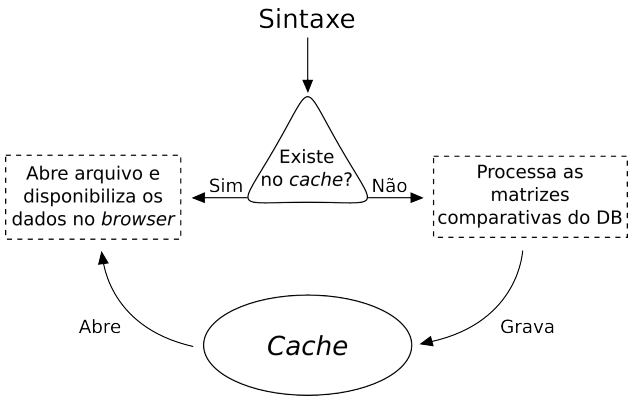

Figura 3.8: *Estratégia híbrida implementada pelo SPLOOCE: dados novos podem ser processados* on-the-fly.

Listing 3.1: Main (define a sintaxe e chama as funções de análise)

```
#!/usr/bin/perl -w
use strict;

#MAIN
my $sintaxe = "-s-e-e-S-";                            #SINTAXE (INPUT)
my @data_files = &exits_data_for_sintaxe($sintaxe);  #CARREGA CACHE

#VERIFICA SE O RESULTADO JA EXISTE NO "CACHE"
if  (! @data_files)
{
    #NAO EXISTE NO CACHE, ENTAO PROCESSA OS DADOS
    my $pattern   = &sintax_to_regex($sintaxe);      #TRANSFORMA SINTAXE PARA REGEX
    my @data_files = &regex_analysis($pattern);      #ANALISA A REGEX E GRAVA CACHE
}

&print_results(@data_files);                         #IMPRIME RESULTADOS (CACHE)

#FIM DO SCRIPT
end;
```

3.6.2.1 Banco de Dados

O banco de dados do SPLOOCE (figura 3.9) possui, basicamente, as sequências denominadas como "Posição" e "Comparativa", além de diversos dados adicionais (sequências de suporte, anotações tecido/patologia, sítios de *splicing* etc). Bancos de dados relacionais são normalmente utilizados para remover a redundância dos dados e otimizar as pesquisas. No SPLOOCE, porém, um simples banco de dados relacional não é rápido o suficiente devido ao grande número de dados nele armazenados. Um dos maiores problemas é que o SPLOOCE procura por padrões de sequências, e o uso direto de *regexes* em todos os dados do DB é uma estratégia altamente ineficiente. Portanto, algumas estratégias de otimização tiveram que ser implementadas.

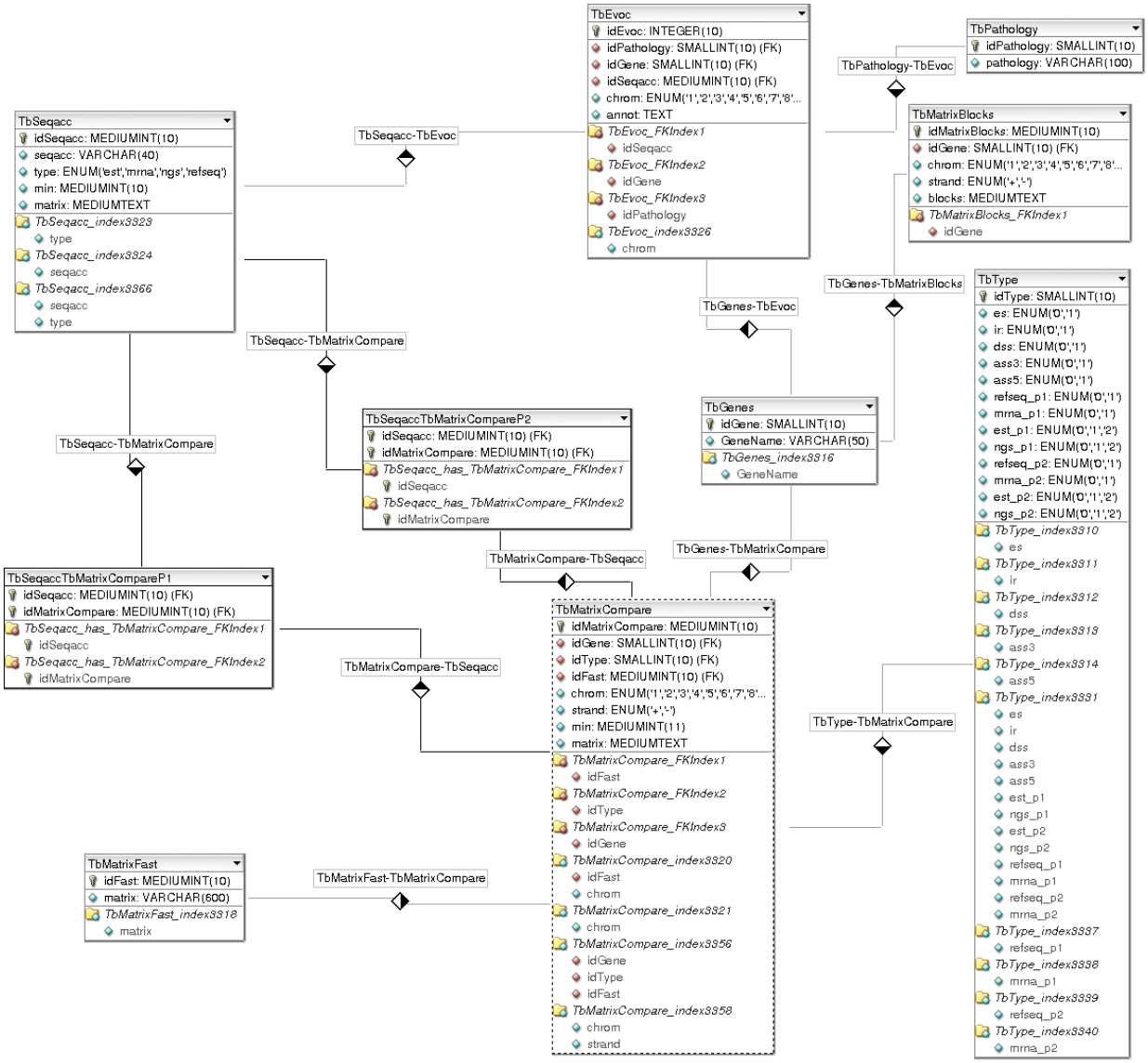

Figura 3.9: *Organização do banco de dados relacional do SPLOOCE, onde diversos tipos de dados são armazenados.*

Uma estratégia foi anotar cada sequência "Comparativa" de forma binária, mostrando quais eventos de *splicing* alternativo estão presentes ou ausentes. Exemplo de sequências e busca:

```
SELECT * FROM Splooce;

-----------------------------------------------------------------------------------------
| Sequencia_Comparativa              | Seq_Reduzida     | 5ASS | 3ASS | ES | IR | DSS |
-----------------------------------------------------------------------------------------
| 111111000001112220000000033311111  | 1012031          | 1    | 1    | 0  | 0  | 0   |
| 111111220001111110000000033311111  | 1201031          | 1    | 1    | 0  | 0  | 0   |
|  1111100000111111000000033311111111 | 101031          | 0    | 1    | 0  | 0  | 0   |
|    11100000333333000000000001111   | 10301            | 0    | 0    | 1  | 0  | 0   |
| 111111000002222223333000000111111 | 102301           | 0    | 0    | 0  | 0  | 1   |
-----------------------------------------------------------------------------------------

//Buscar sequencias que possuem eventos de bordas 5' e 3' Alternativas
Select * FROM Splooce WHERE 5ASS = 1 AND 3ASS = 1;

-----------------------------------------------------------------------------------------
| Sequencia_Comparativa              | Seq_Reduzida     | 5ASS | 3ASS | ES | IR | DSS |
-----------------------------------------------------------------------------------------
| 111111000001112220000000033311111  | 1012031          | 1    | 1    | 0  | 0  | 0   |
| 111111220001111110000000033311111  | 1201031          | 1    | 1    | 0  | 0  | 0   |
-----------------------------------------------------------------------------------------
```

A partir de uma sintaxe é possível determinar quais eventos a mesma possui e, consequentemente, encontrar sequências "Comparativas" com uma maior probabilidade de apresentarem o padrão de *splicing* desejado. Para finalizar a busca, as sequências "Comparativas" são comparadas com padrão de eventos definido pela sintaxe. Essa etapa pode ser feita por meio de *regexes*, porém a comparação simples de *strings* é muito mais eficiente. Uma estratégia de otimização foi, portanto, reduzir a complexidade da sequência "Comparativa" removendo a redundância dos caracteres. A busca pelo padrão de *splicing* alternativo pôde então ser feita utilizando o comando "LIKE", assim como exemplificado a seguir:

```
Select * FROM Splooce WHERE (5ASS = 1 and 3ASS = 1)
                        AND (Sequencia_Reduzida LIKE '%12031%'
                        OR   Sequencia_Reduzida LIKE '%13021%');

-------------------------------------------------------------------------------------
| Sequencia_Comparativa             | Seq_Reduzida    | 5ASS | 3ASS | ES | IR | DSS |
-------------------------------------------------------------------------------------
| 1111110000011122200000033311111  | 1012031         | 1    | 1    | 0  | 0  | 0   |
-------------------------------------------------------------------------------------
```

Resumidamente, quanto maior a complexidade de combinações de ASEs uma sintaxe apresentar, mais rápida será a busca. Uma busca não otimizada, por outro lado, deve apresentar o mesmo tempo de busca para qualquer tipo de sintaxe. Por esse motivo, os eventos mais simples e frequentes foram identificados e processados, e os dados resultantes foram armazenados no *cache*. O *cache* nada mais é do que uma árvore de arquivos e pastas onde os dados referentes à cada sintaxe já processada pelo SPLOOCE é armazenada (incluindo sintaxes novas submetidas por quaisquer usuários).

3.6.2.2 Processando Sequências "Comparativas"

Após o SPLOOCE pesquisar por uma sintaxe que não está presente no *cache* e, o banco de dados retornar diversas sequências "Comparativas", o primeiro passo é, portanto, converter a sintaxe para uma expressão regular (função chamada no script 3.1, linha 12). Nessa etapa, os caracteres da sintaxe são convertidos para *regex* utilizando uma sequência de *regexes* de substituição. Por exemplo, "-S-" é convertido para "0(3+)0", e assim por diante. No final das substituições existem algumas *regexes* responsáveis por corrigir quaisquer erros e redundâncias encontradas. O segundo passo, consequentemente, é processar as sequências "Comparativas" utilizando a *regex* referente à sintaxe (script 3.2).

Listing 3.2: Função regex-analysis (Analisa as seqs "Comparativas")

```
1  sub regex_analysis
   {
3      my $pattern = shift;     #REGEX (INPUT)
       my @res_files = ();      #NOMES DOS ARQUIVOS CACHE
5
       #OBTEM DADOS DO BANCO DE DADOS
7      while (my ($gene, $cmp_sequence) = &got_data_mysql($pattern))
       {
9          #PROCURA POR EVENTOS AO LONGO DA SEQ COMPARATIVA
           while ($cmp_sequence =~ /($pattern)/icg)
11         {
               my $pos_ini = $-[1];           #RETORNA A POSICAO...
13             my $pos_fim = $+[1];           #INICIAL E FINAL DO EVENTO

15             #OBTEM DETALHES DO EVENTO E REGISTRA OS DADOS
               push @res_files, &process_event($gene,$pos_ini,$pos_fim);
17         }
       }
19     return @res_files;       #RETORNA O NOME DOS ARQUIVOS (CACHE)
   }
```

Resumidamente, com base no script, as sequências "Comparativas" são diretamente carregadas do banco de dados por meio da função "got_data_mysql" e então processadas pela *regex*. A posição do evento na sequência binária é obtida e transferida para a função "process_event" (função chamada no script 3.2, linha 18). Nessa etapa, em especial, a sequência "Posição" é utilizada para obter a posição genômica do evento identificado (script 3.3, a partir da linha 9).

Listing 3.3: Função process-event (Retorna informações sobre o evento)

```
sub process_event
{
    #INPUT: GENE E POSICOES DO EVENTO NA SEQUENCIA BINARIA
    my $gene = shift;
    my $pos_ini = shift;
    my $pos_fim = shift;

    #CARREGA MATRIX DE CONVERSAO PARA POSICOES GENOMICAS
    my @gen_pos_matrix = split /,/, &got_matrixpos_mysql($gene);

    #RETORNA POSICOES GENOMICAS
    my $gen_start = $gen_pos_matrix[$pos_ini];
    my $gen_end   = $gen_pos_matrix[$pos_fim];

    #ANALISA DETALHES DIVERSOS
    my @support_seqs = &search_supp_seqs($gene, $gen_start, $gen_end);
    my %annotation   = &load_annotation(@support_seqs);

    #OUTRAS FUNCOES
    .
    .
    .

    #ARMAZENA DADOS
    my $file = &save_data($gene, $gen_start, $gen_end, ...);

    #RETORNA NOME DO ARQUIVO DE DADOS PARA USO POSTERIOR
    return $file;
}
```

Por fim, diferentes funções são chamadas para se obter uma completa lista de informações acerca do evento. Os dados são gravados no *cache* e o nome do respectivo arquivo é retornado para a função "regex_analysis". No final de tudo, todos os nomes de arquivos armazenados são abertos, interpretados e apresentados no *web browser*.

3.7 Discussão

Embora outras metodologias e interfaces já tenham sido anteriormente propostas (Malko *et al.* , 2006; Nagasaki *et al.*, 2006; Sammeth *et al.*, 2008) para o estudo do *splicing* alternativo, todas apresentam limitações, assim como discutido anteriormente (Sammeth *et al.*, 2008). O SPLOOCE, descrito aqui, é uma alternativa eficiente e complementar para a análise de eventos de *splicing* alternativo devido a sua flexibilidade na análise de padrões e variedade de aplicações. Assim como o ASTALAVISTA (Sammeth *et al.*, 2008), o método utilizado pelo SPLOOCE é baseado na comparação de todos os transcritos de um determinado *locus*. O SPLOOCE, porém, utiliza um sistema de anotação baseado em expressões regulares que permite o uso de uma simples e eficiente sintaxe para os eventos de *splicing*. A sintaxe é composta por caracteres intuitivos e é capaz de representar qualquer padrão de CASE, incluindo os que apresentam os raros eventos de sítios duplamente específicos. Por fim, a sintaxe foi implementada com sucesso e pode se tornar um formato padrão para representar CASEs de diferentes complexidades.

Capítulo 4

Retenção de Íntrons

4.1 Introdução

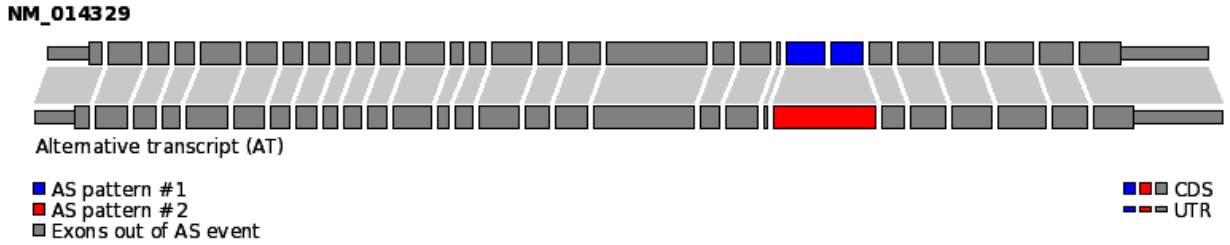

Figura 4.1: *Exemplo "ideal" de um evento de retenção de íntron (SID: 16_E904006409367).*

O *splicing* alternativo (AS) é um dos mecanismos moleculares mais importantes em células eucarióticas (Galante *et al.*, 2004; Modrek e Lee, 2002; Wang *et al.*, 2008), e é encontrado em transcritos da ampla maioria dos genes humanos que apresentam mais de um éxon (Pan *et al.*, 2008; Wang *et al.*, 2008). Muitos eventos de *splicing* alternativo (ASEs) estão envolvidos na etiologia de doenças, incluindo o câncer (Kirschbaum-Slager *et al.*, 2005; Venables, 2004; Wang e Cooper , 2007), isquemia (Daoud *et al.*, 2002) e outras desordens humanas comuns (Garcia-Blanco *et al.* , 2004).

Atualmente, cinco tipos diferentes de ASEs são conhecidos (Zhang *et al.*, 2007). Retenção de íntron (IR) é o tipo de evento mais controverso, uma vez que ele pode ser o resultado de pré-mRNAs não ou parcialmente processados (Galante *et al.*, 2004). Assim como todos tipos de ASEs, IRs podem ser identificados por meio de comparações entre sequências expressas, as quais estão exponencialmente crescendo em termos de disponibilidade. Por exemplo, em 2004 existiam aproximadamente 7,7 milhões de sequências expressas humanas disponíveis no GenBank e, atualmente, existem mais de 20 milhões (Benson *et al.*, 2012). Esses números não incluem as sequências obtidas a partir de sequenciamentos de próxima geração (NGS), os quais estão sendo exaustivamente explorados e podem ser uma excelente fonte para a descoberta de novos ASEs (Kroll *et al.*, 2012; Roberts *et al.*, 2011; Trapnell *et al.*, 2009; Wheeler *et al.*, 2007).

