Rodrigo Fernandes Ramalho

Análise dos sinais de seleção natural em
reguladores de *splicing* exônicos do genoma
humano

Investigating the signature of natural selection on
exonic splicing regulators of the human genome

São Paulo

2012

Rodrigo Fernandes Ramalho

Análise dos sinais de seleção natural em reguladores de
splicing exônicos do genoma humano

Investigating the signature of natural selection on exonic
splicing regulators of the human genome

Versão corrigida da tese apresentada
ao Instituto de Biociências da
Universidade de São Paulo, para a
obtenção de Título de Doutor em
Ciências, na Área de Genética e
Biologia Evolutiva.

A versão original da tese encontra-se
disponível na Biblioteca do IB-USP
assim como na Biblioteca Digital de
Teses e Dissertações da USP.

Orientador: Prof. Dr. Diogo Meyer
Coorientador: Prof. Dr. Sandro J. de
Souza

São Paulo

2012

Ficha catalográfica

F. Ramalho, Rodrigo
Análise dos sinais de seleção natural
em reguladores de *splicing* exônicos do
genoma humano

141 páginas

Tese (Doutorado) - Instituto de
Biociências da Universidade de São Paulo.
Departamento de Genética e Biologia Evolutiva.

1. Seleção natural 2. Evolução
Molecular 3. Reguladores de *splicing*
Universidade de São Paulo. Instituto de
Biociências. Departamento de Genética e
Biologia Evolutiva.

Comissão Julgadora:

Prof. Dr. Sergio Russo Matioli

Prof. Dr. Sergio Roberto Peres Line

Prof. Dr. Wilson Araujo da Silva Junior

Profa. Dra. Carla Columbano de Oliveira

Prof. Dr. Diogo Meyer
Orientador

Dedicatória

À minha esposa, Roberta

Agradecimentos

- Ao meu orientador, Prof. Dr. Diogo Meyer, por sua imensa colaboração em mais um etapa da minha formação como cientista.
- Ao meu co-orientador, Dr. Sandro José de Souza, pela parceria fundamental em diversas etapas do projeto, desde a elaboração do projeto até a redação dos artigos.
- Ao Dr. Jorge Estefano S. de Souza, por me ensinar o básico de programação e me apresentar a diversas outras ferramentas da Bioinformática. Agradeço também pela sua parceria durante todo o projeto.
- Ao Dr. Pedro Galante, por disponibilizar dados fundamentais para a realização do projeto e colaborar diretamente em análises de dados.
- Ao Daniel T. Ohara, pelo valioso auxílio computacional.
- Ao Dr. Douglas Vasconcelos Cancherini, pelas orientações sobre lógica de programação.
- Ao Dr. Sahar Gelfman, da Universidade de Tel Aviv, pela grande colaboração em análises fundamentais para o andamento do projeto.
- Ao Dr. Gil Ast, da Universidade de Tel Aviv, por me receber em seu laboratório e fornecer opiniões de grande importância para a compreensão de aspectos evolutivos do *splicing*.
- À Secretaria do Instituto Ludwig de Pesquisa sobre o Câncer localizado no Hospital Alemão Oswaldo Cruz e do Instituto de Biociências da Universidade de São Paulo.
- Aos meus colegas de laboratório Rodrigo Francisco, Fábio, Maria Helena, Márcia, Kelly, Bárbara e Débora.
- À FAPESP, pela bolsa concedida (Proc. N° 2007/59721-8).
- Aos meus pais e irmãos pela inestimável ajuda.

Índice

CAPÍTULO 1 : ASPECTOS BIOQUÍMICOS E EVOLUTIVOS DO *SPLICING*

PRÓLOGO	1
RESUMO (PARTE A).....	2
RESUMO (PARTE B).....	3
PARTE A) <i>SPLICING</i> : ASPECTOS BIOQUÍMICOS	4
VISÃO GERAL DO MECANISMO DE <i>SPLICING</i>	4
SÍTIOS DE <i>SPLICING</i>	6
PRINCIPAIS FATORES DE <i>SPLICING</i>	8
FORÇA DO SÍTIO DE <i>SPLICING</i>	9
A ETAPA DE RECONHECIMENTO DE ÉXONS (<i>EXON DEFINITION</i>)	10
REGULADORES DE <i>SPLICING</i>	12
EXONIC <i>SPLICING</i> ENHANCERS (ESEs).....	14
IDENTIFICAÇÃO DOS SELEX-ESES.	14
IDENTIFICAÇÃO IN SILICO DE ESES	16
EXONIC <i>SPLICING</i> SILENCERS (ESS).....	16
IDENTIFICAÇÃO IN VITRO DOS ESSs.....	18
EVIDÊNCIAS DE FUNÇÃO ANTAGÔNICA E DEPENDENTE DE CONTEXTO DE CERTOS ESRs.....	18
PARTE B) <i>SPLICING</i> : ASPECTOS BIOLÓGICOS E EVOLUTIVOS	20
ORIGENS DO <i>SPLICING</i>	20
<i>SPLICING</i> ALTERNATIVO COMO UM MECANISMO DE GERAÇÃO DE DIVERSIDADE BIOLÓGICA	22
<i>SPLICING</i> ALTERNATIVO REGULADO E NÃO REGULADO	23
<i>SPLICING</i> ALTERNATIVO E DOENÇAS	24
CLASSIFICAÇÃO DOS ÉXONS ALTERNATIVOS.....	25
ORIGENS DO <i>SPLICING</i> ALTERNATIVO	27
EXONIZAÇÃO DE SEQUÊNCIAS INTRÔNICAS.....	27
DUPLICAÇÃO DE GENES E ÉXONS.....	28
RELAXAMENTO DE SELEÇÃO EM ÉXONS CONSTITUTIVOS	28
SELEÇÃO NATURAL EM REGULADORES DE <i>SPLICING</i> EXÔNICOS (ESRS).....	31
SELEÇÃO NEGATIVA	32
SELEÇÃO POSITIVA.....	36
EVOLUÇÃO DOS ESRs EM GENES PARÁLOGOS	37
SELEÇÃO NATURAL EM ÉXONS CONSTITUTIVOS E ALTERNATIVOS.....	38
TEMPO DE DIVERGÊNCIA INFLUENCIA NA DETECÇÃO DE SELEÇÃO NATURAL EM ESES.....	42
SOBREPOSIÇÃO DOS CÓDIGOS DE <i>SPLICING</i> E PROTEICO.....	43
BIBLIOGRAFIA	45

CAPÍTULO 2: SPLICING ALTERNATIVO E DIVERSIDADE GENÉTICA: SNVS ASSOCIADOS À VARIANTES DE SPLICING SÃO MAIS FREQUENTES EM INIBIDORES DE SPLICING

PRÓLOGO	51
RESUMO	53
ABSTRACT	55
INTRODUCTION	55
MATERIALS AND METHODS	58
RESULTS AND DISCUSSION	64
FINAL REMARKS.....	73
REFERENCES.....	74

CAPÍTULO 3: SOBRE AS ORIGENS DO ÉXON SKIPPING: POLIMORFISMOS FRACAMENTE DELETÉRIOS CRIAM INIBIDORES DE SPLICING EM ÉXONS CONSTITUTIVOS

PRÓLOGO	77
RESUMO	78
ABSTRACT	80
INTRODUCTION	81
MATERIALS AND METHODS	85
RESULTS AND DISCUSSION	90
CONCLUDING REMARKS	99
REFERENCES.....	101
SUPPLEMENTARY MATERIAL	103

CAPÍTULO 4: SELEÇÃO NATURAL EM EXONS HUMANOS COM DIFERENTES NÍVEIS DE INCLUSÃO

PRÓLOGO	108
RESUMO	110
ABSTRACT	113
INTRODUCTION	114
MATERIALS AND METHODS	117
RESULTS AND DISCUSSION	122
CONCLUSIONS	130
REFERENCES.....	130
SUPPLEMENTARY MATERIAL	133

CAPÍTULO 5: DISCUSSÃO GERAL E CONCLUSÕES

DISCUSSÃO GERAL E CONCLUSÕES	134
APÊNDICE 1.....	141

Resumo

O *splicing* é o processo que resulta na remoção dos íntrons e união dos éxons nos genes eucarióticos. Nesse processo diversos elementos em *cis* e *trans* estão envolvidos. Além das sequências em *cis* canônicas (sítios de *splicing*, sítio de ramificação e trato de polipirimidina), os reguladores de *splicing* -- sequências curtas localizadas em éxons e íntrons -- são considerados de grande importância, pois auxiliam na correta determinação das fronteiras éxons/íntrons. Através do *splicing* alternativo pode-se gerar grande diversidade de transcritos e atualmente sabe-se que no mínimo, 80% dos genes humanos apresentam variantes de *splicing*. A evolução dos Reguladores de *Splicing* Exônicos (ESRs) foi principalmente analisada através de abordagens filogenéticas. Embora esses estudos tenham revelado resultados consistentes (por exemplo, evidências de seleção negativa contra mudanças sinônimas que afetam ESRs) outros ainda parecem contraditórios, como a maior conservação filogenética e a maior taxa de evolução não-sinônima dos éxons alternativos em relação aos constitutivos. Nesta tese abordamos questões sobre o regime e a intensidade de seleção que atuam sobre os éxons do genoma humano e seus reguladores através da comparação da variação genética intra e inter-específica. Os resultados da tese demonstram que: 1. Há uma diferença na densidade e na intensidade de seleção negativa que atua sobre os ESRs de éxons constitutivos e alternativos. 2. Os inibidores de *splicing* tem papel principal na origem dos éxons alternativos a partir de éxons constitutivos e também nos casos de uso alternativo de sítios de *splicing*. 3. O nível de inclusão dos éxons está diretamente relacionado com a intensidade de seleção negativa sobre mudanças não-sinônimas.

Abstract

Splicing is the process by which introns are removed from a mRNA precursor and exons are ligated to form a mature mRNA. During this process several *cis* and *trans* factors are involved. Besides the canonical *cis* factors (e.g., splicing sites, branch point and polypyrimidine tract), Splicing Regulators – short sequences located in exons and introns – have an important role in assisting the spliceosome to correctly recognize exon/intron boundaries. Through alternative splicing, great transcript diversity is generated, and currently it is known that more than 80% of human genes present splicing variants. The evolution of Exonic Splicing Regulators (ESRs) has been mainly analyzed by interspecific comparisons. Although these studies have revealed consistent results (evidences of weak negative selection against synonymous variations that affect ESRs), other findings still appear inconsistent, for instance the reports on increased level of conservation and higher non-synonymous evolutionary rate in alternative than constitutive exons. The present thesis investigates the regime and intensity of natural selection on exons from the human genome and their ESRs by comparing intra and interspecific genetic variation. We demonstrate that 1. ESRs from constitutive and alternative exons differ significantly in density and the intensity of negative selection. 2. The exonic splicing silencers have a major role in the origin of exons skipping from constitutive exons, and also on events of alternative usage of splicing sites. 3. There is a positive correlation between the exon inclusion level and the intensity of negative selection against non-synonymous variations.