Em 2004, nosso grupo publicou a primeira análise em larga escala de IR em humanos (Galante *et al.*, 2004). Foi argumentado que uma fração considerável dos eventos de IR poderia ter significado biológico porque a distribuição desses eventos em relação à diferentes parâmetros apresentou um padrão não-aleatório. O aumento da disponibilidade de sequências expressas junto com avanços conceituais na área nos permitiram atualizar esses dados.

4.2 Objetivos

O objetivo geral desse capítulo é identificar e avaliar o potencial funcional dos eventos de retenção de íntrons observados em sequências expressas humanas. Para isso, os seguintes objetivos secundários foram definidos:

- Identificar eventos de retenção de íntrons e analisá-los a partir de aspectos gerais como distribuição no transcrito, conteúdo de CG e domínios. Parte dos resultados serão comparados com os dados obtidos por Galante *et al.* (2004), entre outros;
- Construir computacionalmente proteínas afetadas por eventos de retenção de íntron e validá-las utilizando espectrometria de massas em *tandem*. Essa análise pode mostrar o quão importante as retenções de íntrons são do ponto de vista proteômico;
- Descrever e classificar as alterações proteômicas causadas pelos eventos de retenção de íntrons, no intuito de verificar quais são os padrões mais frequentes e biologicamente viáveis.
- Avaliar o papel dos alvos de miRNA para os eventos de retenção de íntrons não-codificantes, observando a frequência e tipos de alvos encontrados em éxons, íntrons e íntrons retidos;

4.2.1 Justificativa

Os ASEs são um dos componentes significativos que estão envolvidos com a complexidade de organismos multicelulares (Modrek e Lee, 2002; Wang *et al.*, 2008). Assim, é importante predizer se os eventos de IR, que podem advir do processamento parcial do pré-mRNA, também possuem significado biológico. Dessa forma, a validação desses eventos permitirá, por exemplo, o estudo direcionado de proteínas que são realmente codificadas e possuem, provavelmente, algum papel biológico em alguma patologia e/ou em processos celulares de importância terapêutica. Esse trabalho, portanto, pode representar um avanço no estudo desses eventos até então subestimados.

4.3 Materiais & Métodos

Dados públicos

O genoma humano de refeferência (NCBI36/hg18) foi gravado do portal da UCSC *Genome Bioinformatics*[1]. As sequências RefSeq foram gravadas do banco de dados de Sequências de Referência[2]. Sequências do tipo mRNA foram gravadas do portal da UCSC *Genome Bioinformatics* (*Homo sapiens*), e as sequências do tipo ESTs foram gravadas do dbEST[3]. Dados de RNAseq foram gravados do portal SRA[4]; Foram utilizados os IDs SRX003935, SRX003934, SRX003933, SRX003932, SRX003931, SRX003930, SRX003929, SRX003928, SRX003927 e SRX003926. Esses experimentos de NGS totalizam 63 corridas de sequênciamento (Illumina Genome Analyzer) e 211.006.871 *reads* (36 pb cada).

Alinhamento das sequências

Todas sequências dos tipos RefSeq, mRNA e EST foram alinhadas contra o genoma humano usando um protocolo já descrito anteriormente (Galante *et al.*, 2007). Resumidamente, todas as sequências foram mapeadas contra o genoma humano usando o alinhador BLAT (Kent , 2002b). Em seguida, os transcritos apresentando identidade maior que 95% e com uma cobertura maior que 90% foram remapeadas utilizando o SIM4 (Florea *et al.*, 1998). Os dados de RNASeq, por outro lado, foram mapeados contra o genoma humano usando o pipeline Tophat (Trapnell *et al.*, 2009), e todos os *reads* devidamente mapeados foram submetidos

[1] http://genome.ucsc.edu
[2] http://www.ncbi.nlm.nih.gov/RefSeq/
[3] http://www.ncbi.nlm.nih.gov/dbEST/
[4] http://www.ncbi.nlm.nih.gov/sra

para o programa Cufflinks (Trapnell *et al.*, 2010) (os transcritos preditos foram denominados de NGS). Parâmetros padrão foram usados em ambos os programas.

Clusterização das sequências

Todas as sequências mapeadas que compartilharam a mesma região genômica foram agrupadas utilizando uma estratégia orientada ao gene. Primeriamente, todas as sequências RefSeq foram anotadas conforme os nomes oficiais dos genes correspondentes, assim como presente no NCBI-Gene[5]. Em seguida, foram agrupados no mesmo gene todos os transcritos não-RefSeq (mRNAs, ESTs e NGSs) que apresentaram múltiplos éxons e que compartilharam uma ou mais junções de *splicing* com pelo menos uma sequência RefSeq. Por fim, foram agrupados os transcritos não-RefSeq que apresentaram apenas um éxon e uma sobreposição maior que 30 nucleotídeos com um transcrito RefSeq. As informações da clusterização foram armazenadas em um banco de dados relacional.

Dados de espectrometria

Os dados brutos foram obtidos de Geiger *et al.* (2012) e estão disponíveis no *Tranche Network*[6]. Os dados de MS/MS das linhagens celulares A549, GAMG, HEK293, HeLa, HepG2, K562, MCF7, RKO, U2=S, LnCap e Jurkat foram utilizados. Resumidamente, 100 μg de extrato de proteínas de cada linhagem celular foram processados pela enzima tripsina e fracionados pelo método FASP (Wis niewski *et al.*, 2009). Os peptídeos foram então separados por cromatografia de fase reversa utilizando colunas de 20 cm (diâmetro interno de 75 μm) acopladas diretamente ao espectrômetro LTQ-Orbitrap Velos (*Thermo Scientific*).

Análise dos dados de espectrometria de massa

As análises foram feitas utilizando o software MaxQuant (Cox e Mann, 2008) versão 1.3.0.5. O espectro de MS/MS foi comparado ao nosso banco de dados de proteínas (ver seção 4.5). Proteínas do banco de dados IPI-Human versão 3.81 (Kersey *et al.*, 2004) também foram adicionadas às análises com o objetivo de identificar proteínas desconhecidas. As análises do MaxQuant incluiram uma tolerância de massa de 20 ppm, a qual foi utilizada para a recalibração da massa. Na busca principal, as massas precursoras e os fragmentos de massa foram pesquisados com uma tolerância de massa de 6 ppm e 0.5 Da, respectivamente. A busca incluiu modificações de oxidação (Met), acetilação do N-terminal (proteína), e Pyro-Glu (Q e E). Carbamidometil cisteína foi tratada como uma modificação fixa. O tamanho mínimo para o peptídeo foi de 7 amino ácidos e o número máximo de *miscleavages* (clivagens errôneas) permitido foi de 2. O *false discovery rate* (FDR, taxa de falsas descobertas) foi configurado para 0.01 para a identificação de peptídeos e proteínas. Peptídeos compartilhados entre duas proteínas foram combinados e reportados como um grupo de proteínas. A tabela de proteínas foi filtrada para eliminar a identificação de sequências do banco de dados reverso e contaminantes comuns.

[5] http://www.ncbi.nlm.nih.gov/gene/
[6] http://proteomecommons.org

4.4 Resultados

4.4.1 Identificação dos Eventos de Retenção de Íntron

O banco de dados de sequências expressas humanas (ver Materiais & Métodos 4.3) foi processado por meio de um *pipeline* que inclui etapas de mapeamento e clusterização (Galante *et al.*, 2004; Sakabe *et al.*, 2003). Os dados do *Sequence Read Archive* (SRA) (Wheeler *et al.*, 2007), o qual contém dados brutos de plataformas de *next-generation sequencing*, foram primeiramente mapeados utilizando o *pipeline* Tophat (Trapnell *et al.*, 2009). O transcriptoma foi então construído utilizando o algoritmo Cufflinks (Roberts *et al.*, 2011).

O banco de dados inicial de sequências, devidamente mapeadas (ver Materiais & Métodos 4.3), apresentou 30.678 RefSeqs, 258.444 mRNAs, 6.987.423 ESTs e 9.565.439 transcritos criados pelo Cufflinks. Para aprimorar a qualidade do banco de dados, apenas sequências que apresentaram pelo menos dois éxons foram selecionadas. Esse processo permite excluir grande parte das sequências que não sofreram *splicing* e que podem apresentar, portanto, "falsas" retenções de íntrons. Como resultado, apenas 95,7% dos RefSeqs, 82,6% dos mRNAs, 56,4% dos ESTs e 4,9% das sequências do Cufflinks foram selecionadas para as análises posteriores.

Os IRs foram analisados utilizando um método baseado em sequências binárias (Kroll *et al.* , 2012). Resumidamente, as sequências expressas de cada gene foram primeiramente convertidas para sequências binárias. Essas sequências binárias resultantes foram sobrepostas e utilizadas para criar um catálogo de éxons e íntrons, denominado de "catálogo de referência" (figura 4.2), o qual foi, finalmente, comparado par-a-par com cada uma das sequências expressas para a identificação dos eventos de IR.

Foi criado um catálogo de referência baseado em cDNAs (*full-insert*: RefSeqs e mRNAs), e ele apresentou um total de 210.698 íntrons (72,2%, 6,8% e 3,3% no CDS, 5' UTR e 3' UTR, respectivamente). Outro catálogo de referência foi criado utilizando sequências ESTs e um total de 303.764 íntrons foram identificados (69%, 8,3% e 5,3% no CDS, 5' UTR e 3' UTR, respectivamente). Uma pequena porcentagem dos eventos foi mapeada em regiões não completamente anotadas ou mistas (uma parte se encontra na CDS e a outra na UTR), conforme o banco de dados do UCSC *Genome Bioinformatics*.

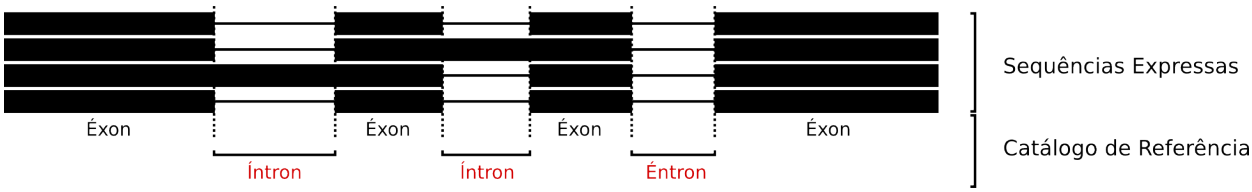

Figura 4.2: *Algoritimo de construção do "catálogo de referência"*

4.4.2 Análise Geral dos Eventos de Retenção de Íntron

No geral, 9.037 de 19.845 (45,54%) genes humanos apresentaram no mínimo um evento de IR (tabela 4.1). O número total de genes afetados triplicou após 8 anos (Galante *et al.*, 2004), assim como esperado devido ao aumento de sequências públicas disponíveis, como ESTs e NGSs. Os dados gerados a partir da comparação entre as sequências *full-insert* foram considerados como de alta confiança (*FxF*), enquanto os dados gerados a partir da comparação entre as sequências *full-insert* com todos os diferentes tipos de sequências expressas foram considerados como sendo de baixa confiança (*FxA*).

Múltiplos eventos de IR que ocorrem em um mesmo gene foram observados. A maioria desses eventos constuma ocorrer consecutivamente, retendo um grupo de íntrons próximos. No banco de dados de baixa confiança, um total de 813 (4,10%) genes mostraram ao menos um evento de IR múltiplo. Esses eventos possuem de 2 até 5 íntrons retidos. Alguns tipos de sequências, como ESTs e NGSs, apresentam dificuldade em detectar eventos abrangendo mais de 3 íntrons devido

	Full-insert cDNA	*EST*	*NGS*	*EST + NGS*	*Total*
Full-insert cDNA	5.327	3.948 (2.672)	2.713 (1.994)	4.822 (3.675)	7.000 (6.114)
EST	7.372	—	—	—	7.372
Total	8.476	3.948 (2.672)	2.713 (1.994)	4.822 (3.675)	9.679 (9.037)

Tabela 4.1: *Número de genes afetados por pelo menos um evento de IR. *() apenas eventos suportados por mais de uma sequência expressa (EST/NGS).*

Comparação	*Região*	*Observado*	*Esperado*	*p-Value*
FxF	CDS	5.158	5.976	$1,831 \times 10^{-26}$
	5' UTR	662	625	$0,090$
	3' UTR	1.164	307	$0,000$
FxA	CDS	5.765	6.478	$4,099 \times 10^{-19}$
	5' UTR	718	677	$0,073$
	3' UTR	1.170	333	$0,00$

Tabela 4.2: *Número observado e esperado de IRs para as diferentes regiões dos transcritos.*

ao tamanho restrito de suas respectivas sequências. Por motivos de confiança para as próximas análises, os eventos não suportados por RefSeqs ou mRNAs foram filtrados, e apenas casos com ao menos 2 ou mais EST/NGSs de suporte foram selecionados.

Em um artigo de 2004, publicado por nosso grupo (Galante *et al.*, 2004), foi encontrada uma tendência de IRs ocorrerem com uma maior frequência em UTRs, especialmente na 3' UTR, sugerindo a existência de uma pressão seletiva contra eventos de IR que afetam a região CDS. Essa análise foi refeita para os nossos dados. O número esperado de retenções de íntrons para o 5' UTR, CDS e 3' UTR foi calculado utilizando a equação $\frac{IR}{Ii}/I$, onde IR é o número total de eventos de retenção, Ii é o número de íntrons na respectiva região do transcrito, e I é o número total de íntrons presentes no banco de dados. A distribuição dos eventos de IR foi confirmada. Existe uma clara escassez de IRs na região CDS e um excesso na região 3' UTR (tabela 4.2).

Uma vez que os IRs observados para a CDS se encontram sob uma forte pressão seletiva, acredita-se, portanto, que eles devem compartilhar algumas caracerísticas com éxons codificadores. Para testar essa hipótese, foi analisado o conteúdo de CG dos éxons, IRs e seus respectivos éxons adjacentes presentes na região CDS. Éxons possuem um maior conteúdo de CG, e os IRs de importância biológica devem consequentemente apresentar um maior conteúdo de CG comparado com íntrons não retidos, assim como apresentar um conteúdo de CG similar ao dos éxons adjancentes (Galante *et al.*, 2004). Para uma melhor comparação, os dados foram classificados por comprimento, uma vez que diferentes tamanhos de íntrons e éxons podem apresentar diferentes valores de CG (Galante *et al.*, 2004; Oliver e Marín, 1996). O conteúdo de CG dos IRs mostrou ser estatísticamente diferente dos íntrons não retidos (FxF: $p = 3,50 \times 10^{-4}, d = 6, t = 7,25$; FxA: $p = 4,25 \times 10^{-3}, d = 6, t = 4,47$), mas estatíscamente não diferente dos éxons adjacentes (FxF: $p = 3,01 \times 10^{-1}, d = 6, t = -1,13$; FxA: $p = 3,01 \times 10^{-1}, d = 6, t = -1,13$), assim como esperado (Galante *et al.*, 2004; Lander *et al.*, 2001).

Foi também observado um excesso de eventos na 3' UTR comparado com a região 5' UTR. Diferenças entre essas duas regiões podem ocorrer porque or IRs em 5' UTR tem a possibilidade de introduzir *start* códons alternativos, afetando, assim, a proteína. Essa possibilidade foi testada verificando a presença de potenciais *start* códons alternativos em IRs encontrados na 5' UTR. De todos os eventos presentes na 5' UTR, 539 (75,07%) apresentaram pelo menos um *start* códon em potencial. Esse número é significativamente menor do que o esperado ao acaso quando comparado com os íntrons não retidos encontrados nessa mesma região ($p = 1,14 \times 10^{-21}, d = 1, \chi^2 = 90,11$). Os eventos encontrados na 5' UTR que não introduziram *start* códons mostraram um maior conteúdo

CG (71%) em comparação com os eventos que introduziram *start* códons (56%) ($p = 0,00, d = 836, t = 17,13$). Essas diferenças de CG estão em grande parte relacionadas com o tamanho do íntron, o que pode explicar a baixa frequência de novos start códons. Possivelmente existe uma força de seleção purificadora que filtra os eventos da 5' UTR que afetam a sequência da proteína, assim como ocorre com os eventos de encontrados na CDS.

Resumidamente, nessa seção, novos eventos de IR foram identificados devido ao aumento de sequências expressas publicamente disponíveis, as quais continuam suportando nossos antigos resultados (Galante *et al.*, 2004). Embora resultados significativamente diferentes não foram observados, o uso de novas metodologias pode ser a chave para novas descobertas.

4.5 Análise do Proteoma

Estudos utilizando espectrometria de massas identificaram proteínas alternativas com relativo sucesso (Chang *et al.*, 2010; Power *et al.*, 2009). Um problema, porém, é que a maioria dos espectros de MS/MS não encontram sequências correspondentes nos diversos bancos de dados de proteínas disponíveis (Johnson *et al.*, 2005). Entre diversos fatores, isso ocorre porque apenas poucos eventos de *splicing* alternativo codificadores são atualmente conhecidos (Chang *et al.*, 2010; Johnson *et al.*, 2005; Power *et al.*, 2009; Roth *et al.*, 2005).

Análises de Espectrometria de Massas em *Tandem* (MS/MS) foram utilizadas para identificar proteínas variantes possivelmente codificadas por eventos de IR. Sabe-se que os ASEs em geral contribuem para a complexidade do proteoma e para a modulação da atividade de diversas proteínas (Galante *et al.*, 2004; Modrek e Lee, 2002; Wang *et al.*, 2008), porém, pouco ainda é conhecido sobre o papel funcional dos eventos de retenção de íntrons. Aqui, foram explorados dois tipos de eventos de IR: os que são observados no transcriptoma e os que não são observado no transcriptoma. Para as análises de MS/MS em especial, foram utilizados os dados e o algoritmo de predição de proteínas do SPLOOCE (Kroll *et al.*, 2012). O SPLOOCE atualmente contém 25.684 RefSeqs, que totalizam 249.724 íntrons.