Capítulo 1

Aspectos Bioquímicos e Evolutivos do *Splicing*

Prólogo

Este capítulo tem como objetivo apresentar aspectos bioquímicos e evolutivos do processo de *splicing*. Não há dúvidas que ambos os aspectos estão relacionados, porém por uma questão didática, eles serão tratados separadamente.

Inicialmente (Parte A) irei revisar conceitos básicos do mecanismo molecular do *splicing* e apresentar os elementos (fatores em *cis* e *trans*) fundamentais para a viabilidade desse processo. O intuito desta secção é fornecer os conceitos fundamentais do mecanismo de *splicing* para a compreensão do conteúdo evolutivo presente na parte B, relevante para o restante da tese.

Na parte B apresento as principais hipóteses sobre a origem e os processos evolutivos relacionados ao *splicing*. Apresento uma revisão detalhada dos principais trabalhos sobre a evolução dos reguladores de *splicing*. Alguns temas de natureza mais ampla, relacionados à evolução do *splicing*, foram incluídos para facilitar a compreensão do tema central da tese. O intuito da parte B é fornecer a informação necessária para a compreensão dos demais capítulos da tese e apresentar hipóteses que serão testadas nesta tese.

***Splicing*: Aspectos Bioquímicos**

Resumo (parte A)

O *splicing* é uma etapa fundamental do processo de expressão dos genes eucarióticos e consiste na separação de duas regiões gênicas que apresentam diferenças marcantes, os éxons e os íntrons. Após o *splicing*, apenas os éxons seguem no processo de expressão gênica enquanto que os íntrons são descartados. A remoção dos íntrons do transcrito primário de um gene (pré-RNA_m), resulta no RNA mensageiro (RNA_m) ou RNA maduro. O processo de *splicing* é mediado pelo spliceossomo que é uma estrutura molecular composta por mais de 100 proteínas distintas e por RNAs conhecidos como snRNAs (*small nuclear RNAs*) que interagem de forma dinâmica determinando as diferentes etapas do processamento. Os componentes do spliceossomo também interagem com sequências de nucleotídeos específicas presentes nas fronteiras entre os éxons e os íntrons durante o processamento do pré-RNA_m. Essas sequências são conhecidas como sítios de *splicing* e reguladores de *splicing* (Hertel, 2008; House & Lynch, 2008).

O *splicing* é um mecanismo que demanda um alto nível de precisão e entender os mecanismos que permitem isso é um objetivo desafiador. Sabe-se que erros na escolha dos sítios de *splicing* podem resultar em um RNA_m não funcional. Além disso, outro desafio é compreender como um pré-RNA_m com dezenas de íntrons pode ser processado ordenadamente, sem omitir nenhum éxon no RNA maduro. Os sítios de *splicing* são importantes na determinação dessa precisão do *splicing*, porém não são suficientes. Sequências similares aos sítios de *splicing* podem ser encontradas ao longo do gene, porém não são

utilizadas como tal. Por isso, sequências auxiliares, conhecidas como *cis*-elementos ou reguladores de *splicing* são fundamentais para o correto processamento do pré-RNA, principalmente nos eucariotos superiores.

Splicing: Aspectos Evolutivos

Resumo (parte B)

A evolução do *splicing* é estudada através da evolução dos íntrons, éxons e dos fatores em *cis* e *trans* que participam deste processo. Devido a sua alta complexidade o *splicing* é estudado em diversos níveis de organização biológica que vão dos mais gerais, por exemplo, a origem da organização éxon/íntron dos genes eucarióticos, até os mais detalhados como, por exemplo, a evolução de um domínio de uma proteína fundamental para o *splicing*, que reconhece uma sequência específica no precursor do RNA.

O *splicing* alternativo representa um dos mecanismos de geração de variabilidade biológica. Por isso, seu estudo é de interesse nas áreas de evolução molecular e médica. Recentemente, os reguladores de *splicing* exônicos ganharam destaque na literatura evolutiva devido a sua sobreposição com os códons. Esses trabalhos chamaram a atenção para os efeitos prejudiciais de mutações sinônimas, geralmente consideradas neutras pela teoria evolutiva, quando localizadas nesses reguladores.

Parte A) *Splicing*: Aspectos Bioquímicos

Visão geral do mecanismo de *splicing*

As etapas do processamento do pré-RNA foram estabelecidas principalmente através de experimentos *in vitro* que usaram como substrato, moléculas de RNA contendo apenas dois éxons separados por um íntron e como meio de reação, o extrato de células inteiras ou apenas de seus núcleos. (Green, *et al.*, 1983).

O desenvolvimento de sistemas eficientes de *splicing in vitro* permitiu a caracterização dos produtos intermediários e finais do *splicing*. Nesses experimentos, dois produtos intermediários foram observados, sendo que ambos continham o íntron. Um desses produtos apresentava o íntron ligado ao segundo éxon e o outro produto era o próprio íntron excisado (Figura 1). Evidências posteriores demonstraram que esses produtos apresentavam uma estrutura em forma de laço. O trabalho de Wallace e Edmonds, (1983), foi fundamental nessa descoberta, pois identificou a presença desses laços apenas no RNA total proveniente do núcleo (local do *splicing*), e não do citoplasma de células HeLa. Além disso, eles também detalharam o tipo de ligação química que ocorre no nucleotídeo do íntron (geralmente contendo uma Adenina) onde ocorre a emenda que forma o laço (Figura 1). Dessa forma foi estabelecido que os produtos do processamento do pré-RNA que contém o íntron estão em forma de laço, e que este laço é formado através da ligação da extremidade 5' do íntron (geralmente uma Guanina) à uma Adenina localizada próxima à extremidade 3' do íntron através de uma ligação 2'-5' fosfodiéster (Figura 1).

Figura 1: Etapas básicas do *splicing* do pré-RNAm. Retângulos representam éxons, a linha representa um íntron. Estão indicados na figura, o dinucleotídeo GU no sítio de *splicing* 5', o AG no sítio de *splicing* 3' e o A no sítio de ramificação (*branch point*). Os fosfatos presentes nos sítios 5' e 3' estão representados por p_1 e p_2 respectivamente. Figura modificada de Green M., (1986).

Os resultados obtidos com os primeiros experimentos de *splicing in vitro* permitiram a elaboração de um modelo de duas etapas para o mecanismo bioquímico do *splicing*. Na primeira etapa, o pré-RNAm é clivado na extremidade 5' do íntron para gerar os seguintes produtos intermediários: Uma molécula de RNA linear composta apenas pelo primeiro éxon do pré-RNAm e uma molécula de RNA não linear, composta pelo íntron em forma de laço unido ao segundo éxon. Na segunda etapa do *splicing* ocorrem a clivagem na extremidade 3' do íntron e a ligação dos dois éxons, resultando na liberação do íntron ainda em forma de laço. Como o fosfato da ligação 2'-5' fosfodiéster que

forma o laço é proveniente da extremidade 5' do íntron, a extremidade 3' do primeiro éxon passa a apresentar um grupo hidroxil. Sendo assim, o fosfato da ligação fosfodiéster convencional (3'-5') que une as extremidades dos dois éxons na etapa final do *splicing* é proveniente do segundo éxon, ou seja, do éxon localizado à jusante do íntron (Figura 1).

Dentre os produtos intermediários da reação de *splicing*, a ausência de formas lineares de RNA contendo o íntron e o segundo éxon, indicaram que a clivagem do sítio 5' e a formação do laço poderiam ser simultâneas. Nesse caso, a clivagem da ligação fosfodiéster convencional (5'-3') na extremidade 5' do íntron aconteceria simultaneamente com a formação da ligação 2'-5' fosfodiéster que caracteriza a formação do laço, numa reação chamada de transesterificação. Similarmente, não foram observados casos em que o segundo éxon ocorre livre do íntron e desligado do primeiro éxon, sugerindo que a clivagem da extremidade 3' do íntron e a ligação dos dois éxons também acontecem simultaneamente através de uma reação de transesterificação.

Os primeiros experimentos de *splicing in vitro* revelaram a surpreendente precisão deste processo na identificação das regiões de fronteira entre éxons e íntrons, ou seja, as regiões onde ocorre a clivagem, também chamadas de sítios de *splicing*.

Sítios de *splicing*

Define-se sítio de *splicing* como a região onde ocorre a ruptura que delimita éxons e íntrons. Essa região inclui nucleotídeos da porção exônica e da intrônica. Os sítios de *splicing* localizados nas extremidades 5' e 3' do íntron são chamados de sítio de *splicing* 5' e 3' respectivamente. O sítio de *splicing* 5'

também é conhecido como sítio doador porque a guanina localizada nesse sítio fornece o fosfato para a ligação 2'-5' fosfodiéster que forma o laço em uma região do íntron chamada de região de ramificação (*branch point*). De acordo com essa nomenclatura o sítio 3' é conhecido como sítio receptor. Vale destacar que os sítios de *splicing* 5' e 3' são originalmente definidos em relação ao íntron. Por isso, em relação ao éxon, o sítio de *splicing* 3' está à montante e o sítio de *splicing* 5' está à jusante (Figura 1).

Um dos primeiros trabalhos que avaliou a presença de algum padrão na sequência de nucleotídeos localizada na fronteira éxon/íntron foi realizado por Breathnach *et al.*, (1978). Comparando a sequência genômica com sequência de cDNA do gene da albumina esses autores identificaram padrões recorrentes de sequências nas fronteiras éxon/íntron desse gene e propuseram duas sequências que representam consensos para os sítios de *splicing* 5' e 3'. Esses resultados representaram a primeira evidência para a hipótese que a sequência de nucleotídeos do pré-RNAm é um determinante principal da escolha da região da fronteira éxon/íntron. Posteriormente, Mount, (1982) compilou e comparou um conjunto de fronteiras éxon/íntron de diversas espécies, incluindo plantas, insetos e mamíferos revelando a existência de uma grande conservação filogenética apenas nos dois primeiros e últimos nucleotídeos dos íntrons. A observação de que os íntrons geralmente começam com GT e terminam com AG justificou a criação da chamada regra GT-AG. Entretanto, além dessas duas posições, pouca conservação filogenética foi observada. Sendo assim, foram determinadas novas sequências consenso para os sítios 5' e 3', dessa vez baseadas em observações interespecíficas. O consenso do sítio de *splicing* 5' foi composto por 9 nucleotídeos sendo 7 invariáveis (C/AAG-

GTA/GAGT, onde “/” representa os nucleotídeos alternativos e “-” representa a fronteira éxon/intron). Por outro lado, o consenso do sítio de *splicing* 3' foi mais degenerado, sendo composto, em sua porção intrônica, por uma região rica em pirimidinas (C,T) e de tamanho variável (conhecida como trato de polipirimidina), seguida por uma posição não conservada, uma pirimidina e os nucleotídeos AG altamente conservados. Em sua porção exônica, o sítio de *splicing* 3' apresenta apenas uma Guanina (G) conservada. Portanto o consenso do sítio de *splicing* 3' pode ser representado por ((T/C)nN(C/T)AG-G).