4.5.1 Validação de Eventos de IR Sem Suporte Transcriptômico

Existem evidências claras de que a expressão de mRNAs não pode ser diretamente corelacionada com a expressão de proteínas (Vogel *et al.*, 2010). Sugere-se, portanto, que alguns eventos de IR podem ser codificados mesmo que evidências transcriptômicas ainda não tenham sido observadas. Para testar essa hipótese, todos os íntrons iguais ou menores que 200 pb foram computacionalmente retidos para cada RefSeq do banco de dados, e novos ORFs foram então preditos. Como resultado, 30.672 íntrons foram retidos e 28.109 proteínas alternativas não redundantes foram criadas. O espectro MS/MS foi comparado contra essas proteínas e adicionalmente contra as proteínas do banco de dados do IPI (ver materiais e métodos 4.3). No total, 22 novos eventos desconhecidos foram identificados no espectro de MS/MS (tabela 4.3, "sem suporte transcriptômico").

4.5.2 Validação de Eventos de IR Com Suporte Transcriptômico

Proteínas variantes modificadas por IRs, suportadas apenas por sequências EST/NGS, também foram submetidas para a análise de MS/MS. No total, 1.886 eventos afetando 1.610 genes foram analisados. As proteínas alternativas foram criadas a partir de RefSeqs, assim como apresentado anteriormente (seção 4.5.1). Resumidamente, múltiplos IRs adjacentes e, IRs encontrados nas regiões CDS e 5' UTR foram utilizados para a criação de RefSeqs variantes. Posteriormente, o maior ORF de cada RefSeq variante foi predito e traduzido para uma sequência de aminoácidos. No total, 6.505 proteínas alternativas não redundantes foram criadas. O espectro MS/MS foi comparado contra essas proteínas e adicionalmente contra as proteínas do banco de dados do IPI (ver materiais e métodos 4.3).

Como resultado, 4.179 (64,24%) proteínas alternativas apresentaram correspondência com pelo menos um peptídeo do espectro. Dessas proteínas, 1.299 (19,97%) compartilharam exatamente os

	Gene	Chr	Fita	Início	Fim
Sem suporte transcriptômico	CGN	CHR1	+	149774784	149774879
	ID3	CHR1	-	23758098	23758204
	NOTCH3	CHR19	-	1515800	15158915
	FAM50A	CHRX	+	153331278	153331424
	RFX1	CHR19	-	13934677	13934805
	TCF3	CHR19	-	1570234	1570314
	CHMP2A	CHR19	-	63755147	63755233
	NXF1	CHR11	-	62316563	62316690
	MYL3	CHR3	-	46874767	46874877
	COL16A1	CHR1	-	31910826	31910911
	LIME1	CHR20	+	61839700	61839769
	NSUN5C	CHR7	-	72061906	72061985
	HDAC6	CHRX	+	48546137	48546221
	SYNJ2	CHR6	+	158405771	158405845
	SLC6A6	CHR3	+	14484469	14484599
	MYBBP1A	CHR17	-	4398102	4398186
	HDGFRP2	CHR19	+	4442680	4442760
	HBE1	CHR11	-	5247483	5247604
	UBAP2	CHR9	-	33912598	33912684
	CTC1	CHR17	-	8072363	8072545
	MEI1	CHR22	+	40510371	40510557
	ACY3	CHR11	-	67169200	67169333
Com suporte transcriptômico	CDV3	CHR3	+	134775659	134775752
	GAPDH	CHR12	+	6516554	6516604
	GPS1	CHR17	+	77603129	77603421
	DDT	CHR22	-	22646497	22646131
	ELMSAN1	CHR14	-	73255762	73255463
	STRA13	CHR17	-	77570459	77570383
	RPL29	CHR3	-	52004482	52004161
	MAT2A	CHR2	+	85621795	85621888
	EDC4	CHR16	+	66473250	66473357
	ASL	CHR7	+	65189809	65190150
	HNRNPD	CHR4	-	83513752	83513682
	CKMT1B	CHR15	+	41673655	41673810
	FTSJ1	CHRX	+	48226127	48226317
	SF1	CHR11	-	64291241	64291120

Tabela 4.3: *Eventos de IR que foram validados no proteoma (posições referentes ao genoma hg18).*

mesmos peptídeos encontrados por pelo menos uma proteína do banco de dados do IPI. Aproximadamente 1.807 (27,78%) proteínas alternativas apresentaram diferentes números de peptídeos e apenas 14 proteínas alternativas codificaram pelo menos um peptídeo evento-específico (tabela 4.3, "com suporte transcriptômico"). Essas últimas proteínas foram posteriormente analisadas e classificadas (seção 4.5.2.1).

4.5.2.1 Classificação dos Eventos de IR

De todos os eventos identificados no espectro de MS/MS, nenhum apresentou IRs múltiplos, e apenas um não truncou a proteína por meio de alterações na fase de leitura ou pela adição de *stop* códons prematuros. No geral, os IRs identificados apresentaram características interessantes e inesperadas, assim como descritas a seguir:

Eventos que não truncam a proteína

Esse evento foi observado, por exemplo, no gene *EDC4* (figura 4.3A). Ele é caracterizado por não quebrar a proteína (não inserir *stop* códons ou alterar fase de leitura) e por inserir novos aminoácidos. Esses eventos são teoricamente esperados para IRs que apresentam algum papel funcional, uma vez que as sequências das proteínas e a estabilidade dos mRNAs não são drasticamente afetadas.

Para melhor entender o papel biológico desse tipo de IR, uma análise de domínios para todos os eventos encontrados na CDS foi feita utilizando o banco de dados Pfam-A e o programa HMMER 3.0 (Finn *et al.*, 2010). Atualmente, o Pfam-A possui aproximadamente 11.912 famílias de domínios devidamente curadas, as quais podem ser encontradas em aproximadamente 75% de todas as proteínas conhecidas (Finn *et al.*, 2010). No total, 28 íntrons retidos mostraram codificar pelo menos um domínio completo, enquanto 107 íntrons retidos mostraram codificar parcialmente pelo menos um domínio. Esses resultados são significativos, visto que números menores foram observados anteriormente (Galante *et al.*, 2004; Hiller *et al.*, 2005). Quase todos os domínios completos e parciais foram identificados pelo banco de dados de alta confiança. Íntrons retidos apresentaram uma significativa alta frequência de domínios comparados com os íntrons não retidos ($p = 0,00, d = 1, \chi^2 = 7948,76$), e uma baixa frequência comparados com éxons ($p = 0,00, d = 1, \chi^2 = 2211,39$).

De 135 eventos apresentando domínios completos ou parciais, nenhum mostrou *stop* códons prematuros, significando que IRs que não quebram a proteína estão diretamente relacionados com processos biológicos. Esse resultado suporta o estudo de Hiller *et al.* (2005), o qual fez predições de IRs baseadas apenas na identificação de domínios proteícos. Essa estratégia pode ser bastante útil, embora seja limitada apenas a casos para os quais existam domínios.

O conteúdo de CG entre IRs apresentando domínios completos e parciais também foi analisado. Domínios completos foram observados com uma grande frequência em IRs de baixo conteúdo CG (média 49%, $p = 9,45 \times 10^{-6}, d = 188, t = -4.55$). IRs apresentando domínios parciais, por outro lado, mostraram um conteúdo CG médio de 58%. Diferentemente do observado, era esperado um maior conteúdo CG para os íntrons que possuiam domínios completos. Uma análise posterior mostrou que o domínio completo mais frequente nos IRs é o *zf-C2H2*, um fator de transcrição muito pequeno e bastante comum em mamíferos (Finn *et al.*, 2010). Aparentemente, o critério "completo" ou "parcial" fornecido pelo HMMER (Eddy, 2011) não é mais importante que o papel do domínio no contexto celular.

***Start* Códons Alternativos**

Nessa categoria de evento, o *start* códon padrão do mRNA deixa de ser utilizado porque o IR insere um *stop* códon prematuro ou altera a fase de leitura da sequência logo no ínicio do mRNA. Porém, um novo *start* códon pode ser inserido pelo IR e utilizado pelo spliceossomo. Como resultado a proteína alternativa resultante não deverá apresentar grandes diferenças quando comparada à proteína original. Esse tipo de evento pôde ser observado no gene *RPL29* (figura 4.4D).

Em um outro exemplo similar, o IR não introduz um *start* códon viável (figura 4.4E). Dessa

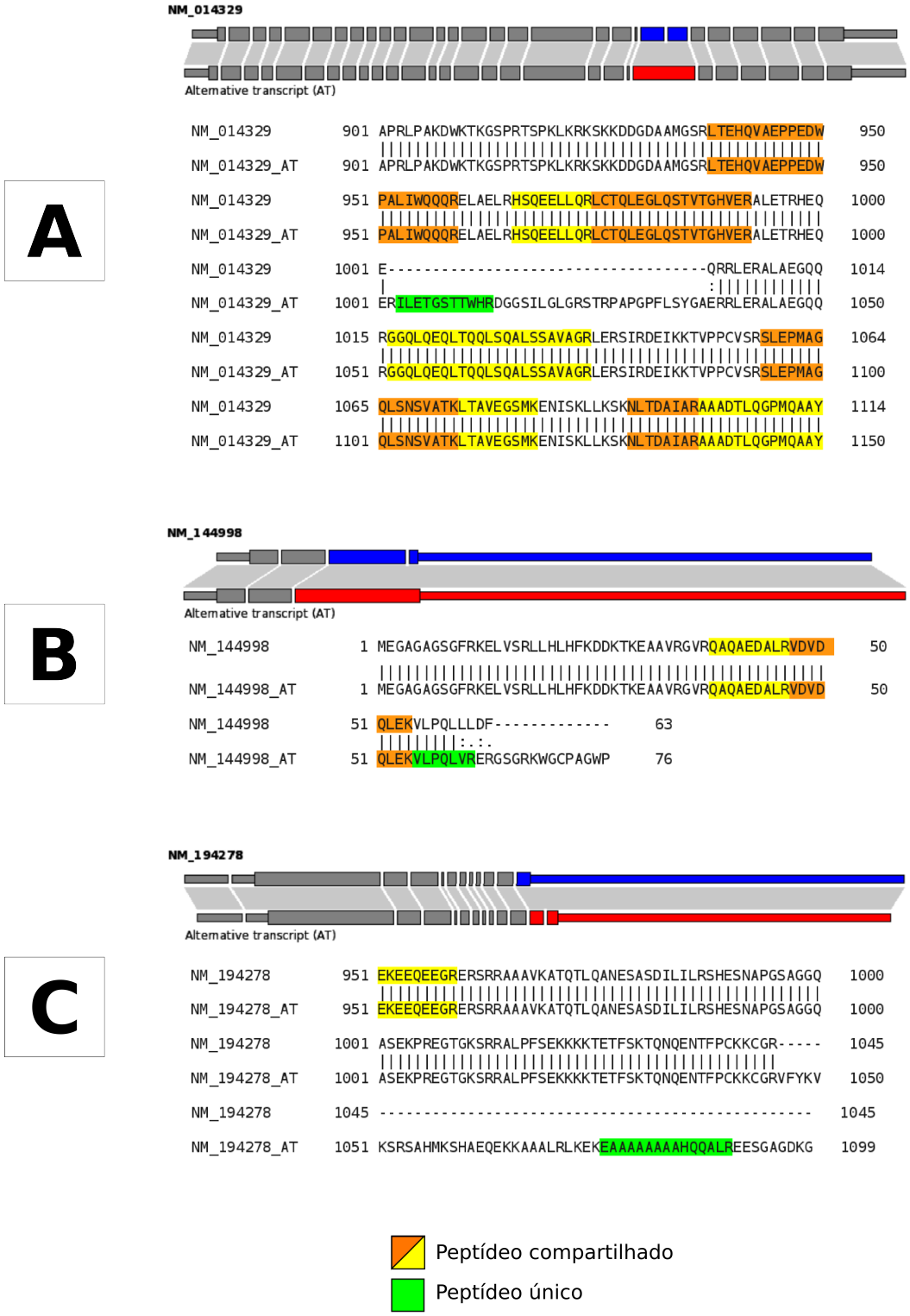

Figura 4.3: *Eventos de IR, suportados por sequências expressas, que foram validados no proteoma. A: gene EDC4; B: gene STRA13 ; C: gene ELMSAN1.*

forma, o spliceossomo utiliza um *start* códon *downstream* ao íntron, normalmente encontrado no éxon adjacente. Esse tipo de evento é difícil de ser corelacionado com IRs, uma vez que as respectivas proteínas alternativas não apresentam peptídeos evento-específicos. Portanto, esses casos podem ser o resultado de diferentes ASEs, mutações ou outros mecanismos de regulação.

Em relação aos IRs encontrados na 5' UTR, não foi possível identificar nenhum evento inserindo algum *start* códon alternativo viável.

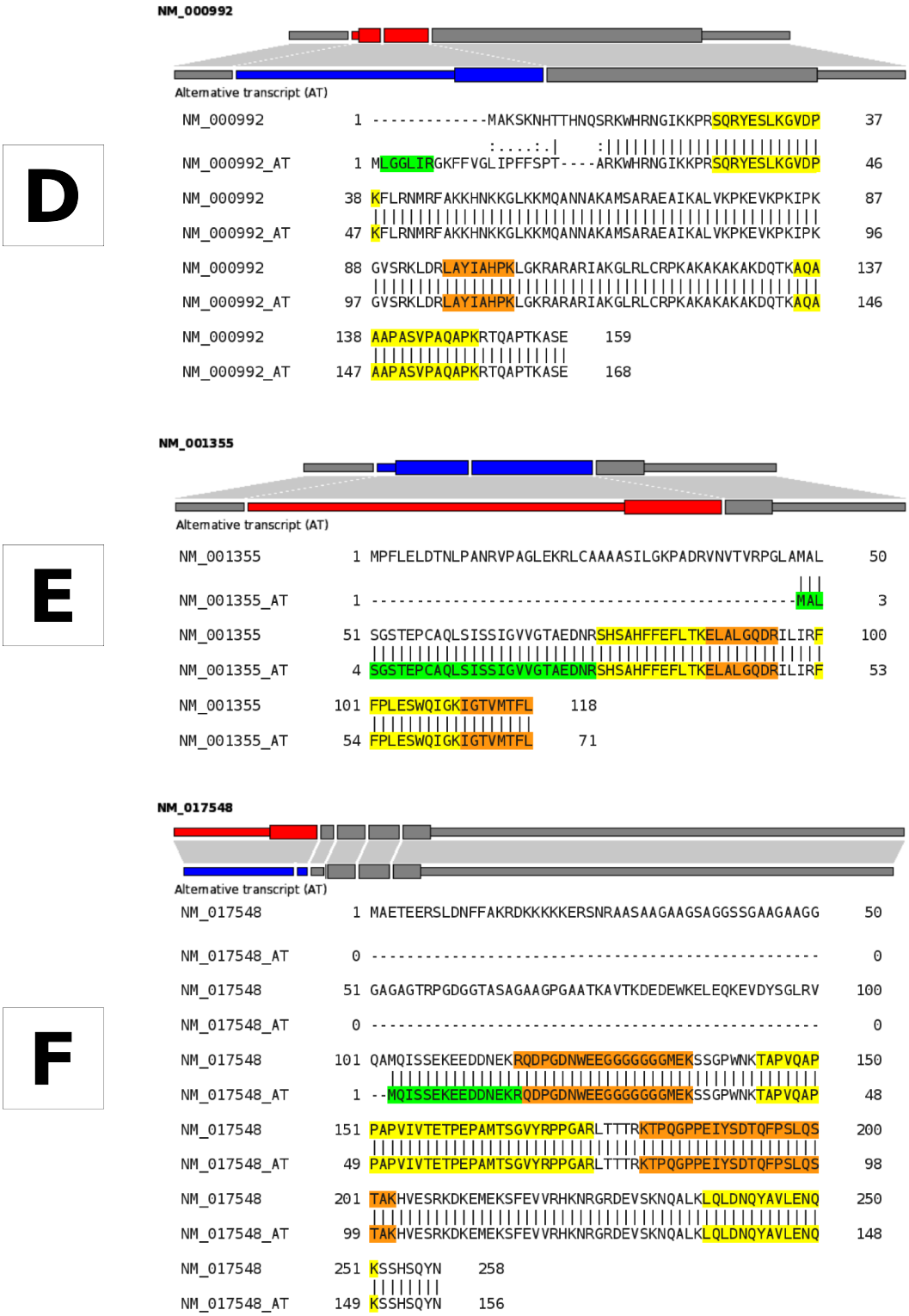

Figura 4.4: *Continuação da figura 4.3. D: gene RPL29; E: gene DDT; F: gene CDV3. Informações adicionais podem ser vistas na tabela 4.3.*

Stop Códons Alternativos

Grande parte dos IRs introduzem *stop* códons prematuros, reduzindo o tamanho da proteína e aumentando o tamanho da sequência 3' UTR. Esse tipo de evento pôde ser observado nos genes *FTSJ1* e *STRA13* (figura 4.3B). Similarmente, alguns IRs podem não introduzir *stop* códons porém podem alterar a fase de leitura do mRNA. Nesses casos, o próximo *stop* códon encontrado no éxon adjacente pode ser utilizado. Esse fenômeno foi observado para o gene *SF1* (tabela

4.3).