Principais fatores de *splicing*

As pequenas ribonucleoproteínas (*Small nuclear ribonucleoproteins*, snRNPs) constituem uma classe de moléculas localizadas no núcleo, que são ubíquas em eucariotos; de insetos aos mamíferos. O nome dessas moléculas deriva da molécula de RNA (snRNA) que cada uma contém. Existem seis principais snRNPs (U1-U6 snRNPs) ; U3 snRNP é específica de nucléolo enquanto que as outras são encontradas no nucleoplasma e estão envolvidas no *splicing*. Estas snRNPs principais, estão presentes em aproximadamente 10^4 - 10^5 cópias por célula. Outras snRNPs que são menos abundantes (U7-U10) também foram identificadas. A análise da sequência de diversos snRNAs revelou alta conservação filogenética, reforçando sua importância funcional (Busch, *et al.*, 1982).

A hipótese de que o snRNP U1 era um fator de *splicing* fundamental, foi baseada na descoberta de uma complementariedade de bases entre a porção terminal do U1 snRNA e o sítio de *splicing* 5'. A ligação do snRNP U1 ao sítio

de *splicing* 5' é uma das primeiras etapas na formação do spliceossomo. Além do snRNP U1 existem diversas propostas para a participação de outras snRNPs, por exemplo a snRNP U2, no *splicing*. O papel fundamental da U2 nesta reação foi estabelecido através de experimentos que demonstraram que extratos nucleares empobrecido de U2 snRNP não poderiam sustentar uma reação de *splicing*. Experimentos de imunoprecipitação indicaram que o U2 snRNP se liga ao sítio de ramificação (*branch point*) e que outro fator, chamado de fator auxiliar da U2 (U2 auxiliary factor, U2AF), interage com o trato de polipirimidina e o sítio de *splicing* 3'. O U2AF é um heterodímero composto pelas proteínas, U2AF65 e U2AF35, que se ligam ao trato de polipirimidina e sítio de *splicing* 3' respectivamente. O recrutamento de três snRNPs (U4,U5 e U6) ao complexo de *splicing* caracteriza a etapa final da montagem do spliceossomo (House & Lynch, 2008).

Força do sítio de *splicing*

O papel funcional dos sítios filogeneticamente conservados (GT-AG) é demonstrado por experimentos de mutação direta nestes sítios. Mudanças em qualquer um desses nucleotídeos inativam completamente os sítios de *splicing*. Por outro lado, mutações em outras posições próximas a estes sítios apresentam efeitos variáveis. Um dos efeitos mais comuns é a ativação de sítios de *splicing* próximos ao sítio normalmente usado. Esses sítios são chamados de sítios crípticos. Esses sítios podem aparecer tanto na porção exônica como na intrônica de uma fronteira éxon/intron padrão. Devido à natureza degenerada dos sítios de *splicing*, as mutações os tornam mais ou menos próximos da sequência consenso que é a sequência com a maior

importância funcional por ter maior afinidade com os fatores do spliceossomo. Sendo assim, o conceito de força do sítio de *splicing* pode ser introduzido, referindo-se ao grau de similaridade de um determinado sítio de *splicing* em relação à sequência consenso. A ativação de sítios críticos pode ser entendida como consequência de uma mutação que tornou uma sequência ao seu redor, ou seja, o novo sítio de *splicing*, mais forte que o sítio comumente usado. Raciocínio similar pode ser aplicado para mutações que diminuem a força do sítio, tornando ativos outros sítios próximos que são mais fortes.

A etapa de reconhecimento de éxons (*Exon definition*)

O modelo de *splicing* descrito até aqui é conhecido como *Intron definition*. Este modelo propõe que os elementos em *cis* presentes no íntron, são fundamentais no processo de montagem do spliceossomo, sugerindo que as regiões gênicas reconhecidas durante a reação de *splicing* são, de modo geral, os íntrons. De acordo com esse modelo, pouca importância é atribuída às sequências exônicas no mecanismo de *splicing*. Entretanto, Robberson, *et al.*, (1990) observaram que: a) o sítio de *splicing* 3' é identificado (ligado pela snRNP U2) logo na etapa inicial da reação. b) que os fatores que reconhecem os sítios 5' e 3' (U1 e U2) interagem nas etapas iniciais da montagem do spliceossomo sugerindo que o reconhecimento desses sítios ocorre de maneira conjunta e não independente. c) que éxons grandes, maiores que 300 nucleotídeos, impedem a formação estável do spliceossomo.

Estas observações levaram os autores a proporem um novo modelo para o processo de *splicing*, onde primeiramente os éxons são reconhecidos e posteriormente ocorrem as etapas descritas pelo modelo *Intron definition*, que

seriam as responsáveis pela remoção dos íntrons de forma ordenada, isto é, sem omitir nenhum éxon, de um pré-RNA com vários éxons (Figura 2).

Figura 2: Esquema do modelo *Exon Definition* de *splicing* do pré-RNAm. snRNPs (círculos) e proteínas SR (triângulos) são mostradas interagindo com os éxons e definindo as bordas éxon/ítron (*Exon definition*). Modificado de Robberson *et al.*, (1990) e Berget, (1995).

Reguladores de *Splicing*

Para a identificação correta de quais são as junções éxon/intron de um pré-RNA, isto é, quais junções de um pré-RNA devem ser usadas para a produção do RNA maduro a célula faz uso de fatores de *splicing* proteicos além dos snRNPs. Entre estes fatores estão as proteínas da família SR.

Essas proteínas têm diversas funções nas várias etapas de montagem do spliceossomo, porém é na etapa inicial que as funções das proteínas SR estão melhor caracterizadas. Estas proteínas facilitam o recrutamento do U1 snRNP ao sítio de *splicing* 5' e dos U2AF e U2 snRNP ao trato de polipirimidina e sítio de *splicing* 3' respectivamente (Chen & Manley, 2009). Além disso, as proteínas RS estabilizam o pareamento das bases entre o U2 snRNA e o *branch point*.

As proteínas SR têm seu nome devido à presença de um domínio RS (domínio rico em Argininas (R) e Serinas (S)) na região C-terminal que permite a interação proteína-proteína. Por este motivo elas são consideradas como os principais responsáveis pela comunicação entre os sítios de *splicing* 5' e 3' prevista pelo modelo *Exon definition*. Além do domínio RS, as proteínas SR contém em sua porção N-terminal um domínio de reconhecimento de RNA, chamado RRM (*RNA recognition motif (RRM)*), que é o domínio que se liga aos reguladores de *splicing* em *cis*, geralmente localizados em éxons (ESRs, *Exonic Splicing Regulators*). Algumas proteínas SR contém mais de um domínio *RRM*, sendo um íntegro e o outro degenerado. Nesses casos, o domínio íntegro seria essencial para a ligação nos reguladores em *cis* e o segundo influenciaria na especificidade da ligação (Fu, 1995).

Curiosamente, as proteínas SR não foram encontradas em *S. cerevisiae*.

Esse resultado foi associado com o fato de que esse organismo apresenta íntrons pequenos dando origem à hipótese de que nesse organismo, o modelo de *splicing* predominante é o de reconhecimento de íntrons (*intron definition*) e não de éxons (*exon definition*) (Figura 2).

Foi proposto que o mecanismo de reconhecimento dos sítios de *splicing* que predomina em um dado organismo é aquele que oferece a menor distância entre os dois sítios adjacentes (Berget, 1995). Isto é, em organismos com íntrons pequenos predomina *Intron Definition* enquanto que nos organismos com éxons pequenos predomina *Exon Definition* (Sternier, *et al.*, 1996).

As proteínas SR foram organizadas em uma família cujos membros apresentam características em comum, como por exemplo, apresentar um fosfoepítipo reconhecido pelo anticorpo monoclonal mAb104, apresentar pelo menos um domínio RRM e um domínio RS e principalmente, serem capazes de complementar extratos citoplasmáticos deficientes para a reação de *splicing in vitro*. Em diversos organismos, como por exemplo, *Drosophila*, *Arabidopsis* e humanos as proteínas SR apresentam tamanhos similares quando avaliadas por eletroforese em gel. A análise da sequência de aminoácidos dessas proteínas revelou alta conservação filogenética (Fu, 1995).

Experimentos de *splicing in vitro* demonstraram que as proteínas SR são essenciais durante os passos iniciais do *splicing*. Entretanto diferentes proteínas SR são capazes de complementar (restaurar) extratos deficientes usados em reações de *splicing in vitro*, sugerindo que as proteínas SR têm funções redundantes no *splicing*. Apesar disso, cada proteína SR tem um grau de afinidade específico ao substrato (pré-RNA_m) que resulta em níveis distintos de eficiência na reação de *splicing* (Fu, 1995).

Exonic splicing enhancers (ESEs)

A especificidade da ligação das proteínas SR em seus substratos, isto é, nos pré-RNAs é baixa (Krainer, *et al.*, 1990). Foi demonstrado que quando isolada, a proteína SR SC35, apresenta baixa especificidade de ligação ao RNA, ou seja, a ligação pode ocorrer em diversos trechos do RNA independentemente da sequência de nucleotídeos presente no local da ligação. Entretanto, quando presente em conjunto com outros fatores de *splicing*, a SC35 foi encontrada próxima do sítio de *splicing* 3'. Esses resultados indicaram que outros componentes do spliceossomo colaboram para determinar a especificidade de ligação da SC35 durante a montagem do spliceossomo. Corroborando esses resultados, os experimentos de seleção *in vitro* para sítios de ligação de alta afinidade, resultaram em centenas de sequências curtas de nucleotídeos que apresentaram apenas uma característica em comum, eram ricas em purinas (A ou G). Essas sequências curtas de nucleotídeos ou *motifs*, foram denominadas *Exonic Splicing Enhancers* (ESEs) pois eram provenientes de sequências exônicas e quando presentes nos éxons favoreciam a inclusão deste no RNA maduro (Fu, 1995).

Identificação dos SELEX-ESEs.

A descoberta de alguns ESEs em estudos de caso (Lavigne, *et al.*, 1993; Watakabe, *et al.*, 1993) abriu a possibilidade de identificação de ESEs em larga escala. Em 1998, Liu e colaboradores modificaram a técnica de seleção *in vitro* SELEX (*Systematic Evolution of Ligands by Exponential Enrichment*) para esse fim. Nessa versão da técnica, um ESE previamente

conhecido, localizado no éxon central de um minigene (arranjo molecular contendo três éxons separados por dois íntrons, sendo o éxon central alternativo e os outros dois constitutivos) é substituído por *motifs* aleatórios. O *splicing* destes minigenes é realizado *in vitro*, em meio suplementado por uma proteína SR específica, para qual se pretende identificar seus sítios de ligação. Os produtos do *splicing* (RNAm) destes minigenes podem ou não conter o éxon alternativo central. Os RNAm que contém o éxon central são então convertidos a DNA e sequenciados para se descobrir os *motifs* com função de ESEs. Devido à baixa especificidade das proteínas SR pela sequência do substrato, uma única realização do experimento gera uma quantidade enorme de ESEs, por isso o experimento pode ser reiterado algumas vezes. Na segunda realização do experimento, os produtos gerados na primeira realização, entram como substratos iniciais do *splicing*. De três a cinco repetições são geralmente suficientes para capturar os ESEs com maiores afinidades pela proteína SR estudada (Liu, *et al.*, 1998; Cartegni, *et al.*, 2002). Na etapa final do método de identificação de ESEs as sequências obtidas com a técnica SELEX são alinhadas para obtenção de consensos. Os alinhamentos podem ser representados na forma de matrizes, conhecidas como PSSM (*Position Specific Scoring Matrices*, PSSM), que atribuem a cada nucleotídeo de cada posição do *motif*, uma nota que é proporcional à frequência desse nucleotídeo no alinhamento. Portanto, nucleotídeos mais e menos comuns em uma dada posição do *motif* recebem notas maiores ou menores respectivamente. A nota de um *motif* completo é obtida através da soma das notas de cada posição.

Identificação in silico de ESEs

Além dos *SELEX-ESEs*, vários outros prováveis ESEs foram identificados através de métodos computacionais.

Esses métodos usaram abordagens estatísticas para procurar sequências curtas de 6-8 p.b. enriquecidas em éxons com características biológicas propícias à dependência de ESEs. Um dos métodos, chamados *RESCUE (Relative Enhancer and Silencer Classification by Unanimous Enrichment)* comparou todos os hexâmeros (sequências de 6 p.b.) presentes em um conjunto de éxons com sítios de *splicing* fraco (uma das características dos éxons alternativos) com todos os *motifs* presentes em um conjunto de íntrons. Aqueles *motifs* que estavam enriquecidos nos éxons com sítios de *splicing* fraco em relação aos íntrons foram considerados ESEs e alguns deles tiveram sua funcionalidade testada (Fairbrother, *et al.*, 2002). Analogamente Zhang & Chasin, (2004) usaram sequências de éxons constitutivos para identificar um conjunto de ESEs, chamados PESE (*Putative exonic splicing enhancers*), enriquecidos em relação à pseudoéxons que não sofrem *splicing* e à regiões não traduzidas (5' *UTRs*) de genes sem íntrons.

Exonic Splicing Silencers (ESS)

Os *Exonic Splicing Silencers* (ESS) foram menos estudados do que os ESEs, porém desempenham função tão importante quanto eles na regulação do *splicing*. Foi demonstrado *in vitro* que um terço das sequências de 50-200 p.b. provenientes do genoma humano demonstra atividade inibitória (ou seja, são capazes de inibir a inclusão do éxon no RNAm) quando inseridas em um éxon de um minigene repórter. Curiosamente, sequências provenientes do

genoma de *E. coli* não apresentaram o mesmo efeito inibitório quando inseridas no mesmo minigene repórter (Fairbrother & Chasin, 2000). Esse resultado sugeriu a existência de diversos éxons potenciais, mantidos silenciados, que podem se tornar éxons reais, mediante aumento da força de seus sítios de *splicing* e/ou aquisição de ESEs eficientes (ou seja, com alta afinidade por proteínas SR). Os *Exonic Splicing Silencers* mais conhecidos são ligados por proteínas da família das ribonucleoproteínas heterogêneas nucleares (*heterogeneous nuclear ribonucleoprotein*, hnRNPs). Esta família é composta por diversas proteínas que estão ligadas ao pré-RNA nascente. Assim como as proteínas SR, as hnRNPs tem uma estrutura modular, apresentando uma ou mais domínios de ligação ao RNA e um domínio auxiliar cuja função é de interação entre proteínas. Uma das hnRNPs mais estudadas é a proteína de ligação ao trato de polipirimidina (*polypyrimidine tract binding protein*, PTB) (Cartegni, *et al.*, 2002).

As hnRNPs podem inibir a identificação do sítio de *splicing* direta ou indiretamente. A forma indireta ocorre quando o ESS está próximo ao sítio de *splicing* ou a um ESE. Nesse caso, a ligação da hnRNP ao ESS, pode obstruir a ligação dos fatores que favorecem o *splicing* como por exemplo snRNP U1 e snRNP U2, ou das proteínas SR. Por exemplo, a PTB se liga ao trato de polipirimidina bloqueando a ligação do U2AF nessa região. No modelo de ação direta as hnRNPs formam um multímero, ou seja, elas se interligam ao longo do éxon, podendo causar uma alteração local na conformação do pré-RNA que aproxima os sítios de *splicing* dos éxons vizinhos, causando a não inclusão (*skipping*) do éxon obliterado. O multímero também pode esconder os sítios de *splicing* do éxon onde ele se encontra favorecendo a omissão do éxon.

Identificação in vitro dos ESSs

Assim como os ESEs, os ESSs também foram experimentalmente identificados em larga escala (Wang, *et al.*, 2004). O método inovador consistiu em usar um minigene composto por três éxons capaz de reportar a inibição do éxon central através da expressão da proteína fluorescente verde GFP (*Green Fluorescent Protein, GFP*) em células humanas transfectadas. O primeiro e o terceiro éxons desse minigene codificam para o RNAm da GFP enquanto que o éxon 2, é um éxon teste contendo um sítio de clonagem onde podem ser inseridos oligonucleotídeos. Nesse minigene, o éxon teste é normalmente incluído no RNAm, resultando em uma proteína não funcional. Entretanto, a inserção de uma sequência com a função de ESS no éxon teste, pode causar a omissão deste éxon no RNAm, resultando na expressão de GFP. As células expressando GFP são separadas das que não expressam usando a técnica FACS (*Fluorescence Activated Cell Sorting, FACS*) e finalmente o DNA dessas células é extraído, amplificado e sequenciado revelando os oligonucleotídeos com função de ESS. O alinhamento das sequências resultantes fornece consensos que representam os *motifs* com maior atividade inibitória.

Evidências de função antagônica e dependente de contexto de certos ESRs

Originalmente os sítios de ligação das proteínas SR localizados em éxons foram descritos como fatores em *cis* importantes para a inclusão do éxon no transcrito maduro. Entretanto, trabalhos posteriores demonstraram que certos SELEX-ESEs apresentavam função oposta, isto é, de inibidores de *splicing*, quando localizados em íntrons (Kanopka, *et al.*, 1996; Jumaa &

Nielsen, 1997; Solis, *et al.*, 2008). Além disso, outros ESRs, ligados pelas proteínas NOVA1 e NOVA2 também apresentaram função antagônica dependendo de sua localização (Ule, *et al.*, 2006). Quando localizados em éxons funcionam como inibidores de inclusão de éxons (ESS) e quando localizados em íntrons funcionam como promotores de inclusão (*enhancers*). Estas descobertas evidenciaram que certos ESRs tem sua função dependente de sua localização no gene.

Outros autores defenderam a hipótese que a função de certos SELEX-ESEs também era dependente da posição destes no exon (Goren, *et al.*, 2006). A função antagônica de diferentes proteínas SR foi revelada explorando-se dois sítios de ligação distintos, alvos das proteínas SR SF2 e SRp40, localizados em um éxon alternativo de um minigene repórter. Quando o sítio de ligação para a proteína SF2 foi introduzido em 26 posições diferentes desse éxon alternativo, na maioria das vezes, eles atuaram como promotores e incluíram o éxon no transcrito maduro. Trocando o sítio de ligação da SF2 por um sítio da proteína SRp40 o efeito predominante foi inibir a inclusão do éxon.

Ainda nessa linha de argumentação, Solis *et al.*, (2008) corroboraram a função inibitória de um sítio de ligação da proteína SR SC35. Estes autores demonstraram que essa proteína pode inibir a inclusão do éxon 3 do gene do hormônio de crescimento (GH1). Eles descreveram uma mutação (A1338G) que cria no éxon, um sítio de ligação para a proteína SC35 e resulta em omissão (*skipping*) do éxon 3. A omissão do éxon 3 no RNAm do gene GH1 está associada com uma rara doença humana caracterizada pela produção insuficiente do hormônio do crescimento (GH).

De acordo com esses resultados, Jumaa & Nielsen, (1997) demonstraram que a inclusão do éxon 4 no RNAm da proteína SRp20 é antagonicamente regulado por duas proteínas SR distintas; a SRp20 (que portanto regula o *splicing* de seu próprio RNAm) e a SF2. A proteína SRp20 promove a inclusão do éxon 4 enquanto que a SF2 inibe essa inclusão. Além disso, foi demonstrado que esse padrão é dependente da força dos sítios de *splicing* do éxon 4. Quando os sítios de *splicing* do éxon 4, que são naturalmente fracos, foram transformados em fortes, o éxon 4 foi constitutivamente incluído e a SF2 não foi capaz de inibir essa inclusão.

Parte B) *Splicing*: Aspectos Biológicos e Evolutivos

Origens do *splicing*

Qual a origem dos íntrons presentes nos genes dos eucariotos e porque eles foram mantidos?

Duas teorias diametralmente opostas foram propostas para explicar a origem dos íntrons nos genes eucarióticos. A primeira, conhecida como *Introns Early*, propõe que a estrutura descontínua dos genes existia no ancestral dos eucariotos. Nesse caso, um mecanismo de *splicing* deveria existir previamente. De acordo com essa hipótese, a inexistência de íntrons nos procariotos atuais seria explicada simplesmente pela perda de íntrons em todas as espécies que compõe esse imenso grupo de organismos. Poucas evidências corroboram esta hipótese (Darnell, 1978; Doolittle, 1978).

A segunda hipótese (*Introns Late*), afirma que os genes do ancestral dos eucariotos não possuíam íntrons e que estes foram sendo inseridos ao longo da evolução. Dois mecanismos de inserção de íntrons foram propostos. Um

deles atribui inserções aos elementos transponíveis que podem ser diretamente inseridos no DNA (Cavalier-Smith, 1985). Essa hipótese se baseia na observação da ampla proliferação desses elementos nos genomas eucariotos. Um argumento contra essa hipótese afirma que a maioria desses eventos de inserção seriam deletérios e que portanto, um mecanismo de *splicing* deveria existir previamente, para remover as primeiras inserções presentes nos pré-RNAs. Um segundo mecanismo propõe que os primeiros íntrons eram do tipo auto-excisáveis (*self-splicing introns*). Esse tipo de íntron é encontrado atualmente nos genomas de organelas celulares de eucariotos, como mitocôndrias e plastos, sugerindo uma origem evolutiva bastante remota (van der Veen, *et al.*, 1986). Através de transcrição reversa do pré-RNA contendo um íntron ainda não excisado, o cDNA formado (contendo a sequência intrônica) seria diretamente inserida no genoma criando o gene com íntrons. Essa hipótese é atraente pois descarta a necessidade de uma maquinaria de *splicing* pré-existente para remoção dos primeiros íntrons já que eles eram auto-excisáveis. Além disso, evidências de que esses íntrons podem funcionar conjuntamente, isto é, de forma colaborativa na reação de *splicing* de outros genes, indica uma similaridade funcional entre os íntrons auto-excisáveis e os snRNAs (componentes das snRNPs), sugerindo que os íntrons auto-excisáveis poderiam ser os precursores do spliceossomo (Malek & Knoop, 1998; Dai & Zimmerly, 2002; Robart & Zimmerly, 2005). Um possível relaxamento de seleção nos íntrons auto-excisáveis devido à presença de fatores em *trans* que auxiliam na remoção do íntron seria um argumento para explicar a manutenção dos íntrons.