Foi observado que IRs próximos aos éxons da região 3' UTR apresentaram uma maior probabilidade de introduzirem *stop* códons prematuros quando comparados com IRs encontrados em diferentes posições ao longo da CDS ($p = 6,08 \times 10^{-2}, d = 1, \chi^2 = 2.75$). Uma explicação é que os eventos próximos à 3' UTR apresentam uma menor tendência de afetar drasticamente a sequência de aminoácidos da proteína e, consequentemente, sua função. Esse resultado é suportado por Ezkurdia *et al.* (2012), o qual mostrou que proteínas variantes normalmente apresentam pequenas diferenças quando comparadas às suas respectivas sequências de referência. Porém, também sabe-se que a frequência de ESTs é maior na região 3' UTR (Hiller *et al.*, 2005), e isso pode contribuir para uma maior frequência de retenções observadas nessa região.

Íntrons Crípticos

Uma alta frequência de íntrons crípticos foi observada. Esses eventos são diferentes de deleções, porque eles são encontrados entre bordas GT-AG ou entre outros sítios de *splicing* variantes, assim como definido pelo programa SIM4 (Florea *et al.*, 1998). Íntrons crípticos são capazes de mudar a fase de leitura do mRNA, porém não podem adicionar *start* ou *stop* códons. No total, quatro eventos distintos desse tipo foram validados, e foram encontrados nos genes CDV3 (figura 4.4F), *ELMSAN1* (figura 4.3C), *CKMT1B* e *HNRNPD*.

4.6 Discussão

Foi apresentada uma análise de larga escala de eventos de retenção de íntron para todos os genes humanos codificantes. Os eventos de retenção foram avaliados em busca de alguma significância biológica, uma vez que diversos eventos podem ser o resultado de pré-mRNAs não/parcialmente processados (Galante *et al.*, 2004).

Os resultados mostram que aproximadamente 48% de todos os genes codificadores possuem pelo menos um evento de retenção de íntron. Essa predição é independente da existência de sequências ortólogas ou outros métodos restringentes, os quais são úteis para identificar eventos de *splicing* alternativo bastante comuns (Sorek e Ast, 2003). O objetivo desse trabalho, porém, foi explorar todas as probabilidades de eventos de retenção, dos quais muitos ainda são desconhecidos. Desde 2004, o número de sequências expressas disponíveis triplicou (Benson *et al.*, 2012), sem considerar as sequências obtidas a partir de sequenciamentos de *next-generation*, mais explorados apenas recentemente (Roberts *et al.*, 2011; Trapnell *et al.*, 2009; Wheeler *et al.*, 2007). Consequentemente, uma alta frequência de genes afetados por eventos de retenção de íntrons já era esperada comparada com trabalhos anteriores (Galante *et al.*, 2004; Kan *et al.*, 2002).

Considerando o exposto, informações discrepantes não foram encontradas (Galante *et al.*, 2004; Hiller *et al.*, 2005; Tan *et al.*, 2007). Sabe-se que bancos de dados públicos de sequência expressas são enriquecidas com bibliotecas de tumor (Kelso *et al.*, 2003), as quais podem contribuir para uma maior frequência de eventos incomuns de *splicing* alternativo (Venables, 2004). Esse problema é atualmente difícil de ser evitado e, portanto, surgem dúvidas sobre o papel biológico da ampla maioria dos eventos já identificados.

Apesar da dificuldade de identificar IRs funcionalmente viáveis, alguns métodos podem ser utilizados para predizê-los (Hiller *et al.*, 2005; Sorek e Ast, 2003). Por exemplo, Hiller *et al.* (2005) fez a predição de IRs com sucesso a partir da identificação de domínios proteícos. Nossas análises suportam essa estratégia, que parece ser bastante útil, porém é limitada apenas para poucos casos nos quais domínios podem ser identificados. Atualmente, análises de espectrometria de massas tem sido utilizadas para explorar ASEs em nível proteômico (Chang *et al.*, 2010; Power *et al.*, 2009). Essas estratégias solucionam antigas limitações e podem, além do mais, revelar novas proteínas. Esse trabalho foi capaz de identificar e classificar proteínas bastante incomuns por meio de análises utilizando MS/MS, sugerindo consequentemente alguma evidência funcional.

Portanto, esse trabalho atualizou os dados estatísticos disponíveis para os eventos de retenção de íntrons utilizando novos dados de sequências expressas. A identificação de eventos de IR que apresentam significância biológica ainda continua sendo uma tarefa difícil. Porém, visto que dife-

rentes tipos de IRs mostraram ser viáveis proteomicamente, sugere-se que a análise de eventos de *splicing* alternativo não deve ser feita de forma muito restritiva. Trabalhos futuros devem focar a expansão dos dados e explorar novas formas de avaliar o papel funcional do eventos de IR e outros tipos de eventos de *splicing* alternativo.

Capítulo 5

Conclusão

Os projetos aqui apresentados abordaram os aspectos mais incomuns relacionados com os eventos de *splicing* alternativo, que é sem dúvida uma ferramenta incrível que contribuiu para o desenvolvimento da complexidade encontrada nos eucariontes superiores. Por meio do SPLOOCE, por exemplo, foi possível verificar que a complexidade encontrada no transcriptoma humano pode ser, em algumas vezes, incomum e/ou incompreensível. Quanto aos eventos de retenção de íntron, alguns casos desconhecidos foram observados no proteoma, sugerindo, consequentemente, um possível papel biológico. Claramente, muito pouco ainda é conhecido sobre o transcriptoma e proteoma humano. Avanços nas tecnologias de RNASeq e espectrometria de massa permitirão, em um futuro breve, um melhor entendimento sobre o funcionamento celular e, o desenvolvimento de novas estratégias terapêuticas e diagnósticas contra uma ampla variedade de doenças.

5.1 Sugestões Futuras

Com base no contexto dessa tese, as seguintes sugestões futuras foram definidas:

- Explorar a complexidade do transcriptoma com o objetivo de melhor entender a existência de determinados eventos complexos de *splicing* alternativo.
- Validar proteomicamente os eventos complexos de *splicing* alternativo do banco de dados do SPLOOCE, assim como já feito para os eventos de retenção de íntron.
- Ampliar o banco de dados do SPLOOCE, que é construído basicamente a partir de sequências clássicas (RefSeqs, mRNAs, ESTs), mas possui apenas poucas sequências provenientes de sequenciamentos de *next-generation*.
- Desenvolver ferramentas para a análise de *splicing* alternativo. Atualmente, exitem poucas ferramentas que permitem o usuário analisar eventos em um *data set* personalizado.
- Explorar os sítios de *splice* duplamente específicos. Na literatura, pouco tem sido discutido sobre o assunto.

Apêndice A

Artigo do SPLOOCE

TECHNICAL PAPER

RNA Biology 9:11, 1–6; November 2012; © 2012 Landes Bioscience

SPLOOCE
A new portal for the analysis of human splicing variants

José Eduardo Kroll,[1,2] Pedro A. F. Galante,[1,†] Daniel T. Ohara,[1] Fábio C. P. Navarro,[1] Lucila Ohno-Machado[3] and Sandro J. de Souza[1,*]

[1]Laboratory of Computational Biology; Ludwig Institute for Cancer Research; São Paulo, Brazil; [2]Inter-institutional Program on Bioinformatics; University of São Paulo; São Paulo, Brazil; [3]Bioinformatics Group; Division of Biomedical Informatics; University of California San Diego; La Jolla, CA USA

[†]Current Affiliation: Group of Bioinformatics; Instituto de Ensino e Pesquisa - Hospital Sírio-Libanês; São Paulo, Brazil

Keywords: bioinformatics, alternative splicing, combined alternative splicing events, database, method of analysis, regular expressions, next-generation sequencing

Abbreviations: ASE, alternative splicing event; CASE, complex alternative splicing event; RNA, ribonucleic acid; mRNA, messenger ribonucleic acid; UCSC, University of California, Santa Cruz; NCBI, National Center for Biotechnology Information; EST, expressed sequence tag; RNA-seq, whole transcriptome shotgun sequencing; SRA, sequence read archive; RefSeq, reference sequence; BLAT, BLAST-like alignment tool; cDNA, complementary deoxyribonucleic acid; NGS, next-generation sequencing; eVOC, expressed sequence annotation for humans; ORF, open reading frame; Pfam, protein families; ES, exon skipping; DSS, dual-specific splicing; Regex, regular expression; SIM4, Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence; HMMER, Computer Program for Biosequence Analysis Using Profile Hidden Markov Models

Understanding alternative splicing is crucial to elucidate the mechanisms behind several biological phenomena, including diseases. The huge amount of expressed sequences available nowadays represents an opportunity and a challenge to catalog and display alternative splicing events (ASEs). Although several groups have faced this challenge with relative success, we still lack a computational tool that uses a simple and straightforward method to retrieve, name and present ASEs. Here we present SPLOOCE, a portal for the analysis of human splicing variants. SPLOOCE uses a method based on regular expressions for retrieval of ASEs. We propose a simple syntax that is able to capture the complexity of ASEs.

Introduction

Alternative splicing events (ASEs) are present in almost all multi-exonic human genes[1,2] and are believed to be one of the most significant components behind the complexity of multi-cellular organisms.[1,3,4] Furthermore, ASEs are clearly involved in the etiology of a wide variety of diseases, including cancer,[5-7] ischemia[8] and other common human disorders.[9] Recently, several studies have shown that constitutive and alternative splicing are regulated by a complex network of cellular elements, which include a set of trans-acting factors and cis-acting sequences found in the primary RNAs.[10-18]

The complex regulation of splicing and the high frequency of ASEs explain the appearance of Complex Alternative Splicing Events (CASEs), which are composed by a regulated combination of two or more single ASEs in transcripts from the same gene, or even in the same transcript. The most striking example of CASE is the Dscam in Drosophila, a gene containing a cluster of 48 mutually exclusive exons that, in principle, can generate thousands of splicing variants.[19]

In humans, some ASEs and CASEs occurring in oncogenes and tumor suppressor genes have already been associated to cancer.[6,20-23] For example, the gene NTRK1 (nerve growth factor) has a sequence variant, TrkAIII, which is common in certain tumors and lacks three exons that affect a regulatory immunoglobulin-like domain.[24] Further CASE examples include the gene CD44, which is a known marker of malignancy and invasiveness and has about ten ASEs that can occur in different combinations in its region coding for the extra-cellular portion of the protein.[20,21,23]

In spite of the efforts of other groups,[25-27] a simple and efficient nomenclature to take into account all the variability generated by alternative splicing, especially for CASEs, is still missing. Here, we present a web portal, SPLOOCE, which uses a method based on regular expressions with an associated syntax. SPLOOCE provides a series of tools that allow users to profile splicing variants and analyze their functional impacts.

*Correspondence to: Sandro J. de Souza; Email: sandro@i2bio.org
Submitted: 06/12/12; Accepted: 09/11/12
http://dx.doi.org/10.4161/rna.22182

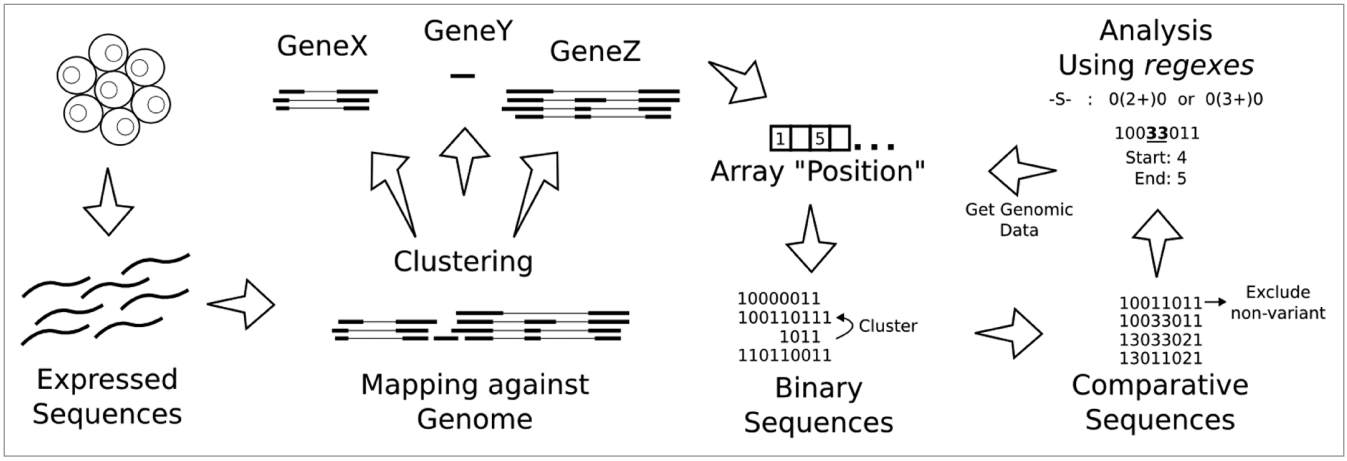

Figure 1. General strategy to process and analyze CASEs in a set of expressed sequences.

Results and Discussion

For the sake of space and clarity, we opted to describe the method used in this report as supplemental material, although an overview is present in **Figure 1**. In this section we present an implementation of the method in a computational tool, SPLOOCE and illustrate the use of SPLOOCE discussing a few examples.

Implementation. To make the method described in the supplemental material available to the community in an easy way, a web tool called SPLOOCE was implemented. SPLOOCE is available at http://www.bioinformatics-brazil.org/splooce. Data sources include RefSeqs, mRNAs, ESTs and data from NGS.

To help the users, SPLOOCE provides in the query box a quick reference table explaining the syntax and showing some illustrative examples. All ASEs and CASEs identified by SPLOOCE can also be displayed for a specific gene by simply typing the gene name between quotes in the query box (**Fig. 2**). SPLOOCE also provides advanced options for querying. Filters for chromosome, strand, gene name and sequence type are provided. Users can also evaluate the specificity of ASEs and CASEs expression regarding both tissue and pathology. A score for expression specificity is provided, which is a simple X^2 distribution analysis done among the expressed sequences supporting the corresponding variant. The analysis is based on the annotation provided by eVOC[33] for ESTs and manual curation for NGS sequences.

Results provided by SPLOOCE can be downloaded in a GFF file format. By default, results are shown in a table containing chromosome, genomic position, gene name and a pictorial view of the respective ASE or CASE, followed by the amount of their respective supporting sequences. SPLOOCE also provides a link to the UCSC Genome Browser (with tracks), and a local link for additional information.

When a Reference Sequence (RefSeq) is involved in an ASE, it is used as a template for creating a new mRNA sequence containing the specified event. For each of these new sequences, SPLOOCE predicts its open reading frame (ORF), which is then translated to a protein. Moreover, aiming to infer additional biological significance, SPLOOCE analyzes the protein domains using Pfam data and HMMER 3.0 program.[34] All these data are shown graphically (**Fig. 2C**).

Analysis. To illustrate the use of SPLOOCE, some basic questions about the frequency and mode of alternative splicing events were addressed. **Table 1** shows the frequency of all types of ASEs in our data set. As expected, exon skipping (ES) is the most frequent type of alternative splicing. One interesting aspect that our method allowed us to explore is the distribution of dual-specific splicing (DSS) events. This type of event was found in 53 (0.27%) human genes (**Table S2**). In a less restrictive analysis, without considering the number of supporting sequences, the number of genes showing this type of event increased to 577. DSSs events were found to occur frequently in genes such as DIABLO, IRF3 and MAG, and they can occur together with other events as shown in **Table 2**.

Another interesting feature that our method allows us to evaluate is the combination of events occurring in the same mRNA molecule. Each pair-wise comparative sequence in our pipeline contains on average 1.6 events and 46.87% and 12.21% of these pairwise comparisons report more than one or two events, respectively.

To better understand what influences the frequency of ASEs and CASEs, some patterns were further explored. For example, the combination of two adjacent ES events is significantly more frequent among all sets of CASEs (**Table 3**). This excess is absent in situations when both events are not adjacent, like in the patterns –s-E-s- and –s-E-E-s- . Do these adjacent events tend to maintain the phase of an ORF? When adjacent, 60.78% of -E-E-s-s-E-E- maintains the ORF. This is significantly higher than what one would expect by chance based on all pairs of exons in the human genome that maintain an ORF ($p < 0.001$). This strongly suggests that adjacent ES events are under selection to maintain the ORF. The same pattern is not observed for other types of CASEs (data not shown).