A hipótese *Introns Late*, não descarta possíveis vantagens adaptativas

dos íntrons, como por exemplo, a de que aumentam a variabilidade de produtos gênicos através de *splicing* alternativo e servem de substrato molecular para regiões regulatórias importantes que favorecem ou inibem a transcrição gênica ou o *splicing*.

Finalmente, resta a questão de porque os procariotos também não adquiriram íntrons ao longo da evolução. Especula-se que a intensidade de seleção natural em organismos procariotos é mais intensa que em eucariotos, porque o efeito de deriva genética nos procariotos é pequeno devido ao seu grande tamanho populacional efetivo. Sendo assim, assumindo que os íntrons tendem a ser onerosos à curto prazo, pois sobrecarregam a etapa de transcrição, eles foram eficientemente removido nos procariotos, mas não em eucariotos. (Green, 1986; Lynch, 2007).

***Splicing* alternativo como um mecanismo de geração de diversidade biológica**

Desde a descoberta de que os genes eucarióticos são formados por éxons e íntrons, hipóteses sobre a importância do *splicing* foram levantadas. Em um artigo seminal, Gilbert W., (1978) destacou o potencial impacto funcional de mutações genéticas que alteram a correta identificação das bordas éxon/íntron. Sendo assim, as mutações de ponto poderiam não só alterar um único aminoácido, como também alterar uma sequência deles, devido ao *splicing* modificado do gene. Além disso, o autor destacou que a descoberta dos íntrons quebrou o dogma “um gene uma proteína”, pois um único transcrito primário de uma região gênica pode dar origem a mais de um polipeptídeo dependendo das bordas éxons/íntrons usadas no *splicing*. Também levantou as

hipóteses de que possíveis falhas no mecanismo de *splicing* poderiam ser uma fonte de diversidade em escala de um único gene já que novos RNAm, possivelmente funcionais, seriam criados. Além disso, a presença dos íntrons facilitaria a troca de éxons entre genes através de recombinação não-homóloga (*exon shuffling*) aumentando a diversidade em escala genômica.

Atualmente sabemos que o *splicing* é um gerador de diversidade genética em eucariotos pois pode aumentar o número de produtos gênicos além de possibilitar uma gama de mecanismos de regulação da expressão gênica. A importância biológica do *splicing* alternativo foi logo exemplificada através do trabalho de Gebauer *et al.*, (1998) que demonstrou que a determinação do sexo em *Drosophila* é consequência de *splicing* alternativo.

Splicing alternativo regulado e não regulado

Nos genes humanos a abrangência do *splicing* alternativo é enorme. Wang *et al.*, (2008) estimaram que 95% dos genes humanos apresenta pelo menos uma variante de RNAm decorrente de *splicing* alternativo. Além disso, foi demonstrado que aproximadamente um terço dos eventos de *splicing* alternativo em humanos introduzem nos RNAm, códons de terminação prematuros, que levam à degradação do RNAm através do mecanismo *nonsense mediated decay* afetando a expressão gênica (Lewis, *et al.*, 2003).

Nem todos os eventos de *splicing* alternativos são regulados e uma parcela desses eventos é produto de erros ou falhas. Embora exista essa afirmativa seja amplamente aceita, a prevalência do *splicing* regulado e não-regulado em um dado organismo, é uma questão a ser definida.

O *splicing* alternativo regulado é aquele cujos produtos são

potencialmente funcionais e recorrentes no organismo ou tipo celular. O *splicing* alternativo não regulado é aquele cujas variantes são esporádicas e acontecem independentemente do tecido, da fase do ciclo celular ou da fase do desenvolvimento do organismo em que o gene é preferencialmente expresso (Green, 1986).

A utilização diferencial de fatores de *splicing* em *trans* (snRNPs, proteínas SR e hnRNPs) é o mecanismo mais provável de regulação do *splicing* alternativo (Auweter, *et al.*, 2006; Ule, *et al.*, 2006). Por outro lado, a utilização diferencial de fatores de *splicing* em *cis* (sítios de *splicing*, ESEs e ESS), por um mesmo conjunto de fatores em *trans*, é um dos principais mecanismos envolvidos no *splicing* alternativo não regulado (Green, 1986).

Dentre os eventos de *splicing* alternativo tecido-específicos (regulados), a maioria é conservada filogeneticamente (Wang, *et al.*, 2008). Portanto, assume-se que o *splicing* alternativo regulado seja filogeneticamente mais conservado que o não regulado. Partindo dessa hipótese, Yeo *et al.*, (2005) compararam milhares de éxons, ortólogos entre humanos e camundongos, e estimaram que a maioria deles não apresentou eventos de *splicing* alternativo conservados entre as duas espécies. Esse resultado demonstrou que a maioria dos eventos de *splicing* alternativo é espécie-específica e apenas sugeriu que muitos devem ser não regulados.

***Splicing* alternativo e doenças**

O *splicing* alternativo pode ser a causa direta ou provocar alterações na severidade de doenças humanas. Além disso, ele pode alterar a susceptibilidade dos indivíduos à doenças. Em todos esses casos, alterações em fatores de *splicing*, em *cis* ou em *trans*, são as principais responsáveis pela

alteração no *splicing* normal. As estimativas da fração de mutações que causam alterações no *splicing* e potencialmente estão associadas a doenças ainda são muito variáveis (15%-60%) (Wang & Cooper, 2007). Recentemente, foi estimado que no banco de dados de mutações humanas associadas a doenças (HGMD, *Human Gene Mutation Database*) existem quatro vezes mais alterações em ESEs do que em um banco de SNPs de uma população saudável (Sanford, *et al.*, 2009). A grande variação nesta estimativa reflete o conhecimento ainda incompleto dos fatores determinantes do *splicing*.

Alterações no *splicing* tem um importante papel no desenvolvimento e progressão do câncer. Genes associados com migração celular, regulação do crescimento celular, responsividade a hormônios e controle da apoptose apresentam alterações no *splicing* em células tumorais (Wang & Cooper, 2007).

Classificação dos éxons alternativos

O *splicing* alternativo pode dar origem a quatro tipos principais de éxons alternativos: i) éxon omitido (*exon skipped*) quando o éxon e os íntrons adjacentes são removidos do precursor do RNAm, ii) éxon com uso alternativo do sítio de *splicing* 5' (sítios crípticos 5') , iii) éxon com uso alternativo de sítio de *splicing* 3' (sítios crípticos 3'), iv) e íntron retido, quando um íntron não é removido do precursor do RNAm e portanto adquire status de éxon (Figura 3). Os éxons com uso alternativo de sítio de *splicing* também são chamados de éxons com bordas alternativas. Os éxons que sempre são incluídos no RNAm e que não apresentam uso alternativo de bordas são chamados éxons constitutivos.

Figura 3: Modos de *splicing* alternativo. Os quatro principais tipos de *splicing* alternativo estão ilustrados. Linhas indicam íntrons, regiões alteradas devido ao *splicing* alternativo estão sombreadas e linhas tracejadas indicam as alternativas do *splicing*. Modificado de Kim *et al.*, (2007).

Éxons alternativos do tipo *skipped* podem ser classificados de acordo com sua frequência de inclusão no *pool* de RNAm. O critério usado para essa classificação é a fração de cDNAs (geralmente *expressed sequence tags*, *ESTs*) que reportam o éxon alternativo. Éxons *skipped* de baixa inclusão são aqueles cuja fração de cDNAs que reporta a inclusão do éxon é menor que a fração que não reporta o éxon. Por outro lado, éxons *skipped* de alta inclusão são aqueles cuja fração de cDNA que reporta a inclusão do éxon é maior que a fração que não reporta. Em humanos, os éxons de alta inclusão representam a maioria dos casos de éxons *skipped* (Xing & Lee, 2005b; Xing & Lee, 2006). Arbitrariamente os limiares: um terço, entre um terço e dois terços e maior que dois terços, foram usados para definir as categorias de baixa, média e alta inclusão respectivamente (Xing & Lee, 2005a).

Os diferentes tipos de éxons alternativos ocorrem em diferentes frequências no genoma. Em humanos os éxons dos tipos *skipped* e borda alternativa correspondem a aproximadamente 50% do total de éxons

alternativos, sendo que as frequências de cada tipo são similares ($\approx 25\%$). A frequência dos íntrons retidos corresponde a menos de 5% do total. Combinações desses diferentes tipos de *splicing* alternativo (os tipos complexos) e alguns tipos raros como, por exemplo, éxons mutuamente exclusivos representam a fração restante do total de éxons alternativos (Lewis, *et al.*, 2003).

Considerando apenas os éxons alternativos conservados entre humanos e camundongos, os éxons *skipped* correspondem à maioria (38%). Éxons *skipped* conservados entre várias espécies tendem a ter um tamanho múltiplo de três, de modo que eles podem ser retirados ou inseridos do RNAm sem afetar a fase de leitura dos códons. Além disso, esses éxons estão estatisticamente associados aos casos de *splicing* tecido-específico (Xing & Lee, 2005b). Esses dois resultados sugerem que o *splicing* alternativo desses éxons é regulado por ser funcionalmente importante.

Origens do *splicing* alternativo

Como o *splicing* alternativo evoluiu ? Três mecanismos são descritos como possíveis geradores de éxons potencialmente alternativos.

Exonização de sequências intrônicas: Essa hipótese se baseia na evidência de que os elementos retrotransponíveis *Alu* foram encontrados apenas em éxons alternativos e não em constitutivos. As sequências *Alu* são retrotransposons que ocupam aproximadamente 10% do genoma humano. Esses elementos não codificam nenhuma proteína, porém podem ter efeitos na expressão de outros genes. Os *Alus* são raros em éxons, mas comuns em

íntrons e regiões intergênicas. Em humanos, todos os Alus exonizados foram encontrados em éxons alternativos e a maioria deles foi exonizada através de *splicing* alternativo (isto é estavam nos íntrons adjacentes ao novo éxon). Essas observações fundamentaram a hipótese de que a origem dos éxons alternativos seria através de modificação de elementos Alus intrônicos (Sorek, *et al.*, 2002).

Duplicação de genes e éxons: Os tipos de *splicing* podem variar consideravelmente após a duplicação de genes ou mesmo de éxons. Alterações no modo de *splicing* de genes duplicados foram demonstradas em *Arabidopsis thaliana* (Zhang, *et al.*, 2010). Em humanos, existem evidências de que praticamente todos os éxons duplicados, especialmente aqueles duplicados lado-a-lado (em *tandem*), sofrem *splicing* alternativo. Foi observado que dentre os éxons duplicados em *tandem*, apenas um é inserido no RNAm, ou seja, sofrem *splicing* do tipo mutuamente exclusivo (Letunic, *et al.*, 2002). Esses resultados sugerem que a duplicação gênica assim como a exônica podem ser mecanismos que favorecem o surgimento de *splicing* alternativo.

Relaxamento de seleção em éxons constitutivos: Apesar de interessantes, as duas hipóteses acima podem explicar apenas a minoria dos éxons alternativos do genoma humano, uma vez que menos de 5% desses éxons possuem Alus exonizados e menos de 20% estão associados com éxons duplicados. A hipótese de que os éxons alternativos surgem de éxons constitutivos poderia explicar a maioria dos casos. A comparação de éxons ortólogos de diversas espécies sugeriu que tanto os éxons alternativos

contendo sítios de *splicing* crípticos como os éxons do tipo *skipped*, podem ter origem em éxons constitutivos.

De acordo com Koren *et al.*, (2007), os éxons com sítios crípticos apresentam certas características como tamanho, conservação filogenética e manutenção da fase de leitura que se assemelham com os éxons constitutivos. Segundo essa hipótese, mutações próximas ao sítio de *splicing* podem resultar em novos sítios que competem com os sítios originais, levando ao uso alternativo de bordas. Pensando na evolução de um gene qualquer presente em várias espécies, é mais provável que tais mutações apareçam nas espécies onde a seleção purificadora sobre o sítio de *splicing* original é menor.

Através da comparação dos consensos dos sítios de *splicing* 5' de organismos que apresentam apenas éxons constitutivos (*S. cerevisiae* e *S. pombe*) e outros com éxons alternativos (humanos e camundongos) foi observada maior variação nos consensos de humanos e camundongos do que nos de organismos unicelulares (Ast, 2004). Esse resultado sugeriu que em eucariotos superiores os sítios de *splicing* são mais variáveis do que em eucariotos basais. Segundo esse autor, a maior variação nos sítios de *splicing* poderia ser consequência do relaxamento de seleção purificadora nesses sítios. Essa atenuação de seleção purificadora seria atribuída à perda parcial da importância funcional dos sítios de *splicing* que seria explicada pelo relativo aumento de importância de outros fatores em *cis*, por exemplo ESEs e ESS, envolvidos no *splicing*. Dessa forma, o *splicing* alternativo poderia ter se originado a partir de variações nos sítios de *splicing* de éxons constitutivos tornando-os mais fracos.

Resultados de experimentos *in vitro* também sugerem que os éxons

skipped, principalmente aqueles que são espécie-específicos, se originaram de éxons constitutivos. Foi demonstrado que éxons *skipped* conservados em humanos e camundongos, apresentam *splicing* constitutivo em galinha e peixe. Além disso, a transição constitutivo-alternativo pode ser reconstruída através de mutações no sítios de *splicing* 5' e em ESEs próximos desse sítio (Lev-Maor, *et al.*, 2007).

Finalmente, resultados obtidos através da genômica comparativa reforçam a hipótese de relaxamento de seleção em éxons alternativos. Foi observado que os éxons constitutivos humanos possuem mais ortólogos em espécies de vertebrados (87%) do que os éxons alternativos (68.5%). Além disso, parece existir uma tendência de aumento no tamanho dos íntrons ao longo da evolução dos vertebrados, especialmente dos mamíferos. Foi estimado que o tamanho de um íntron típico do ancestral comum dos mamíferos era quatro vezes menor que os íntrons do genoma humano. Sabe-se que os íntrons adjacentes aos éxons constitutivos humanos são menores que os íntrons adjacentes aos éxons alternativos. Portanto é possível que o tamanho dos íntrons adjacentes aos éxons constitutivos seja mantido por seleção purificadora enquanto que os íntrons adjacentes aos éxons alternativos não estejam sob grande restrição evolutiva (Gelfman, *et al.*, 2012).

Curiosamente, em mamíferos foi observado que os éxons constitutivos com sítios de *splicing* fracos apresentam íntrons menores que os éxons constitutivos com sítios de *splicing* fortes (Gelfman, *et al.*, 2012). Esse resultado sugeriu que a seleção purificadora contra o aumento dos íntrons é mais forte nos casos em que os íntrons estão ao redor de éxons constitutivos de sítio fraco do que de sítios fortes. Segundo essa lógica, a presença de sítios

de *splicing* fortes garantiria a precisão do *splicing*, e nesse contexto o tamanho do íntrons adjacentes não seria importante. Por outro lado, em éxons constitutivos de sítio fraco, a presença de íntrons grandes poderia prejudicar o reconhecimento desses éxons, de forma que os íntrons adjacentes teriam seu aumento de tamanho controlado por seleção natural (Gelfman, *et al.*, 2012).

Seleção natural em reguladores de *splicing* exônicos (ESRs)

Nesta seção serão resumidos resultados dos principais trabalhos que tiveram como foco evidenciar a seleção natural em reguladores de *splicing* exônicos (ESEs e ESS). Esses trabalhos usaram tanto abordagens filogenéticas como a populacionais. A abordagem filogenética (ou interespecífica) faz uso de mudanças genéticas (substituições ou divergências) acumuladas entre espécies. Essa abordagem possui as vantagens de usar um número relativamente alto de substituições conferindo-lhe maior poder estatístico e de ser bastante robusta aos efeitos da história demográfica, que geralmente é pouco conhecida. Os métodos populacionais fazem uso de dados de polimorfismos (por exemplo, SNPs) e são geralmente mais sensíveis aos efeitos demográficos, tornando-se necessário o uso de controles apropriados. Os métodos de criação de tais controles baseiam-se na teoria de Genética de Populações e destacam-se pela capacidade de modelar cenários de evolução neutros (evolução sem ação de seleção natural). Embora a abordagem populacional apresente poder estatístico reduzido em relação à abordagem filogenética, ela é mais sensível aos eventos seletivos recentes, principalmente aos de seleção positiva (Sabeti, *et al.*, 2006).

Finalmente, a comparação das duas abordagens, filogenética e

populacional é um recurso muito interessante para se distinguir a intensidade de seleção. Por exemplo, a seleção negativa de baixa intensidade, que atua sobre mutações fracamente deletérias, apresenta um padrão caracterizado por um déficit de substituições em relação aos polimorfismos. Esse relativo excesso de polimorfismo se deve ao menor tempo de ação da seleção negativa sobre a variação genética intrapopulacional em relação à variação interespecífica. Por outro lado, o excesso de substituições em relação aos polimorfismos, geralmente é uma característica atribuída aos *loci* sob seleção positiva (McDonald & Kreitman, 1991). Esse tipo de abordagem praticamente não foi utilizado na literatura referente aos reguladores de *splicing* exônicos e ela será extensivamente explorada no capítulo 3 da tese.

A seguir, iremos apresentar resultados relevantes de artigos que estudaram a evolução dos reguladores de *splicing*. Serão apresentadas evidências de seleção natural em ESRs, obtidas através de diferentes abordagens. Detalhes como, por exemplo, a diferença na intensidade e regime dessa seleção, dependendo dos diferentes tipos de mutações (e.g., ganho ou perda de função) e da sobreposição destas mutações com mutações que alteram os ou não os aminoácidos das proteínas também serão apresentados.

Seleção negativa

Como os reguladores de *splicing* e os códons encontram-se sobrepostos na sequência exônica, tanto as mutações não-sinônimas como as sinônimas podem alterar os ESRs. Sendo assim, é possível, por exemplo, classificar mudanças como não-sinônimas que mantêm ESEs ou como sinônimas que alteram ESEs, etc...

A razão (d_N/d_S) entre a taxa de substituição não-sinônima (d_N) e a taxa

de substituição sinônima (d_S) é frequentemente usada como uma medida de seleção natural. Como o efeito fenotípico de mudanças não-sinônimas na proteína é em média maior que o de mudanças sinônimas, essas últimas tornam-se mais comuns que as mudanças não-sinônimas, que são removidas por seleção natural negativa (ou purificadora). Sendo assim, a razão d_N/d_S apresenta um valor menor que um para diversos genes ortólogos de mamíferos (Bielawski, *et al.*, 2000) e para a maioria dos genes humanos (Bustamante, *et al.*, 2005).

Com o intuito de avaliar a diferença na intensidade de seleção natural sobre as mudanças não-sinônimas e sobre as que alteram ESEs foi proposta uma modificação do d_N/d_S padrão. Nesta nova versão, estima-se o d_{Nperda} e o d_{Sperda} (onde "perda" refere-se a perda de ESEs) (Ke, *et al.*, 2008). Comparando éxons constitutivos ortólogos entre humano e macaco-rhesus (*M. mulatta*) foram observados $d_N/d_S = 0.204$ e $d_{Nperda}/d_{Sperda} = 0.192$. A redução de 6% foi atribuída à seleção natural sobre mutações associadas à perda de ESEs.

Outra abordagem, descartou os fortes sinais de seleção negativa que sabidamente existem sobre mudanças não-sinônimas e, nesse caso, foram propostas duas medidas de divergência de ESRs baseadas apenas nas mudanças sinônimas.

Em uma delas, proposta por Ke *et al.*, (2008), a taxa de substituição sinônima de perda de ESRs, foi ponderada pela taxa sinônima associada à perda de *motifs* de um grupo controle, formado por sequências não-ESEs. A taxa sinônima de perda de ESEs foi menor do que a de perda de *motifs* do grupo controle. A redução nessa taxa foi de 13% e 17% para éxons

constitutivos e alternativos respectivamente. Uma maior redução na taxa sinônima de substituição dos ESEs de éxons alternativos em relação aos de constitutivos foi sugerida por Parmley *et al.*, (2006) e representa uma evidência de maior funcionalidade dos ESRs nos éxons alternativos.

Curiosamente, a taxa de substituição calculada a partir das mudanças não-sinônimas foi praticamente a mesma ($d_{Nperda} \sim 0.005$) tanto para as mudanças associadas à perda de ESRs como para as mudanças associadas à perda do conjunto controle. O mesmo foi observado para éxons constitutivos e alternativos analisados separadamente (Ke, *et al.*, 2008), sugerindo que o sinal de seleção natural em mudanças não-sinônimas predomina sobre o sinal de seleção em ESRs.

A segunda medida foi criada partindo-se da hipótese que as sequências ESEs são funcionais quando presentes em éxons, mas não em íntrons. Sendo assim, o d_S/d_i pondera a taxa sinônima pela taxa intrônica de evolução de ESEs. Para as mudanças de perda de ESE, a taxa sinônima foi reduzida em aproximadamente 40% em relação à taxa intrônica, resultando em $d_{S\ perda}/d_{i\ perda}$ menor que 1 (Ke *et al.*, 2008), fornecendo forte evidência de seleção purificadora sobre ESEs.

Evidências de seleção negativa em ESEs também foram descritas ao nível intraespecífico, usando dados de SNPs humanos. Nas regiões das bordas de éxons a densidade de SNPs é menor do que na região central (Majewski & Ott, 2002). Por outro lado, o oposto foi verificado para a distribuição dos ESEs, que ocorrem em maior densidade nas regiões de borda do que nas de centro (Fairbrother, *et al.*, 2004).