Material and Methods

Public data. The human genome reference sequence (NCBI36/hg18) was downloaded from UCSC Genome Bioinformatics

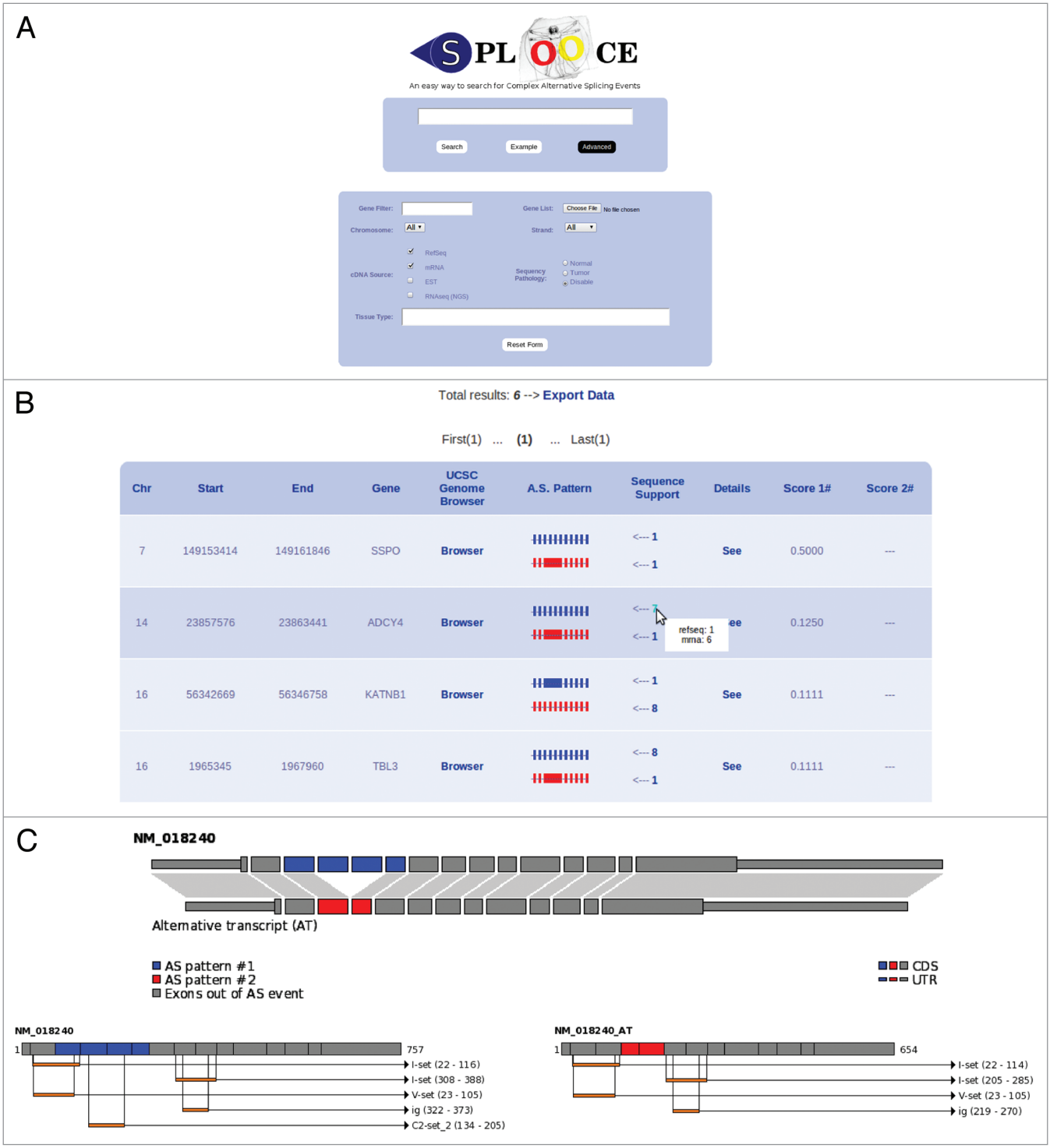

Figure 2. (A) SPLOOCE query form also showing a tab for advanced parameters. (B) Example of the table of results for a query. (C) Some results that can be found in the section "Details" provided by SPLOOCE. In this example, a double skipping (syntax: -e-s-s-e-) that codifies a protein domain is shown.

portal (http://genome.ucsc.edu). RefSeq sequences were downloaded from the Reference Sequence database (release 25; http://www.ncbi.nlm.nih.gov/RefSeq/). A total of 203,649 mRNAs sequences were downloaded from UCSC Genome Bioinformatics portal (file mrna.fa, for Homo sapiens). ESTs sequences were downloaded from dbEST (http://www.ncbi.nlm.nih.gov/dbEST/). RNA-seq reads were downloaded from SRA database (http://www.ncbi.nlm.nih.gov/sra; IDs: SRX003935,

Table 1. Frequency of the major types of ASEs

Simple alternative splicing event	Genes	Total Events	Events per Gene
Exon skipping	10125 (51,77%)	38060	1,95
Alternative 3' splice site	7490 (38,30%)	30172	1,54
Alternative 5' splice site	7258 (37,11%)	27585	1,41
Intron retention	6565 (33,57%)	12632	0,65
Dual-specific splice site	53 (0,27%)	112	0,0057

Table 2. Frequency of DSS events coupled to other types of ASEs

Syntax	Frequency (Genes)	Pattern (Simple)
d	577 (2.95%)	23 or 32
-d-	181 (0.93%)	0230 or 0320
-d-s-	85 (0.43%)	023030 or 032020
-Ed-	57 (0.29%)	01230 or 01320
-d-T	57 (0.29%)	023031 or 032021
-dE-	56 (0.28%)	02310 or 03210
f-d-T	26 (0.13%)	12023031 or 13032021
E-d-T	20 (0.10%)	1023031 or 1032021
-EdE-	17 (0.087%)	012310 or 013210
-d-t	11 (0.06%)	023021 or 032031
-d-S-	8 (0.04%)	023020 or 032030
-df-	5 (0.025%)	023120 or 032130
-tD-	3 (0.015%)	021320 or 031230
f-D-T	2 (0.01%)	12032021 or 13023031

SRX003934, SRX003933, SRX003932, SRX003931, SRX003930, SRX003929, SRX003928, SRX003927, SRX003926).

Sequence alignment. All RefSeqs, mRNAs and ESTs sequences were aligned to the human genome using the protocol described previously.[28] In brief, first all long sequences (RefSeq, mRNAs and ESTs) were mapped against the human genome using the BLAT alignment tool[29] and only the best alignment for each sequence was selected. Next, those transcripts showing alignment identity greater than 95% and a covering more than 90% of its sequence length were remapped using SIM4[30] and stored in a relational database. RNA-seq data were mapped to human genome using Tophat-based pipeline,[31] and all mapped reads were submitted to Cufflinks.[32] Default parameters were used in both algorithms.

Sequence clustering. All mapped sequences sharing the same genomic region were grouped together using a gene-oriented strategy as described previously.[4,28] First, all RefSeq sequences were annotated based on the corresponding "official gene name" from NCBI-Gene (http://www.ncbi.nlm.nih.gov/gene/). Second, all non-RefSeq transcripts (mRNAs, ESTs and RNA-seq) presenting multiple exons and sharing one or more exon-intron boundaries (splice junctions) with a RefSeq sequence were merged together. Third, the remaining non-RefSeq transcripts showing only one exon and overlapping greater than 30 nt with a RefSeq transcript were grouped together. All clustering information was stored in a relational database.

Alternative splicing. Analyses of ASEs and CASEs were done using regular expressions, as detailed in the supplemental material. In the analysis of ASEs, only cDNA clusters containing RefSeqs and/or mRNAs or more than 10 ESTs/RNA-seq were used. The redundancy of ASEs, such as exon skipping and intron retention, was eliminated by comparing all events of the same type from each gene. Afterwards, all events showing position overlap were clustered together and counted as one event. All other events, such as 3'/5' alternative splice sites and dual-specific splice sites, were counted by verifying the position of the alternative splice site. The number of genes affected by specific CASEs was defined through the analysis of the full list of comparative matrices, not considering the number of supporting sequences.

Conclusion

Although previous methods and interfaces have been proposed[25-27] for the study of alternative splicing, all of them present limitations as discussed before.[27] SPLOOCE, described here, is an efficient and complementary alternative for the analysis of alternative splicing events due to its high flexibility in the querying patterns and variety of applications.

Like ASTALAVISTA,[27] the method used by SPLOOCE is based in a comparison of all transcripts for a given locus. SPLOOCE, however, uses a notation system based on regular expressions to provide a simple and straightforward syntax for splicing events. The design of the syntax was developed to provide a set of simple and intuitive characters, and is actually capable of representing any CASE pattern, including those that have rare DSS events. The proposed syntax was successfully implemented and it may become a standard way for representing ASEs and CASEs.

Acknowledgments

Part of this work was supported by grants from the Fogarty International Center, National Institute of Health (D43TW007015 to L.O.M.); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); and from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (2007/55790–5 to S.J.S. and 2009/53853–5 to S.J.S.). Funding for open access charge: Fundação de Amparo à Pesquisa do Estado de São Paulo.

Supplemental Material

Supplemental material may be found here:
www.landesbioscience.com/journals/rna/article/22181

Table 3. Number of genes affected by different types of ASEs and CASEs

Syntax	Freq. Genes	Syntax	Freq. Genes	Syntax	Freq. Genes
-s-	14461 (73.04%)	-s-F-	3258 (16.66%)	-s-S-s-	597 (3.05%)
-f-	11220 (57.37%)	-s-f-	3052 (15.60%)	-t-E-t-	591 (3.02%)
-s-s-	10020 (51.23%)	-s-E-S-	2718 (13.90%)	-f-E-f-	549 (2.81%)
-s-s-s-	6844 (34.99%)	-s-S-S-	2627 (13.43%)	-f-E-F-	541 (2.77%)
-f-S-	5733 (29.31%)	-t-S-	2563 (13.10%)	-rR-	430 (2.20%)
-s-S-	5532 (28.28%)	-s-s-S-	2168 (11.08%)	-r-r-	272 (1.39%)
-s-T-	5302 (27.11%)	-s-E-E-s-	1627 (8.32%)	-r-R-	228 (1.17%)
-r-	4687 (23.96%)	-s-E-E-S-	1580 (8.08%)	-f-E-t-	226 (1.16%)
-s-t-	3934 (20.11%)	-t-T-	1150 (5.88%)	-rrr-	224 (1.15%)
-f-s-	3710 (18.97%)	-f-F-	1033 (5.28%)	-f-E-T-	218 (1.11%)
-f-T-	2866 (14.65%)	-t-t-	1022 (5.23%)	-s-E-S-E-s-	73 (0.37%)
-f-t-	2800 (14.32%)	-f-f-	1021 (5.22%)	-rRR-	42 (0.21%)
-s-E-s-	2768 (14.15%)	-rr-	928 (4.74%)	-rrR-	37 (0.19%)
-E-r-E-	2222 (11.36%)	-t-E-T-	629 (3.22%)	-rRr-	27 (0.14%)

References

1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature 2008; 456:470-6; PMID:18978772; http://dx.doi.org/10.1038/nature07509.
2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008; 40:1413-5; PMID:18978789; http://dx.doi.org/10.1038/ng.259.
3. Modrek B, Lee C. A genomic view of alternative splicing. Nat Genet 2002; 30:13-9; PMID:11753382; http://dx.doi.org/10.1038/ng0102-13.
4. Galante PA, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ. Detection and evaluation of intron retention events in the human transcriptome. RNA 2004; 10:757-65; PMID:15100430; http://dx.doi.org/10.1261/rna.5123504.
5. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet 2007; 8:749-61; PMID:17726481; http://dx.doi.org/10.1038/nrg2164.
6. Venables JP. Aberrant and alternative splicing in cancer. Cancer Res 2004; 64:7647-54; PMID:15520162; http://dx.doi.org/10.1158/0008-5472.CAN-04-1910.
7. Kirschbaum-Slager N, Parmigiani RB, Camargo AA, de Souza SJ. Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. Physiol Genomics 2005; 21:423-32; PMID:15784694; http://dx.doi.org/10.1152/physiolgenomics.00237.2004.
8. Daoud R, Mies G, Smialowska A, Oláh L, Hossmann KA, Stamm S. Ischemia induces a translocation of the splicing factor tra2-beta 1 and changes alternative splicing patterns in the brain. J Neurosci 2002; 22:5889-99; PMID:12122051.
9. Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. Nat Biotechnol 2004; 22:535-46; PMID:15122293; http://dx.doi.org/10.1038/nbt964.
10. Alló M, Buggiano V, Fededa JP, Petrillo E, Schor I, de la Mata M, et al. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. Nat Struct Mol Biol 2009; 16:717-24; PMID:19543290; http://dx.doi.org/10.1038/nsmb.1620.
11. Schor IE, Rascovan N, Pelisch F, Alló M, Kornblihtt AR. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. Proc Natl Acad Sci USA 2009; 106:4325-30; PMID:19251664; http://dx.doi.org/10.1073/pnas.0810666106.
12. Muñoz MJ, Pérez Santangelo MS, Paronetto MP, de la Mata M, Pelisch F, Boireau S, et al. DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation. Cell 2009; 137:708-20; PMID:19450518; http://dx.doi.org/10.1016/j.cell.2009.03.010.
13. Ip JY, Schmidt D, Pan Q, Ramani AK, Fraser AG, Odom DT, et al. Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. Genome Res 2011; 21:390-401; PMID:21163941; http://dx.doi.org/10.1101/gr.111070.110.
14. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. Science 2010; 327:996-1000; PMID:20133523; http://dx.doi.org/10.1126/science.1184208.
15. Batsché E, Yaniv M, Muchardt C. The human SWI/SNF subunit Brm is a regulator of alternative splicing. Nat Struct Mol Biol 2006; 13:22-9; PMID:16341228; http://dx.doi.org/10.1038/nsmb1030.
16. Saint-André V, Batsché E, Rachez C, Muchardt C. Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. Nat Struct Mol Biol 2011; 18:337-44; PMID:21358630; http://dx.doi.org/10.1038/nsmb.1995.
17. Sakabe NJ, de Souza SJ. Sequence features responsible for intron retention in human. BMC Genomics 2007; 8:59; PMID:17324281; http://dx.doi.org/10.1186/1471-2164-8-59.
18. de Souza JES, Ramalho RF, Galante PAF, Meyer D, de Souza SJ. Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders. Nucleic Acids Res 2011; 39:4942-8; PMID:21398627; http://dx.doi.org/10.1093/nar/gkr081.
19. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, et al. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 2000; 101:671-84; PMID:10892653; http://dx.doi.org/10.1016/S0092-8674(00)80878-8.
20. Hayes GM, Dougherty ST, Davis PD, Dougherty GJ. Molecular mechanisms regulating the tumor-targeting potential of splice-activated gene expression. Cancer Gene Ther 2004; 11:797-807; PMID:15359288; http://dx.doi.org/10.1038/sj.cgt.7700759.
21. Galiana-Arnoux D, Del Gatto-Konczak F, Gesnel MC, Breathnach R. Intronic UGG repeats coordinate splicing of CD44 alternative exons v8 and v9. Biochem Biophys Res Commun 2005; 336:667-73; PMID:16137657; http://dx.doi.org/10.1016/j.bbrc.2005.08.153.
22. Tanko Q, Franklin B, Lynch H, Knezetic J. A hMLH1 genomic mutation and associated novel mRNA defects in a hereditary non-polyposis colorectal cancer family. Mutat Res 2002; 503:37-42; PMID:12052501; http://dx.doi.org/10.1016/S0027-5107(02)00031-3.
23. Venables JP. Unbalanced alternative splicing and its significance in cancer. Bioessays 2006; 28:378-86; PMID:16547952; http://dx.doi.org/10.1002/bies.20390.
24. Tacconelli A, Farina AR, Cappabianca L, Desantis G, Tessitore A, Vetuschi A, et al. TrkA alternative splicing: a regulated tumor-promoting switch in human neuroblastoma. Cancer Cell 2004; 6:347-60; PMID:15488758; http://dx.doi.org/10.1016/j.ccr.2004.09.011.
25. Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O. Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. Bioinformatics 2006; 22:1211-6; PMID:16500940; http://dx.doi.org/10.1093/bioinformatics/btl067.
26. Malko DB, Makeev VJ, Mironov AA, Gelfand MS. Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes. Genome Res 2006; 16:505-9; PMID:16520458; http://dx.doi.org/10.1101/gr.4236606.
27. Sammeth M, Foissac S, Guigó R. A general definition and nomenclature for alternative splicing events. PLoS Comput Biol 2008; 4:e1000147; PMID:18688268; http://dx.doi.org/10.1371/journal.pcbi.1000147.
28. Galante PAF, Vidal DO, de Souza JE, Camargo AA, de Souza SJ. Sense-antisense pairs in mammals: functional and evolutionary considerations. Genome Biol 2007; 8:R40; PMID:17371592; http://dx.doi.org/10.1186/gb-2007-8-3-r40.
29. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res 2002; 12:656-64; PMID:11932250.
30. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 1998; 8:967-74; PMID:9750195.
31. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25:1105-11; PMID:19289445; http://dx.doi.org/10.1093/bioinformatics/btp120.
32. Trapnell C, Williams BA, Pertea G, Mortazavi AM, 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; 28:511-5; PMID:20436464; http://dx.doi.org/10.1038/nbt.1621.

33. Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, et al. eVOC: a controlled vocabulary for unifying gene expression data. Genome Res 2003; 13(6A):1222-30; PMID:12799354; http://dx.doi.org/10.1101/gr.985203.

34. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res 2010; 38(Database issue):D211-22; PMID:19920124; http://dx.doi.org/10.1093/nar/gkp985.

Referências Bibliográficas

Aaltomaa *et al.* **(2001)** S Aaltomaa, P Lipponen, M Ala-Opas e V-M Kosma. Expression and prognostic value of cd44 standard and variant v3 and v6 isoforms in prostate cancer. *European urology*, 39(2):138–144. URL http://www.karger.com/Article/Fulltext/52428. Citado na pág. 9

Alló *et al.* **(2009)** Mariano Alló, Valeria Buggiano, Juan P Fededa, Ezequiel Petrillo, Ignacio Schor, Manuel de la Mata, Eneritz Agirre, Mireya Plass, Eduardo Eyras, Sherif Abou Elela *et al.* Control of alternative splicing through sirna-mediated transcriptional gene silencing. *Nature structural & molecular biology*, 16(7):717–724. URL http://www.nature.com/nsmb/journal/v16/n7/abs/nsmb.1620.html. Citado na pág. 11

Ameur *et al.* **(2010)** Adam Ameur, Anna Wetterbom, Lars Feuk, Ulf Gyllensten *et al.* Global and unbiased detection of splice junctions from rna-seq data. *Genome Biol*, 11(3):R34. Citado na pág. 6

Au *et al.* **(2010)** Kin Fai Au, Hui Jiang, Lan Lin, Yi Xing e Wing Hung Wong. Detection of splice junctions from paired-end rna-seq data by splicemap. *Nucleic Acids Research*, 38(14):4570–4578. URL http://nar.oxfordjournals.org/content/38/14/4570.short. Citado na pág. 6

Batsché *et al.* **(2005)** Eric Batsché, Moshe Yaniv e Christian Muchardt. The human swi/snf subunit brm is a regulator of alternative splicing. *Nature structural & molecular biology*, 13(1): 22–29. URL http://www.nature.com/nsmb/journal/vaop/ncurrent/full/nsmb1030.html. Citado na pág. 11

Beaudoing *et al.* **(2000)** Emmanuel Beaudoing, Susan Freier, Jacqueline R Wyatt, Jean-Michel Claverie e Daniel Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome research*, 10(7):1001–1010. URL http://genome.cshlp.org/content/10/7/1001.short. Citado na pág. 5

Benson *et al.* **(2012)** Dennis A Benson, Ilene Karsch-Mizrachi, Karen Clark, David J Lipman, James Ostell e Eric W Sayers. Genbank. *Nucleic acids research*, 40(D1):D48–D53. URL http://nar.oxfordjournals.org/content/40/D1/D48.short. Citado na pág. 5, 25, 35

Bertone *et al.* **(2004)** Paul Bertone, Viktor Stolc, Thomas E Royce, Joel S Rozowsky, Alexander E Urban, Xiaowei Zhu, John L Rinn, Waraporn Tongprasit, Manoj Samanta, Sherman Weissman *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306(5705):2242–2246. URL http://www.sciencemag.org/content/306/5705/2242.short. Citado na pág. 5

Black e Grabowski (2003) DL Black e PJ Grabowski. Alternative pre-mrna splicing and neuronal function. Em *Regulation of Alternative Splicing*, páginas 187–216. Springer. URL http://link.springer.com/chapter/10.1007/978-3-662-09728-1_7. Citado na pág. 4

Black (2003) Douglas L. Black. Mechanisms of alternative pre-messenger rna splicing. *Annu Rev Biochem*, 72:291–336. doi: 10.1146/annurev.biochem.72.121801.161720. URL http://dx.doi.org/10.1146/annurev.biochem.72.121801.161720. Citado na pág. 3, 4, 5

Blencowe (2000) B. J. Blencowe. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci*, 25(3):106–110. Citado na pág. 4

Blencowe *et al.* (2006) Benjamin J Blencowe *et al.* Alternative splicing: new insights from global analyses. *Cell*, 126(1):37–48. Citado na pág. 5

Breitbart *et al.* (1987) Roger E Breitbart, Athena Andreadis e Bernardo Nadal-Ginard. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annual review of biochemistry*, 56(1):467–495. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.bi.56.070187.002343. Citado na pág. 5

Brett *et al.* (2000) David Brett, Jens Hanke, Gerrit Lehmann, Sabine Haase, Sebastian Delbrück, Steffen Krueger, Jens Reich e Peer Bork. Est comparison indicates 38% of human mrnas contain possible alternative splice forms. *FEBS letters*, 474(1):83–86. Citado na pág. 5

Brett *et al.* (2002) David Brett, Heike Pospisil, Juan Valcárcel, Jens Reich, Peer Bork *et al.* Alternative splicing and genome complexity. *Nature genetics*, 30(1):29. URL http://europepmc.org/abstract/MED/11743582. Citado na pág. 5

Broeks *et al.* (2003) Annegien Broeks, Jos HM Urbanus, Peter de Knijff, Peter Devilee, Marion Nicke, Karin Klöpper, Thilo Dörk, Arno N Floore e Laura J van't Veer. Ivs10–6t> g, an ancient atm germline mutation linked with breast cancer. *Human mutation*, 21(5):521–528. URL http://onlinelibrary.wiley.com/doi/10.1002/humu.10204/full. Citado na pág. 7

Cáceres e Kornblihtt (2002) Javier F Cáceres e Alberto R Kornblihtt. Alternative splicing: multiple control mechanisms and involvement in human disease. *TRENDS in Genetics*, 18(4): 186–193. URL http://www.sciencedirect.com/science/article/pii/S0168952501026269. Citado na pág. 4

Cartegni *et al.* (2002) Luca Cartegni, Shern L. Chew e Adrian R. Krainer. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*, 3(4):285–298. doi: 10.1038/nrg775. URL http://dx.doi.org/10.1038/nrg775. Citado na pág. 4

Chang *et al.* (2010) Kung-Yen Chang, D Ryan Georgianna, Steffen Heber, Gary A Payne e David C Muddiman. Detection of alternative splice variants at the proteome level in aspergillus flavus. *Journal of proteome research*, 9(3):1209–1217. URL http://pubs.acs.org/doi/abs/10.1021/pr900602d. Citado na pág. 30, 35

Chen *et al.* (2011) Geng Chen, Ruiyuan Li, Leming Shi, Junyi Qi, Pengzhan Hu, Jian Luo, Mingyao Liu e Tieliu Shi. Revealing the missing expressed genes beyond the human reference genome by rna-seq. *BMC genomics*, 12(1):590. Citado na pág. 6

Clarke *et al.* (2000) Luka A Clarke, Isabel Veiga, Gloria Isidro, Peter Jordan, José Silva Ramos, Sergio Castedo e Maria Guida Boavida. Pathological exon skipping in an hnpcc proband with mlh1 splice acceptor site mutation. *Genes, Chromosomes and Cancer*, 29(4):367–370. URL http://onlinelibrary.wiley.com/doi/10.1002/. Citado na pág. 7

Collins *et al.* (2004) FS Collins, ES Lander, J Rogers, RH Waterston e IHGS Conso. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945. Citado na pág. 4

Cork *et al.* (2012) David MW Cork, Thomas WJ Lennard e Alison J Tyson-Capper. Progesterone receptor (pr) variants exist in breast cancer cells characterised as pr negative. *Tumor Biology*, 33 (6):2329–2340. URL http://link.springer.com/article/10.1007/s13277-012-0495-z. Citado na pág. 4

Coulson *et al.* (2000) Judy M Coulson, Jodie L Edgson, Penella J Woll e John P Quinn. A splice variant of the neuron-restrictive silencer factor repressor is expressed in small cell lung cancer: a potential role in derepression of neuroendocrine genes and a useful clinical marker.

Cancer research, 60(7):1840–1844. URL http://cancerres.aacrjournals.org/content/60/7/1840.short. Citado na pág. 8

Cox e Mann (2008) Jürgen Cox e Matthias Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–1372. URL http://www.nature.com/nbt/journal/v26/n12/abs/nbt.1511.html. Citado na pág. 27

Da Cunha *et al.* **(2009)** JPC Da Cunha, PAF Galante, JE De Souza, RF De Souza, PM Carvalho, DT Ohara, RP Moura, SM Oba-Shinja, SKN Marie, WA Silva *et al.* Bioinformatics construction of the human cell surfaceome. *Proceedings of the National Academy of Sciences*, 106(39):16752–16757. URL http://www.pnas.org/content/106/39/16752.short. Citado na pág. 8

Daoud *et al.* **(2002)** Rosette Daoud, GÃ¼nter Mies, Agata Smialowska, Laszlo OlÃ¡h, Konstantin-Alexander Hossmann e Stefan Stamm. Ischemia induces a translocation of the splicing factor tra2-beta 1 and changes alternative splicing patterns in the brain. *J Neurosci*, 22(14):5889–5899. doi: 20026571. URL http://dx.doi.org/20026571. Citado na pág. 4, 25

De Klein *et al.* **(1998)** Annelies De Klein, Peter HJ Riegman, Emilia K Bijlsma, Anneliek Heldoorn, Manja Muijtjens, Michael A den Bakker, Cees JJ Avezaat e Ellen C Zwarthoff. Ag a transition creates a branch point sequence and activation of a cryptic exon, resulting in the hereditary disorder neurofibromatosis 2. *Human molecular genetics*, 7(3):393–398. URL http://hmg.oxfordjournals.org/content/7/3/393.short. Citado na pág. 7

de la Mata *et al.* **(2003)** Manuel de la Mata, Claudio R Alonso, Sebastián Kadener, Juan P Fededa, Matıas Blaustein, Federico Pelisch, Paula Cramer, David Bentley e Alberto R Kornblihtt. A slow rna polymerase ii affects alternative splicing in vivo. *Molecular cell*, 12(2):525–532. Citado na pág. 6, 8

Deckert *et al.* **(2006)** Jochen Deckert, Klaus Hartmuth, Daniel Boehringer, Nastaran Behzadnia, Cindy L. Will, Berthold Kastner, Holger Stark, Henning Urlaub e Reinhard LÃ¼hrmann. Protein composition and electron microscopy structure of affinity-purified human spliceosomal b complexes isolated under physiological conditions. *Mol Cell Biol*, 26(14):5528–5543. doi: 10.1128/MCB.00582-06. URL http://dx.doi.org/10.1128/MCB.00582-06. Citado na pág. 3

DiFeo *et al.* **(2009)** Analisa DiFeo, John A Martignetti e Goutham Narla. The role of klf6 and its splice variants in cancer therapy. *Drug resistance updates*, 12(1-2):1–7. URL http://cat.inist.fr/?aModele=afficheN&cpsidt=21387742. Citado na pág. 7

Eddy (2011) Sean R Eddy. Accelerated profile hmm searches. *PLoS computational biology*, 7(10): e1002195. Citado na pág. 32

Ezkurdia *et al.* **(2012)** Iakes Ezkurdia, Angela del Pozo, Adam Frankish, Jose Manuel Rodriguez, Jennifer Harrow, Keith Ashman, Alfonso Valencia e Michael L Tress. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Molecular biology and evolution*, 29(9):2265–2283. URL http://mbe.oxfordjournals.org/content/29/9/2265.short. Citado na pág. 35

Fairbrother *et al.* **(2002)** William G. Fairbrother, Ru-Fang Yeh, Phillip A. Sharp e Christopher B. Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583): 1007–1013. doi: 10.1126/science.1073774. URL http://dx.doi.org/10.1126/science.1073774. Citado na pág. 4

Faustino e Cooper (2003) Nuno André Faustino e Thomas A Cooper. Pre-mrna splicing and human disease. *Genes & development*, 17(4):419–437. URL http://genesdev.cshlp.org/content/17/4/419.short. Citado na pág. 4

Feng *et al.* **(2012)** Huijuan Feng, Zhiyi Qin e Xuegong Zhang. Opportunities and methods for studying alternative splicing in cancer with rna-seq. *Cancer letters*. URL http://www.sciencedirect.com/science/article/pii/S030438351200657X. Citado na pág. 5, 6

Finn *et al.* **(2010)** Robert D Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund *et al.* The pfam protein families database. *Nucleic acids research*, 38(suppl 1):D211–D222. URL http://nar.oxfordjournals.org/content/38/suppl_1/D211.short. Citado na pág. 18, 32

Florea *et al.* **(1998)** Liliana Florea, George Hartzell, Zheng Zhang, Gerald M Rubin e Webb Miller. A computer program for aligning a cdna sequence with a genomic dna sequence. *Genome research*, 8(9):967–974. URL http://genome.cshlp.org/content/8/9/967.short. Citado na pág. 5, 13, 26, 35

Frischmeyer *et al.* **(2002)** Pamela A Frischmeyer, Ambro van Hoof, Kathryn O'Donnell, Anthony L Guerrerio, Roy Parker e Harry C Dietz. An mrna surveillance mechanism that eliminates transcripts lacking termination codons. *Science*, 295(5563):2258–2261. URL http://www.sciencemag.org/content/295/5563/2258.short. Citado na pág. 6

Galante *et al.* **(2007)** Pedro AF Galante, Daniel O Vidal, Jorge E de Souza, Anamaria A Camargo e Sandro J de Souza. Sense-antisense pairs in mammals: functional and evolutionary considerations. *Genome biology*, 8(3):R40. Citado na pág. 13, 26

Galante *et al.* **(2004)** Pedro Alexandre Favoretto Galante, Noboru Jo Sakabe, Natanja Kirschbaum-Slager e Sandro José de Souza. Detection and evaluation of intron retention events in the human transcriptome. *Rna*, 10(5):757–765. URL http://rnajournal.cshlp.org/content/10/5/757.short. Citado na pág. 1, 11, 14, 25, 26, 28, 29, 30, 32, 35

Galiana-Arnoux *et al.* **(2005)** Delphine Galiana-Arnoux, Fabienne Del Gatto-Konczak, Marie-Claude Gesnel e Richard Breathnach. Intronic ugg repeats coordinate splicing of cd44 alternative exons v8 and v9. *Biochemical and biophysical research communications*, 336(2):667–673. URL http://www.sciencedirect.com/science/article/pii/S0006291X0501870X. Citado na pág. 11

Garcia-Blanco *et al.* **(2004)** Mariano A Garcia-Blanco, Andrew P Baraniak e Erika L Lasda. Alternative splicing in disease and therapy. *Nature biotechnology*, 22(5):535–546. URL http://www.nature.com/nbt/journal/v22/n5/abs/nbt964.html. Citado na pág. 4, 11, 25

Geiger *et al.* **(2012)** Tamar Geiger, Anja Wehner, Christoph Schaab, Juergen Cox e Matthias Mann. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & Cellular Proteomics*, 11(3). URL http://www.mcponline.org/content/11/3/M111.014050.short. Citado na pág. 27

Ghigna *et al.* **(1998)** Claudia Ghigna, Mauro Moroni, Camillo Porta, Silvano Riva e Giuseppe Biamonti. Altered expression of heterogeneous nuclear ribonucleoproteins and sr factors in human colon adenocarcinomas. *Cancer research*, 58(24):5818–5824. URL http://cancerres.aacrjournals.org/content/58/24/5818.short. Citado na pág. 7

Graveley (2000) B. R. Graveley. Sorting out the complexity of sr protein functions. *RNA*, 6(9):1197–1211. Citado na pág. 4

Graveley (2001) Brenton R Graveley. Alternative splicing: increasing diversity in the proteomic world. *TRENDS in Genetics*, 17(2):100–107. URL http://www.sciencedirect.com/science/article/pii/S0168952500021764. Citado na pág. 3

Guttman *et al.* **(2010)** Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum *et al.*

Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, 28(5):503–510. URL http://www.nature.com/nbt/journal/v28/n5/abs/nbt.1633.html. Citado na pág. 6

Halin *et al.* **(2001)** Cornelia Halin, Luciano Zardi e Dario Neri. Antibody-based targeting of angiogenesis. *Physiology*, 16(4):191–194. URL http://physiologyonline.physiology.org/content/16/4/191.short. Citado na pág. 8, 9

Hartmuth *et al.* **(2002)** Klaus Hartmuth, Henning Urlaub, Hans-Peter Vornlocher, Cindy L. Will, Marc Gentzel, Matthias Wilm e Reinhard LÃ$\frac{1}{4}$hrmann. Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. *Proc Natl Acad Sci U S A*, 99(26): 16719–16724. doi: 10.1073/pnas.262483899. URL http://dx.doi.org/10.1073/pnas.262483899. Citado na pág. 3

Hayes *et al.* **(2004)** Gregory M Hayes, Shona T Dougherty, Peter D Davis e Graeme J Dougherty. Molecular mechanisms regulating the tumor-targeting potential of splice-activated gene expression. *Cancer gene therapy*, 11(12):797–807. URL http://www.nature.com/cgt/journal/v11/n12/abs/7700759a.html. Citado na pág. 11

He *et al.* **(2007)** X He, M Pool, KM Darcy, SB Lim, N Auersperg, JS Coon e WT Beck. Knockdown of polypyrimidine tract-binding protein suppresses ovarian tumor cell growth and invasiveness in vitro. *Oncogene*, 26(34):4961–4968. URL http://www.nature.com/onc/journal/vaop/ncurrent/full/1210307a.html. Citado na pág. 7

Hellmich *et al.* **(2000)** Mark R Hellmich, Xian-Liang Rui, Helen L Hellmich, RY Declan Fleming, B Mark Evers e Courtney M Townsend. Human colorectal cancers express a constitutively active cholecystokinin-b/gastrin receptor that stimulates cell growth. *Journal of Biological Chemistry*, 275(41):32122–32128. URL http://www.jbc.org/content/275/41/32122.short. Citado na pág. 9

Hiller *et al.* **(2005)** Michael Hiller, Klaus Huse, Matthias Platzer e Rolf Backofen. Non-est based prediction of exon skipping and intron retention events using pfam information. *Nucleic acids research*, 33(17):5611–5621. URL http://nar.oxfordjournals.org/content/33/17/5611.abstract. Citado na pág. 32, 35

Holmila *et al.* **(2003)** R Holmila, C Fouquet, J Cadranel, G Zalcman e T Soussi. Splice mutations in the p53 gene: case report and review of the literature. *Human mutation*, 21(1):101–102. URL http://onlinelibrary.wiley.com/doi/10.1002/humu.9104/full. Citado na pág. 7

Hsu e Hertel (2009) Shu-Ning Hsu e Klemens J Hertel. Spliceosomes walk the line: Splicing errors and their impact on cellular function. *RNA biology*, 6(5):526–530. URL http://www.landesbioscience.com/journals/rnabiology/HsuRNA6-5.pdf. Citado na pág. 4