Para testar diretamente a hipótese de seleção natural em ESE,

Fairbrother *et al.*, (2004), analisaram as frequências (ou densidades) de SNPs de três categorias principais: a) SNPs associados a alteração de ESEs (subdivididos em SNPs de ganho e perda de ESE); b) associados com manutenção de ESEs e c) não associados a ESEs, isto é, fora de ESEs. Partindo da premissa de que alelos deletérios tendem a estar subrepresentados no conjunto total de alelos de uma amostra populacional (pois a seleção negativa os remove com maior eficiência da população), em relação aos neutros, essa comparação testou se há menor frequência de SNPs na categoria de alteração de ESE em relação às outras duas.

Usando um conjunto de 2561 SNPs exônicos esses autores observaram que 60.3% dos SNPs pertenciam à categoria fora de ESEs, 15.8% à categoria ganho de ESE, 12.9% à categoria manutenção e 10.9% à categoria perda. Para testar a hipótese de que esses valores diferem daqueles esperados, num cenário onde a seleção natural está ausente, os autores utilizaram uma simulação para prever a frequência esperada de SNPs em cada classe, em função apenas do processo mutacional. A comparação das mutações observadas empiricamente com as mutações simuladas, livres da ação da seleção natural, é uma maneira de se estudar a pressão seletiva ao nível molecular. Conforme esperado, eles observaram nos dados empíricos menor densidade de SNPs de perda de ESEs em relação aos dados simulados (10.9% vs. 13.6%). Esse resultado corroborou a hipótese de seleção natural negativa atuando sobre mutações que afetam ESEs. Vale destacar que essa análise foi feita usando mutações sinônimas e não-sinônimas separadamente e os resultados foram similares. Este resultado foi questionado por Carlini & Genut, (2006), pois para esses autores o sinal de seleção purificadora em

ESEs é mais forte em SNPs sinônimos do que nos não-sinônimos.

Em suma, é possível afirmar que as evidências disponíveis na literatura indicam que as mutações associadas à perda de ESE estão sob seleção purificadora. Embora os sinais dessa seleção sejam mais fracos que aqueles observados para as mutações não-sinônimas, eles são detectáveis através de abordagens populacionais e filogenéticas.

Seleção positiva

No contexto molecular, seleção positiva (ou darwiniana) é um tipo de seleção natural que atua sobre qualquer modalidade de variação genética (mutação de ponto, duplicação, recombinação, transposição, etc...) cujo efeito traz vantagens para o indivíduo que a possui em relação aos indivíduos que não a possui. Essa variação aumenta de frequência na população, pois os indivíduos portadores deixam mais descendentes que os demais (Nielsen, 2005).

Embora existam diversas evidências de seleção negativa sobre mudanças que alteram ESRs, há poucas de seleção positiva. Será destacado apenas um trabalho, que embora tenha apresentado uma fraca evidência, argumentou diretamente a favor de seleção positiva sobre esses elementos.

Segundo Ke *et al.*, (2008), as mutações sinônimas que criam ESEs em éxons constitutivos apresentam sinais de seleção positiva. Foi observado que, em éxons constitutivos, a taxa sinônima de ganho de ESEs ($d_{\text{Sganho}} \sim 0.017$) era ligeiramente maior que a taxa intrônica de ganho de ESEs ($d_{\text{iganho}} \sim 0.016$), resultando em um $d_{\text{Sganho}} / d_{\text{iganho}}$ maior que 1. Esse resultado foi interpretado como um sinal de seleção positiva.

Ainda neste trabalho foram demonstradas evidências de que a seleção positiva sobre ESRs poderia ser atribuída a um efeito compensatório, fundamental para manutenção do *splicing*, isto é, mudanças com efeitos negativos sobre a inclusão do éxon (perda de ESE, ganho de ESS e enfraquecimento dos sítios de *splicing*) seriam compensadas por mutações com efeitos positivos sobre a inclusão do éxon (ganho de ESE, perda de ESS e fortalecimento dos sítios de *splicing*). A importância de mutações compensatórias em outros fatores de *splicing* em *cis* (trato de polipirimidina e *branch site*) e *trans* foi demonstrada por outros (Dewey, *et al.*, 2006; Plass, *et al.*, 2008; Schwartz, *et al.*, 2008).

Evolução dos ESRs em genes parálogos

A evolução dos ESRs também foi abordada através da comparação dos éxons de genes parálogos (Zhang, *et al.*, 2009). A diferença na densidade de reguladores de *splicing* em *cis* (ESEs e ESSs) entre os éxons duplicados fornece uma quantificação da diferença de potencial regulador acumulada após a duplicação dos genes, já a taxa de substituição sinônima medida entre eles, fornece uma estimativa do tempo de divergência entre as cópias.

Conforme esperado, foi observada uma correlação positiva entre a diferença de densidade de ESRs e a taxa de substituição sinônima dos parálogos. Isto é, com o aumento do tempo desde a duplicação, a diferença na densidade de reguladores entre as cópias também aumentou.

Embora esta correlação positiva também tenha sido observada para um conjunto controle, formado por sequências não-ESRs, os ESRs apresentaram diferenças importantes. Independentemente do tempo de divergência dos

parálogos, a diferença de densidade de ESEs entre as cópias foi menor que a observada para os do conjunto controle. Esse resultado sugere que mesmo que duplicação gênica resulte, logo após a duplicação, em um relaxamento de seleção negativa em uma das cópias (Ohno, *et al.*, 1968), a seleção sobre ESEs ainda pode ser detectada usando parálogos de diversas idades.

Curiosamente, um padrão oposto foi observado para os ESSs, isto é, a diferença de densidade de ESSs entre as cópias, foi maior que a observada para os *motifs* do conjunto controle, principalmente para os pares de parálogos mais antigos, isto é que divergiram há mais tempo.

Esse resultado sugere que após a duplicação gênica, ESEs e ESSs experimentam regimes de seleção opostos. É possível que os ESSs sejam evitados nos éxons de genes de cópia única e que após a duplicação, o relaxamento dessa seleção permita o aumento na densidade de ESSs.

Os resultados ficam ainda mais ilustrativos quando foram considerados apenas os casos em que os éxons ancestrais eram constitutivos e os derivados alternativos; observou-se uma diminuição na taxa sinônima de criação de ESEs e um aumento na taxa sinônima de criação de ESSs.

Seleção natural em éxons constitutivos e alternativos

A categoria de éxons alternativos é muito heterogênea pois inclui éxons com diferentes modos de *splicing* (*skipping*, sítio críptico, retenção de íntron e eventos complexos), de conservação filogenética (conservado ou não conservado) e de níveis de inclusão. Além disso, a classificação de um éxon alternativo como “conservado” pode se referir a duas instâncias; conservado no genoma (DNA) ou no transcriptoma (cDNA). No primeiro caso, o éxon está

presente nos genomas das espécies comparadas, ou seja, apresenta ortólogos, porém não necessariamente apresenta variantes de cDNA em todas as espécies comparadas, podendo ser alternativo em apenas uma. No segundo caso, o éxon é alternativo em todas as espécies comparadas. Sendo assim, não é surpreendente que essa categoria de éxon apresente resultados contrastantes na literatura, principalmente em relação a diferenças na taxa de evolução.

A maioria dos estudos direcionados à compreensão do padrão evolutivo dos éxons alternativos e constitutivos utilizou dados de sequências de éxons *skipped* conservados nos genomas de várias espécies (humano, camundongo e chimpanzé). Esses estudos demonstram claramente que os éxons das categorias constitutivos, alternativos e as diferentes subcategorias de éxons alternativos evoluem sob regimes de seleção diferentes. Os éxons constitutivos evoluem sob um regime de seleção purificadora mais forte do que os éxons alternativos, principalmente ao nível da proteína (Xing & Lee, 2006). As observações que corroboram essa hipótese são: maior proporção de ortólogos (Haerty & Golding, 2009; Gelfman, *et al.*, 2012), menor taxa de evolução não-sinônima (Iida & Akashi, 2000; Xing & Lee, 2005a; Plass & Eyra 2006), maior viés para o uso de códons (Iida & Akashi, 2000; Haerty & Golding, 2009), maior proporção dentro de domínios proteicos (Haerty & Golding, 2009) e menor proporção em regiões não-estruturadas da proteína (Haerty & Golding, 2010). Estes resultados corroboram a hipótese de que, por serem mais expressos que os éxons alternativos (os éxons constitutivos estão sempre presentes no RNAm), os éxons constitutivos estão sujeitos a um regime de seleção purificadora mais forte que os alternativos (Carlini & Genut, 2006; Xing & Lee,

2006; Haerty & Golding, 2009; Haerty & Golding, 2010).

O único resultado que desafia a hipótese de maior restrição evolutiva em éxons constitutivos é a maior taxa de evolução sinônima desses éxons em relação, principalmente, a uma categoria específica de éxons alternativos; os éxons *skipped* de baixa inclusão (Iida & Akashi, 2000; Parmley, *et al.*, 2006; Plass & Eyra, 2006). Na comparação entre humanos e camundongos, a taxa de substituição sinônima dos éxons de baixa inclusão foi aproximadamente 4 vezes menor que dos éxons de alta inclusão e resultados similares foram observados para comparações entre humano e chimpanzé (Xing & Lee, 2005a). Esses resultados sugerem haver maior seleção natural conservando os elementos reguladores de *splicing* dos éxons de baixa do que de alta inclusão. Os éxons *skipped* de baixa inclusão representam a menor parte dos éxons *skipped* humanos (Lev-Maor, *et al.*, 2007). A análise em larga escala do genoma de camundongo revelou que os éxons de baixa inclusão representam aproximadamente 10% do total de éxons *skipped* analisados. Além disso, os éxons *skipped* de baixa inclusão possuem menos ortólogos em humanos do que os de alta inclusão (30% vs. 60% respectivamente) e portanto, apresentam conservação filogenética menor que os éxons de alta inclusão (Pan, *et al.*, 2004).

Análises baseadas na comparação interespecífica indicam os éxons de baixa inclusão, são alvos de seleção positiva (Xing & Lee, 2005a; Ramensky, *et al.*, 2008). Além disso, os éxons de baixa inclusão também apresentaram taxa de evolução não-sinônima, aproximadamente 4 vezes maior que a dos éxons de alta inclusão, resultando em maior D_n/D_s para os éxons de baixa inclusão do que para os éxons de alta inclusão. A maior taxa de evolução não-sinônima

poderia ser um sinal de seleção positiva atuando na proteína criada através da inclusão do éxon alternativo (Xing & Lee, 2005a; Xing & Lee, 2006).

Foi demonstrado, em camundongos, que éxons *skipped* com grandes diferenças de inclusão entre tecidos tendem a manter a fase de leitura dos códons, isto é, tem um tamanho em nucleotídeos que é múltiplo de três (Xing & Lee, 2005b). É provável que uma parcela dos éxons de baixa inclusão com sinais de seleção positiva, faça parte do conjunto de éxons com grande diferença de inclusão entre tecidos (éxons tecido-específicos). Éxons podem ser classificados como de baixa inclusão em relação a um conjunto de transcritos vindos de diversos tecidos, porém uma parcela desses éxons pode ser considerada de alta inclusão em relação ao total de transcritos de um determinado tecido.