Huang *et al.* **(2003)** Ruimin Huang, Zhigang Xing, Zhidong Luan, Tangming Wu, Xin Wu e Gengxi Hu. A specific splicing variant of svh, a novel human armadillo repeat protein, is up-regulated in hepatocellular carcinomas. *Cancer research*, 63(13):3775–3782. URL http://cancerres.aacrjournals.org/content/63/13/3775.short. Citado na pág. 8

Iczkowski *et al.* **(2003)** Kenneth A Iczkowski, SHAN Bai, Cooley G Pantazis *et al.* Prostate cancer overexpresses cd44 variants 7-9 at the messenger rna and protein level. *Anticancer research*, 23 (4):3129. URL http://europepmc.org/abstract/MED/12926045. Citado na pág. 9

Ip *et al.* **(2011)** Joanna Y Ip, Dominic Schmidt, Qun Pan, Arun K Ramani, Andrew G Fraser, Duncan T Odom e Benjamin J Blencowe. Global impact of rna polymerase ii elongation inhibition on alternative splicing regulation. *Genome research*, 21(3):390–401. URL http://genome.cshlp.org/content/21/3/390.short. Citado na pág. 11

Jang (2002) Jun-Hyeog Jang. Identification and characterization of soluble isoform of fibroblast growth factor receptor 3 in human saos-2 osteosarcoma cells. *Biochemical and biophysical research communications*, 292(2):378–382. URL http://www.sciencedirect.com/science/article/pii/S0006291X02966684. Citado na pág. 9

Jang *et al.* **(2001)** Jun-Hyeog Jang, Ki-Hyuk Shin e Jae-Gahb Park. Mutations in fibroblast growth factor receptor 2 and fibroblast growth factor receptor 3 genes associated with human gastric and colorectal cancers. *Cancer research*, 61(9):3541–3543. URL http://cancerres.aacrjournals.org/content/61/9/3541.short. Citado na pág. 9

Jensen *et al.* **(2009)** CathyJ Jensen, BrianJ Oldfield e JustinP Rubio. Splicing, cis genetic variation and disease. *Biochemical Society Transactions*, 37(6):1311. Citado na pág. 6

Johnson *et al.* **(2003)** Jason M Johnson, John Castle, Philip Garrett-Engele, Zhengyan Kan, Patrick M Loerch, Christopher D Armour, Ralph Santos, Eric E Schadt, Roland Stoughton e Daniel D Shoemaker. Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. *Science*, 302(5653):2141–2144. URL http://www.sciencemag.org/content/302/5653/2141.short. Citado na pág. 4

Johnson *et al.* **(2005)** Richard S Johnson, Michael T Davis, J Alex Taylor e Scott D Patterson. Informatics for protein identification by mass spectrometry. *Methods*, 35(3):223–236. URL http://www.sciencedirect.com/science/article/pii/S104620230400204X. Citado na pág. 30

Jurica e Moore (2003) Melissa S Jurica e Melissa J Moore. Pre-mrna splicing: awash in a sea of proteins. *Molecular cell*, 12(1):5–14. URL http://www.sciencedirect.com/science/article/pii/S1097276503002703. Citado na pág. 3

Kan *et al.* **(2002)** Zhengyan Kan, Warren Gish *et al.* Selecting for functional alternative splices in ests. *Genome research*, 12(12):1837–1845. URL http://genome.cshlp.org/content/12/12/1837.short. Citado na pág. 35

Kelso *et al.* **(2003)** Janet Kelso, Johann Visagie, Gregory Theiler, Alan Christoffels, Soraya Bardien, Damian Smedley, Darren Otgaar, Gary Greyling, C Victor Jongeneel, Mark I McCarthy *et al.* evoc: a controlled vocabulary for unifying gene expression data. *Genome research*, 13(6a):1222–1230. URL http://genome.cshlp.org/content/13/6a/1222.short. Citado na pág. 18, 35

Kent (2002a) W James Kent. Blat – the blast-like alignment tool. *Genome research*, 12(4):656–664. URL http://genome.cshlp.org/content/12/4/656.short. Citado na pág. 5

Kent (2002b) W James Kent. Blat: the blast-like alignment tool. *Genome research*, 12(4):656–664. URL http://genome.cshlp.org/content/12/4/656.short. Citado na pág. 13, 26

Kerbel (2000) Robert S Kerbel. Tumor angiogenesis: past, present and the near future. *Carcinogenesis*, 21(3):505–515. URL http://carcin.oxfordjournals.org/content/21/3/505.short. Citado na pág. 9

Kersey *et al.* **(2004)** Paul J Kersey, Jorge Duarte, Allyson Williams, Youla Karavidopoulou, Ewan Birney e Rolf Apweiler. The international protein index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985–1988. URL http://onlinelibrary.wiley.com/doi/10.1002/pmic.200300721/full. Citado na pág. 27

Khan *et al.* **(2012)** Dilshad H Khan, Sanzida Jahan e James R Davie. Pre-mrna splicing: Role of epigenetics and implications in disease. *Advances in Biological Regulation*. URL http://www.sciencedirect.com/science/article/pii/S2212492612000577. Citado na pág. 7

Kim *et al.* **(2007)** Eddo Kim, Alon Magen e Gil Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, 35(1):125–131. URL http://nar.oxfordjournals.org/content/35/1/125.short. Citado na pág. 4

Kirschbaum-Slager *et al.* **(2005)** Natanja Kirschbaum-Slager, Raphael Bessa Parmigiani, Anamaria Aranha Camargo e Sandro José de Souza. Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. *Physiological genomics*, 21(3): 423–432. URL http://physiolgenomics.physiology.org/content/21/3/423.short. Citado na pág. 4, 14, 25

Krawczak *et al.* **(1992)** Michael Krawczak, Jochen Reiss e David N Cooper. The mutational spectrum of single base-pair substitutions in mrna splice junctions of human genes: causes and consequences. *Human genetics*, 90(1-2):41–54. URL http://link.springer.com/article/10.1007/BF00210743. Citado na pág. 6

Krawczak *et al.* **(2007)** Michael Krawczak, Nick ST Thomas, Bernd Hundrieser, Matthew Mort, Michael Wittig, Jochen Hampe e David N Cooper. Single base-pair substitutions in exon–intron junctions of human genes: nature, distribution, and consequences for mrna splicing. *Human mutation*, 28(2):150–158. URL http://onlinelibrary.wiley.com/doi/10.1002/humu.20400/full. Citado na pág. 6

Kroll *et al.* **(2012)** José Eduardo Kroll, Pedro A F. Galante, Daniel T. Ohara, FÃ¡bio C P. Navarro, Lucila Ohno-Machado e Sandro J. de Souza. Splooce: a new portal for the analysis of human splicing variants. *RNA Biol*, 9(11):1339–1343. doi: 10.4161/rna.22182. URL http://dx.doi.org/10.4161/rna.22182. Citado na pág. 5, 12, 25, 28, 30

Kurahashi *et al.* **(1995)** Hiroki Kurahashi, Koji Takami, Takaharu Oue, Takeshi Kusafuka, Akira Okada, Akio Tawa, Shintaro Okada e Isamu Nishisho. Biallelic inactivation of the apc gene in hepatoblastoma. *Cancer research*, 55(21):5007–5011. URL http://cancerres.aacrjournals.org/content/55/21/5007.short. Citado na pág. 7

Kwabi-Addo *et al.* **(2001)** Bernard Kwabi-Addo, Frederic Ropiquet, Dipak Giri e Michael Ittmann. Alternative splicing of fibroblast growth factor receptors in human prostate cancer. *The Prostate*, 46(2):163–172. Citado na pág. 9

Lander *et al.* **(2001)** Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. URL http://www.nature.com/nature/journal/v409/n6822/abs/409860a0.html. Citado na pág. 29

Lee *et al.* **(2000)** David W Lee, Kejian Zhang, Zhi-Qiang Ning, Eric H Raabe, Suzanne Tintner, Regina Wieland, Benjamin J Wilkins, Julia M Kim, Ruthann I Blough e Robert J Arceci. Proliferation-associated snf2-like gene (pasg): A snf2 family member altered in leukemia1. *Cancer research*, 60(13):3612–3622. URL http://cancerres.aacrjournals.org/content/60/13/3612.short. Citado na pág. 8

Lee *et al.* **(2008)** Hwa Jin Lee, Brian Wall e Suzie Chen. G-protein-coupled receptors and melanoma. *Pigment cell & melanoma research*, 21(4):415–428. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1755-148X.2008.00478.x/full. Citado na pág. 9

Letunic *et al.* **(2002)** Ivica Letunic, Richard R Copley e Peer Bork. Common exon duplication in animals and its role in alternative splicing. *Human molecular genetics*, 11(13):1561–1567. URL http://hmg.oxfordjournals.org/content/11/13/1561.short. Citado na pág. 5

Levin *et al.* **(2009)** Joshua Z Levin, Michael F Berger, Xian Adiconis, Peter Rogov, Alexandre Melnikov, Timothy Fennell, Chad Nusbaum, Levi A Garraway, Andreas Gnirke *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol*, 10(10):R115. Citado na pág. 6

Li *et al.* **(2006)** Hai-Ri Li, Jessica Wang-Rodriguez, T Murlidharan Nair, Joanne M Yeakley, Young-Soo Kwon, Marina Bibikova, Christina Zheng, Lixin Zhou, Kui Zhang, Tracy Downs *et al.* Two-dimensional transcriptome profiling: identification of messenger rna isoform signatures in prostate cancer from archived paraffin-embedded cancer specimens. *Cancer Research*, 66(8): 4079–4088. URL http://cancerres.aacrjournals.org/content/66/8/4079.short. Citado na pág. 5

Lim e Burge (2001) Lee P Lim e Christopher B Burge. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences*, 98(20):11193–11198. URL http://www.pnas.org/content/98/20/11193.short. Citado na pág. 3

Liu *et al.* **(2012)** Jinfeng Liu, William Lee, Zhaoshi Jiang, Zhongqiang Chen, Suchit Jhunjhunwala, Peter M Haverty, Florian Gnad, Yinghui Guan, Houston N Gilbert, Jeremy Stinson *et al.* Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. *Genome research*, 22(12):2315–2327. URL http://genome.cshlp.org/content/22/12/2315.short. Citado na pág. 7

Luco *et al.* **(2010)** Reini F Luco, Qun Pan, Kaoru Tominaga, Benjamin J Blencowe, Olivia M Pereira-Smith e Tom Misteli. Regulation of alternative splicing by histone modifications. *Science Signaling*, 327(5968):996. URL http://stke.sciencemag.org/cgi/content/abstract/sci;327/5968/996. Citado na pág. 11

Luther *et al.* **(2003)** Thomas Luther, Matthias Kotzsch, Axel Meye, Thomaz Langerholc, Susanne Füssel, Natalie Olbricht, Sybille Albrecht, Detlev Ockert, Bernd Muehlenweg, Katrin Friedrich *et al.* Identification of a novel urokinase receptor splice variant and its prognostic relevance in breast cancer. *Thrombosis and haemostasis*, 89(4):705–717. URL http://cat.inist.fr/?aModele=afficheN&cpsidt=14680908. Citado na pág. 8

Maas *et al.* **(2001)** Stefan Maas, Stephan Patt, Michael Schrey e Alexander Rich. Underediting of glutamate receptor glur-b mrna in malignant gliomas. *Proceedings of the National Academy of Sciences*, 98(25):14687–14692. URL http://www.pnas.org/content/98/25/14687.short. Citado na pág. 7

Malko *et al.* **(2006)** Dmitry B Malko, Vsevolod J Makeev, Andrey A Mironov e Mikhail S Gelfand. Evolution of exon–intron structure and alternative splicing in fruit flies and malarial mosquito genomes. *Genome research*, 16(4):505–509. URL http://genome.cshlp.org/content/16/4/505.short. Citado na pág. 11, 24

Malone e Oliver (2011) John H Malone e Brian Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9(1):34. Citado na pág. 5

Maniatis e Tasic (2002) Tom Maniatis e Bosiljka Tasic. Alternative pre-mrna splicing and proteome expansion in metazoans. *Nature*, 418(6894):236–243. URL http://www.nature.com/nature/journal/v418/n6894/abs/418236a.html. Citado na pág. 4

Markovic e Challiss (2009) Danijela Markovic e RA John Challiss. Alternative splicing of g protein-coupled receptors: physiology and pathophysiology. *Cellular and molecular life sciences*, 66(20):3337–3352. URL http://link.springer.com/article/10.1007/s00018-009-0093-4. Citado na pág. 9

Markovic e Grammatopoulos (2009) Danijela Markovic e Dimitris K Grammatopoulos. Focus on the splicing of secretin gpcrs transmembrane-domain 7. *Trends in biochemical sciences*, 34 (9):443–452. URL http://www.sciencedirect.com/science/article/pii/S0968000409001352. Citado na pág. 4

Martin e Wang (2011) Jeffrey A Martin e Zhong Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10):671–682. URL http://www.nature.com/nrg/journal/vaop/ncurrent/full/nrg3068.html. Citado na pág. 3

McGlincy e Smith (2008) Nicholas J McGlincy e Christopher WJ Smith. Alternative splicing resulting in nonsense-mediated mrna decay: what is the meaning of nonsense? *Trends in biochemical sciences*, 33(8):385–393. URL http://www.sciencedirect.com/science/article/pii/S0968000408001436. Citado na pág. 6

Modrek e Lee (2002) Barmak Modrek e Christopher Lee. A genomic view of alternative splicing. *Nature genetics*, 30(1):13–19. URL http://www.nature.com/ng/journal/v30/n1/abs/ng0102-13.html. Citado na pág. 5, 11, 25, 26, 30

Mukherji *et al.* **(2006)** Mridul Mukherji, Laurence M Brill, Scott B Ficarro, Garret M Hampton e Peter G Schultz. A phosphoproteomic analysis of the erbb2 receptor tyrosine kinase signaling pathways. *Biochemistry*, 45(51):15529–15540. URL http://pubs.acs.org/doi/abs/10.1021/bi060971c. Citado na pág. 7

Muñoz *et al.* **(2009)** Manuel J Muñoz, M Santangelo, Maria P Paronetto, Manuel de la Mata, Federico Pelisch, Stéphanie Boireau, Kira Glover-Cutter, Claudia Ben-Dov, Matías Blaustein, Juan J Lozano *et al.* Dna damage regulates alternative splicing through inhibition of rna polymerase ii elongation. *Cell*, 137(4):708–720. URL http://www.sciencedirect.com/science/article/pii/S0092867409002700. Citado na pág. 11

Mutz *et al.* **(2012)** Kai-Oliver Mutz, Alexandra Heilkenbrinker, Maren Lönne, Johanna-Gabriela Walter e Frank Stahl. Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology*. URL http://www.sciencedirect.com/science/article/pii/S0958166912001310. Citado na pág. 3

Nagasaki *et al.* **(2006)** Hideki Nagasaki, Masanori Arita, Tatsuya Nishizawa, Makiko Suwa e Osamu Gotoh. Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. *Bioinformatics*, 22(10):1211–1216. URL http://bioinformatics.oxfordjournals.org/content/22/10/1211.short. Citado na pág. 11, 16, 24

Naor *et al.* **(2002)** David Naor, Shlomo Nedvetzki, Itshak Golan, Lora Melnik e Yoram Faitelson. Cd44 in cancer. *Critical reviews in clinical laboratory sciences*, 39(6):527–579. URL http://informahealthcare.com/doi/abs/10.1080/10408360290795574. Citado na pág. 7

Narla *et al.* **(2005)** Goutham Narla, Analisa DiFeo, Helen L Reeves, Daniel J Schaid, Jennifer Hirshfeld, Eldad Hod, Amanda Katz, William B Isaacs, Scott Hebbring, Akira Komiya *et al.* A germline dna polymorphism enhances alternative splicing of the klf6 tumor suppressor gene and is associated with increased prostate cancer risk. *Cancer research*, 65(4):1213–1222. URL http://cancerres.aacrjournals.org/content/65/4/1213.short. Citado na pág. 7

Oliver e Marín (1996) José L Oliver e Antonio Marín. A relationship between gc content and coding-sequence length. *Journal of molecular evolution*, 43(3):216–223. URL http://link.springer.com/article/10.1007/BF02338829. Citado na pág. 29

Ozsolak e Milos (2010) Fatih Ozsolak e Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98. URL http://www.nature.com/nrg/journal/vaop/ncurrent/full/nrg2934.html. Citado na pág. 6

Pagani e Baralle (2004) Franco Pagani e Francisco E Baralle. Genomic variants in exons and introns: identifying the splicing spoilers. *Nature Reviews Genetics*, 5(5):389–396. URL http://www.nature.com/nrg/journal/v5/n5/abs/nrg1327.html. Citado na pág. 4

Pajares *et al.* **(2007)** María J Pajares, Teresa Ezponda, Raúl Catena, Alfonso Calvo, Ruben Pio e Luis M Montuenga. Alternative splicing: an emerging topic in molecular and clinical oncology. *The lancet oncology*, 8(4):349–357. URL http://www.sciencedirect.com/science/article/pii/S1470204507701043. Citado na pág. 6

Pan *et al.* **(2008)** Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey e Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415. doi: 10.1038/ng.259. URL http://dx.doi.org/10.1038/ng.259. Citado na pág. 11, 25

Passos *et al.* **(2009)** Dario O Passos, Meenakshi K Doma, Christopher J Shoemaker, Denise Muhlrad, Rachel Green, Jonathan Weissman, Julie Hollien e Roy Parker. Analysis of dom34 and its function in no-go decay. *Molecular biology of the cell*, 20(13):3025–3032. URL http://www.molbiolcell.org/content/20/13/3025.short. Citado na pág. 6

Pind e Watson (2003) Molly T Pind e Peter H Watson. Sr protein expression and cd44 splicing pattern in human breast tumours. *Breast cancer research and treatment*, 79(1):75–82. URL http://link.springer.com/article/10.1023/A%3A1023338718974. Citado na pág. 7

Pohl *et al.* **(2013)** Martin Pohl, Ralf H. Bortfeldt, Konrad Grutzmann e Stefan Schuster. Alternative splicing of mutually exclusive exons-a review. *Biosystems*. doi: 10.1016/j.biosystems.2013.07.003. URL http://dx.doi.org/10.1016/j.biosystems.2013.07.003. Citado na pág. 12

Power *et al.* **(2009)** Karen A Power, James P McRedmond, Andreas de Stefani, William M Gallagher e Peadar O Gaora. High-throughput proteomics detection of novel splice isoforms in human platelets. *PLoS One*, 4(3):e5001. Citado na pág. 30, 35

Rajan *et al.* **(2009)** Prabhakar Rajan, David J Elliott, Craig N Robson e Hing Y Leung. Alternative splicing and biological heterogeneity in prostate cancer. *Nature Reviews Urology*, 6(8):454–460. URL http://www.nature.com/nrurol/journal/v6/n8/abs/nrurol.2009.125.html. Citado na pág. 5

Reiter *et al.* **(2001)** Ronald Reiter, Anton Wellstein e Anna Tate Riegel. An isoform of the coactivator aib1 that increases hormone and growth factor sensitivity is overexpressed in breast cancer. *Journal of Biological Chemistry*, 276(43):39736–39741. URL http://www.jbc.org/content/276/43/39736.short. Citado na pág. 8

Roberts *et al.* **(2011)** Adam Roberts, Harold Pimentel, Cole Trapnell e Lior Pachter. Identification of novel transcripts in annotated genomes using rna-seq. *Bioinformatics*, 27(17):2325–2329. URL http://bioinformatics.oxfordjournals.org/content/27/17/2325.short. Citado na pág. 25, 28, 35

Roth *et al.* **(2005)** Michael J Roth, Andrew J Forbes, Michael T Boyne, Yong-Bin Kim, Dana E Robinson e Neil L Kelleher. Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry. *Molecular & Cellular Proteomics*, 4(7):1002–1008. URL http://www.mcponline.org/content/4/7/1002.short. Citado na pág. 30

Saint-André *et al.* **(2011)** Violaine Saint-André, Eric Batsché, Christophe Rachez e Christian Muchardt. Histone h3 lysine 9 trimethylation and hp1γ favor inclusion of alternative exons. *Nature structural & molecular biology*, 18(3):337–344. URL http://www.nature.com/nsmb/journal/v18/n3/abs/nsmb.1995.html. Citado na pág. 11

Saito *et al.* **(2002)** Yoshimasa Saito, Yae Kanai, Michiie Sakamoto, Hidetsugu Saito, Hiromasa Ishii e Setsuo Hirohashi. Overexpression of a splice variant of dna methyltransferase 3b, dnmt3b4, associated with dna hypomethylation on pericentromeric satellite regions during human hepatocarcinogenesis. *Proceedings of the National Academy of Sciences*, 99(15):10060–10065. URL http://www.pnas.org/content/99/15/10060.short. Citado na pág. 8

Sakabe e de Souza (2007) Noboru J Sakabe e Sandro J de Souza. Sequence features responsible for intron retention in human. *BMC genomics*, 8(1):59. Citado na pág. 11

Sakabe *et al.* **(2003)** Noboru Jo Sakabe, Jorge ES de Souza, Pedro AF Galante, Paulo SL de Oliveira, Fábio Passetti, Helena Brentani, Elisson C Osório, André C Zaiats, Maarten R Leerkes, João Paulo Kitajima *et al.* Orestes are enriched in rare exon usage variants affecting the encoded proteins. *Comptes Rendus Biologies*, 326(10):979–985. URL http://www.sciencedirect.com/science/article/pii/S1631069103002233. Citado na pág. 28

Sakurai *et al.* **(2013)** Naoto Sakurai, Shotaro Iwamoto, Yoshihiro Miura, Tomoki Nakamura, Akihiko Matsumine, Junji Nishioka, Kaname Nakatani e Yoshihiro Komada. Novel p53 splicing site mutation in li-fraumeni-like syndrome with osteosarcoma. *Pediatrics International*, 55(1): 107–111. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1442-200X.2012.03641.x/full. Citado na pág. 7

Sammeth *et al.* **(2008)** Michael Sammeth, Sylvain Foissac e Roderic Guigó. A general definition and nomenclature for alternative splicing events. *PLoS computational biology*, 4(8):e1000147. Citado na pág. 11, 24

Sanford *et al.* **(2005)** J. R. Sanford, J. Ellis e J. F. CÃ¡ceres. Multiple roles of arginine/serine-rich splicing factors in rna processing. *Biochem Soc Trans*, 33(Pt 3):443–446. doi: 10.1042/BST0330443. URL http://dx.doi.org/10.1042/BST0330443. Citado na pág. 4

Scheurlen e Senf (1995) Wolfram G Scheurlen e Leonore Senf. Analysis of the gap-related domain of the neurofibromatosis type 1 (nf1) gene in childhood brain tumors. *International journal of cancer*, 64(4):234–238. URL http://onlinelibrary.wiley.com/doi/10.1002/ijc.2910640404/abstract. Citado na pág. 8

Schmucker *et al.* **(2000)** Dietmar Schmucker, James C Clemens, Huidy Shu, Carolyn A Worby, Jian Xiao, Marco Muda, Jack E Dixon e S Lawrence Zipursky. < i> drosophila</i> dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6):671–684. URL http://www.sciencedirect.com/science/article/pii/S0092867400808788. Citado na pág. 5, 11

Schor *et al.* **(2009)** Ignacio E Schor, Nicolás Rascovan, Federico Pelisch, Mariano Alló e Alberto R Kornblihtt. Neuronal cell depolarization induces intragenic chromatin modifications affecting ncam alternative splicing. *Proceedings of the National Academy of Sciences*, 106(11):4325–4330. Citado na pág. 11

Shibayama *et al.* **(2009)** Masaki Shibayama, Satona Ohno, Takashi Osaka, Reiko Sakamoto, Akinori Tokunaga, Yuhki Nakatake, Mitsuharu Sato e Nobuaki Yoshida. Polypyrimidine tract-binding protein is essential for early mouse development and embryonic stem cell proliferation. *FEBS Journal*, 276(22):6658–6668. Citado na pág. 7

Singh e Valcárcel (2005) Ravinder Singh e Juan Valcárcel. Building specificity with nonspecific rna-binding proteins. *Nat Struct Mol Biol*, 12(8):645–653. doi: 10.1038/nsmb961. URL http://dx.doi.org/10.1038/nsmb961. Citado na pág. 4

Sorek e Ast (2003) Rotem Sorek e Gil Ast. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Research*, 13(7):1631–1637. URL http://genome.cshlp.org/content/13/7/1631.short. Citado na pág. 35

Srebrow e Kornblihtt (2006) Anabella Srebrow e Alberto R Kornblihtt. The connection between splicing and cancer. *Journal of Cell Science*, 119(13):2635–2641. URL http://jcs.biologists.org/content/119/13/2635.short. Citado na pág. 6, 9

Stickeler *et al.* **(1999)** Elmar Stickeler, Frances Kittrell, Daniel Medina, Susan M Berget *et al.* Stage-specific changes in sr splicing factors and alternative splicing in mammary tumorigenesis. *Oncogene*, 18(24):3574. URL http://europepmc.org/abstract/MED/10380879. Citado na pág. 7

Sugnet *et al.* **(2004)** C. W. Sugnet, W. J. Kent, M Ares, Jr e D. Haussler. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput*, páginas 66–77. Citado na pág. 4

Sultan *et al.* **(2008)** Marc Sultan, Marcel H Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960. URL http://www.sciencemag.org/content/321/5891/956.short. Citado na pág. 6

Tacconelli *et al.* **(2004)** Antonella Tacconelli, Antonietta R Farina, Lucia Cappabianca, Giuseppina DeSantis, Alessandra Tessitore, Antonella Vetuschi, Roberta Sferra, Nadia Rucci, Beatrice Argenti, Isabella Screpanti *et al.* Trka alternative splicing: a regulated tumor-promoting switch in human neuroblastoma. *Cancer cell*, 6(4):347–360. URL http://www.sciencedirect.com/science/article/pii/S1535610804002715. Citado na pág. 11

Tacke e Manley (1999) Roland Tacke e James L Manley. Functions of sr and tra2 proteins in pre-mrna splicing regulation. Em *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, NY)*, volume 220, páginas 59–63. Royal Society of Medicine. URL http://ebm.rsmjournals.com/content/220/2/59.short. Citado na pág. 7

Tan *et al.* **(2007)** Sheng Tan, Jiaming Guo, Qianli Huang, Xueping Chen, Jesse Li-Ling, Qingwei Li e Fei Ma. Retained introns increase putative microrna targets within 3? utrs of human mrna. *FEBS letters*, 581(6):1081–1086. Citado na pág. 35

Tanaka *et al.* **(2001)** Shinji Tanaka, Keishi Sugimachi, Hiroshi Saeki, Junko Kinoshita, Takefumi Ohga, Mitsuo Shimada, Yoshihiko Maehara e Keizo Sugimachi. A novel variant of wisp1 lacking a von willebrand type c module overexpressed in scirrhous gastric carcinoma. *Oncogene*, 20(39): 5525–5532. URL http://cat.inist.fr/?aModele=afficheN&cpsidt=14154440. Citado na pág. 8

Tang *et al.* **(2013)** Jen-Yang Tang, Jin-Ching Lee, Ming-Feng Hou, Chun-Lin Wang, Chien-Chi Chen, Hurng-Wern Huang e Hsueh-Wei Chang. Alternative splicing for diseases, cancers, drugs, and databases. *The Scientific World Journal*, 2013. URL http://www.hindawi.com/journals/tswj/2013/703568/abs/. Citado na pág. 12

Tanko *et al.* **(2002)** Q Tanko, B Franklin, H Lynch e J Knezetic. A *hMLH1* genomic mutation and associated novel mrna defects in a hereditary non-polyposis colorectal cancer family. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 503(1):37–42. URL http://www.sciencedirect.com/science/article/pii/S0027510702000313. Citado na pág. 7, 11

Taylor *et al.* **(2000)** Michael D Taylor, Nalan Gokgoz, Irene L Andrulis, Todd G Mainprize, James M Drake e James T Rutka. Familial posterior fossa brain tumors of infancy secondary to germline mutation of the< i> hsnf5</i> gene. *The American Journal of Human Genetics*, 66 (4):1403–1406. Citado na pág. 7

Trapnell *et al.* **(2009)** Cole Trapnell, Lior Pachter e Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111. URL http://bioinformatics.oxfordjournals.org/content/25/9/1105.short. Citado na pág. 6, 13, 25, 26, 28, 35

Trapnell *et al.* **(2010)** Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold e Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515. URL http://www.nature.com/nbt/journal/v28/n5/abs/nbt.1621.html. Citado na pág. 6, 13, 27

Venables (2002) Julian P Venables. Alternative splicing in the testes. *Current opinion in genetics & development*, 12(5):615–619. URL http://www.sciencedirect.com/science/article/pii/S0959437X02003477. Citado na pág. 4

Venables (2004) Julian P Venables. Aberrant and alternative splicing in cancer. *Cancer research*, 64(21):7647–7654. URL http://cancerres.aacrjournals.org/content/64/21/7647.short. Citado na pág. 4, 6, 9, 11, 25, 35

Venables (2006) Julian P Venables. Unbalanced alternative splicing and its significance in cancer. *Bioessays*, 28(4):378–386. URL http://onlinelibrary.wiley.com/doi/10.1002/bies.20390/full. Citado na pág. 5, 6, 11

Vickers *et al.* **(2002)** Selwyn M Vickers, Zhi-Qiang Huang, LeeAnn MacMillan-Crow, Jessica S Greendorfer e John A Thompson. Ligand activation of alternatively spliced fibroblast growth factor receptor-1 modulates pancreatic adenocarcinoma cell malignancy. *Journal of gastrointestinal surgery*, 6(4):546–553. URL http://link.springer.com/article/10.1016/S1091-255X(02)00036-7. Citado na pág. 9

Vogel *et al.* **(2010)** Christine Vogel, Raquel de Sousa Abreu, Daijin Ko, Shu-Yun Le, Bruce A Shapiro, Suzanne C Burns, Devraj Sandhu, Daniel R Boutz, Edward M Marcotte e Luiz O Penalva. Sequence signatures and mrna concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular systems biology*, 6(1). URL http://www.nature.com/msb/journal/v6/n1/synopsis/msb201059.html. Citado na pág. 30

Walsh (2010) Gary Walsh. Biopharmaceutical benchmarks 2010. *Nature biotechnology*, 28(9):917. Citado na pág. 8

Wang *et al.* **(2008)** Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth e Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476. doi: 10.1038/nature07509. URL http://dx.doi.org/10.1038/nature07509. Citado na pág. 5, 11, 25, 26, 30

Wang e Cooper (2007) Guey-Shin Wang e Thomas A Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, 8(10):749–761. URL http://www.nature.com/nrg/journal/v8/n10/abs/nrg2164.html. Citado na pág. 4, 6, 25

Wang *et al.* **(2003)** Lei Wang, Lindsay Duke, Peter S Zhang, Ralph B Arlinghaus, W Fraser Symmans, Aysegul Sahin, Richard Mendez e Jia Le Dai. Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. *Cancer research*, 63(15):4724–4730. URL http://cancerres.aacrjournals.org/content/63/15/4724.short. Citado na pág. 8

Wang *et al.* **(2010)** Liguo Wang, Yuanxin Xi, Jun Yu, Liping Dong, Laising Yen e Wei Li. A statistical method for the detection of alternative splicing using rna-seq. *PloS one*, 5(1):e8529. Citado na pág. 6

Ward e Cooper (2010) Amanda J Ward e Thomas A Cooper. The pathobiology of splicing. *The Journal of pathology*, 220(2):152–163. URL http://onlinelibrary.wiley.com/doi/10.1002/path.2649/full. Citado na pág. 6, 7

Weiss *et al.* **(1996)** Mordechai Weiss, Amos Baruch, Iafa Keydar e Daniel H Wreschner. Preoperative diagnosis of thyroid papillary carcinoma by reverse transcriptase polymerase chain reaction of the muc1 gene. *International journal of cancer*, 66(1):55–59. Citado na pág. 8

Wheeler *et al.* **(2007)** David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen *et al.* Database resources of the national center for biotechnology information. *Nucleic*

acids research, 35(suppl 1):D5–D12. URL http://nar.oxfordjournals.org/content/35/suppl_1/D5.short. Citado na pág. 25, 28, 35

Wis niewski *et al.* **(2009)** Jacek R Wis niewski, Alexandre Zougman e Matthias Mann. Combination of fasp and stagetip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *Journal of proteome research*, 8(12):5674–5678. URL http://pubs.acs.org/doi/abs/10.1021/pr900748n. Citado na pág. 27

Woodley e Valcárcel (2002) Louise Woodley e Juan Valcárcel. Regulation of alternative pre-mrna splicing. *Briefings in Functional Genomics & Proteomics*, 1(3):266–277. URL http://bfg.oxfordjournals.org/content/1/3/266.short. Citado na pág. 4

Yeh *et al.* **(2003)** Brian K Yeh, Makoto Igarashi, Anna V Eliseenkova, Alexander N Plotnikov, Ifat Sher, Dina Ron, Stuart A Aaronson e Moosa Mohammadi. Structural basis by which alternative splicing confers specificity in fibroblast growth factor receptors. *Proceedings of the National Academy of Sciences*, 100(5):2266–2271. URL http://www.pnas.org/content/100/5/2266.short. Citado na pág. 9

Zhang *et al.* **(2007)** Chaolin Zhang, Michelle L Hastings, Adrian R Krainer e Michael Q Zhang. Dual-specificity splice sites function alternatively as 5 and 3 splice sites. *Proceedings of the National Academy of Sciences*, 104(38):15028–15033. URL http://www.pnas.org/content/104/38/15028.short. Citado na pág. 4, 16, 25

Zhang *et al.* **(2012)** Xiang-Zhong Zhang, Ai-Hua Yin, XY Zhu, QIAN Ding, Chun-Huai Wang, Yun-Xian Chen *et al.* Using an exon microarray to identify a global profile of gene expression and alternative splicing in k562 cells exposed to sodium valproate. *Oncology reports*, 27(4):1258–1265. Citado na pág. 5

Zhang *et al.* **(2013)** ZhongFa Zhang, Sharmistha Pal, Yingtao Bi, Julia Tchou e Ramana V Davuluri. Isoform-level expression profiles provide better cancer signatures than gene-level expression profiles. *Genome medicine*, 5(4):33. URL http://genomemedicine.com/content/5/4/33/abstract. Citado na pág. 5

Zhou *et al.* **(2002)** Zhaolan Zhou, Lawrence J Licklider, Steven P Gygi e Robin Reed. Comprehensive proteomic analysis of the human spliceosome. *Nature*, 419(6903):182–185. URL http://www.nature.com/nature/journal/v419/n6903/abs/nature01031.html. Citado na pág. 3

Zhu *et al.* **(1997)** Xiang Zhu, Angela AI Daffada, Christina MW Chan e Mitchell Dowsett. Identification of an exon 3 deletion splice variant androgen receptor mrna in human breast cancer. *International journal of cancer*, 72(4):574–580. Citado na pág. 8