A combinação de uma maior taxa não-sinônima de evolução dos éxons de baixa inclusão, e uma maior manutenção da fase de leitura dos éxons com grandes diferenças de expressão entre tecidos (éxons tecido-específicos), fortalecem a hipótese de seleção positiva sobre a nova isoforma da proteína formada pelo *splicing* alternativo.

Em suma, pode-se afirmar que apesar das controvérsias sobre o padrão evolutivo de éxons constitutivos e alternativos (Ermakova, *et al.*, 2006), evidências vindas de abordagens metodológicas independentes, sugerem que os éxons constitutivos estão sob maior seleção purificadora do que éxons alternativos. Embora exista grande heterogeneidade nos padrões evolutivos dos éxons alternativos, os éxons alternativos tecido-específicos, provavelmente representam a categoria de éxons alternativos com o maior sinal de seleção positiva que provavelmente atua ao nível da proteína. Evidências de que genes

com expressão tecido-específica estão sob menor restrição evolutiva do que os genes de expressão generalizada (Zhang & Li, 2004) favorecem esta hipótese.

Tempo de divergência influencia na detecção de seleção natural em ESEs.

Conforme apresentado na seção anterior, Xing e Lee, (2005a), observaram menor taxa de substituição sinônima nos éxons *skipped* de baixa inclusão do que nos de alta inclusão. Esse resultado sugeriu que os éxons *skipped* de baixa inclusão estão sob maior restrição evolutiva ao nível do RNA do que os de alta inclusão. Entretanto, essa diferença foi maior quando a comparação foi feita entre éxons ortólogos de espécies filogeneticamente mais distantes (humano e camundongo, divergência estimada em 90 milhões de anos atrás) do que de entre éxons de espécies mais próximas (humano e chimpanzé, divergência estimada em 5 milhões de anos atrás).

Especificamente, na comparação entre humano e chimpanzé, a taxa de substituição sinônima observada para os éxons de baixa inclusão foi apenas 37% menor que para os éxons de alta inclusão. Já na comparação entre humano e camundongo, a diferença nessa taxa foi de quatro vezes. Este resultado também foi validado através de uma segunda medida de restrição evolutiva: a manutenção da fase de leitura dos códons. A fração de éxons *skipped* que mantém a fase de leitura é maior para os éxons de baixa inclusão do que para os de alta inclusão, entretanto essa diferença é maior para comparações entre éxons ortólogos de humano e camundongo do que para éxons ortólogos de humano e chimpanzé. Esse resultado sugere que a força da seleção natural nas mutações sinônimas que afetam o *splicing* é

relativamente fraca e portanto seus efeitos ficam mais evidentes apenas após um longo período de tempo (Xing & Lee, 2005b).

Sobreposição dos códigos de *splicing* e proteico **As hipóteses de evolução “oportunista” vs. “conflitante”**

A densidade dos reguladores de *splicing* em *cis* (ESEs) aumenta em direção às bordas dos éxons. Além disso, sabe-se que os ESEs apresentam uma composição nucleotídica rica em purinas (A e G) (Blencowe, 2000). Sendo assim, não é surpreendente que as bordas dos éxons codificadoras apresentem um viés para certos aminoácidos cujos códons contenham mais purinas e menos pirimidinas, especialmente a citosina que é praticamente duas vezes menos frequente nos ESEs do que nos éxons codificadores do genoma humano. Por exemplo, o aminoácido mais comum nas bordas dos éxons codificadores é lisina, codificado por AAA e AAG. Por outro lado, entre os aminoácidos menos comuns está prolina, codificado por quatro códons que contém, no total, apenas uma adenina e uma guanina (Parmley, *et al.*, 2007).

Portanto, os ESEs parecem influenciar não apenas no uso de aminoácidos em borda como também no uso de códons sinônimos desses aminoácidos. Por exemplo, o aminoácido arginina é codificado por seis códons que podem ser divididos em dois grupos, o primeiro contém quatro códons do tipo CGX (onde X são os quatro nucleotídeos possíveis) e o segundo grupo contém apenas os códons AGA e AGG. Foi observado que os códons do segundo grupo são mais comuns nas bordas dos éxons do que os códons do primeiro grupo (Parmley, *et al.*, 2007).

Essa diferença na frequência de aminoácidos entre as bordas e as regiões centrais dos éxons é refletida em taxas de evolução menores nas

bordas do que nas regiões centrais dos éxons. Foi estimado que, dentro de um mesmo gene, as bordas dos éxons evoluem aproximadamente 2.3 vezes mais lentamente que a região central independentemente da classe funcional da proteína (Parmley, *et al.*, 2007). Esses resultados levaram os autores a afirmar que o tamanho do éxon é um dos principais determinantes da evolução das proteínas, com efeito similar a outro conhecido fator determinante; a quantidade de tecidos em que um gene é expresso (Zhang & Li, 2004).

Embora os resultados mostrem claramente a influência dos ESEs na determinação de quais aminoácidos irão ocupar as bordas dos éxons, não existem evidências suficientes que corroborem o “modelo conflitante” citado por Warnecke *et al.*, (2009). De acordo com esse modelo a influência dos ESEs deve ser prejudicial à função da proteína gerando um situação de conflito entre os dois códigos, o proteico (genético) e código do *splicing* formado por ESRs.

Diversos trabalhos mostraram que nem sempre a troca de aminoácidos tem efeitos deletérios para a função da proteína, pois aspectos bioquímicos como, por exemplo, a posição do resíduo na estrutura secundária, as modificações pós-traducionais (por exemplo, fosforilação) e características químicas como solubilidade devem ser levadas em consideração.

Portanto, em relação ao padrão evolutivo de códigos sobrepostos, o chamado “modelo oportunista” é mais plausível que o modelo conflitante. No modelo oportunista a flexibilidade de um dos códigos é usada (de forma oportuna) para a evolução do outro código. É razoável assumir que o código genético e de *splicing* evoluam de forma a minimizar a interferência mútua. A degeneração do código genético é um exemplo disso pois permite oportunidades de ajustes para outros códigos.

Bibliografia

- Ast, G. (2004). "How did alternative splicing evolve?" Nat Rev Genet **5**(10): 773-782.
- Auweter, S. D., R. Fasan, et al. (2006). "Molecular basis of RNA recognition by the human alternative splicing factor Fox-1." EMBO J **25**(1): 163-173.
- Berget, S. M. (1995). "Exon recognition in vertebrate splicing." J Biol Chem **270**(6): 2411-2414.
- Bielawski, J. P., K. A. Dunn, et al. (2000). "Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions." Genetics **156**(3): 1299-1308.
- Blencowe, B. (2000). "Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases." Trends Biochem Sci **25**(3): 106-110.
- Breathnach, R., C. Benoist, et al. (1978). "Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries." Proc Natl Acad Sci U S A **75**(10): 4853-4857.
- Busch, H., R. Reddy, et al. (1982). "SnRNAs, SnRNPs, and RNA processing." Annu Rev Biochem **51**: 617-654.
- Bustamante, C., A. Fledel-Alon, et al. (2005). "Natural selection on protein-coding genes in the human genome." Nature **437**(7062): 1153-1157.
- Carlini, D. B. and J. E. Genut (2006). "Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers." J Mol Evol **62**(1): 89-98.
- Cartegni, L., S. Chew, et al. (2002). "Listening to silence and understanding nonsense: exonic mutations that affect splicing." Nat Rev Genet **3**(4): 285-298.
- Cavalier-Smith, T. (1985). "Selfish DNA and the origin of introns." Nature **315**(6017): 283-284.
- Chen, M. and J. L. Manley (2009). "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches." Nat Rev Mol Cell Biol **10**(11): 741-754.
- Dai, L. and S. Zimmerly (2002). "Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior." Nucleic Acids Res **30**(5): 1091-1102.
- Darnell, J. E. (1978). "Implications of RNA-RNA splicing in evolution of eukaryotic cells." Science **202**(4374): 1257-1260.
- Dewey, C. N., I. B. Rogozin, et al. (2006). "Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns." BMC Genomics **7**: 311.
- Doolittle, W. F. (1978). Genes in pieces: were they ever together? Nature. **272**: 581-582.
- Ermakova, E. O., R. N. Nurtdinov, et al. (2006). "Fast rate of evolution in alternatively spliced coding regions of mammalian genes." BMC Genomics **7**: 84.

- Fairbrother, W., R. Yeh, et al. (2002). "Predictive identification of exonic splicing enhancers in human genes." Science **297**(5583): 1007-1013.
- Fairbrother, W.G., and Chasin, L.A. (2000). "Human genomic sequences that inhibit splicing". Mol Cell Biol **20**: 6816-6825
- Fairbrother, W. G., D. Holste, et al. (2004). "Single nucleotide polymorphism-based validation of exonic splicing enhancers." PLoS Biol **2**(9): E268.
- Fu, X. D. (1995). "The superfamily of arginine/serine-rich splicing factors." RNA **1**(7): 663-680.
- Gebauer, F., L. Merendino, et al. (1998). "The Drosophila splicing regulator sex-lethal directly inhibits translation of male-specific-lethal 2 mRNA." RNA **4**(2): 142-150.
- Gelfman, S., D. Burstein, et al. (2012). "Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons." Genome Res **22**(1): 35-50.
- Gilbert, W. (1978). "Why genes in pieces?" Nature **271**(5645): 501.
- Goren, A., O. Ram, et al. (2006). "Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers." Mol Cell **22**(6): 769-781.
- Green, M. R. (1986). "Pre-mRNA splicing." Annu Rev Genet **20**: 671-708.
- Green, M. R., T. Maniatis, et al. (1983). "Human beta-globin pre-mRNA synthesized in vitro is accurately spliced in *Xenopus* oocyte nuclei." Cell **32**(3): 681-694.
- Haerty, W. and B. Golding (2009). "Similar selective factors affect both between-gene and between-exon divergence in *Drosophila*." Mol Biol Evol **26**(4): 859-866.
- Haerty, W. and G. B. Golding (2010). "Genome-wide evidence for selection acting on single amino acid repeats." Genome Res **20**(6): 755-760.
- Hertel, K. (2008). "Combinatorial control of exon recognition." J Biol Chem **283**(3): 1211-1215.
- House, A. E. and K. W. Lynch (2008). "Regulation of alternative splicing: more than just the ABCs." J Biol Chem **283**(3): 1217-1221.
- Iida, K. and H. Akashi (2000). "A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes." Gene **261**(1): 93-105.
- Jumaa, H. and P. J. Nielsen (1997). "The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation." EMBO J **16**(16): 5077-5085.
- Kanopka, A., O. Mühlemann, et al. (1996). "Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA." Nature **381**(6582): 535-538.
- Ke, S., X. H. Zhang, et al. (2008). "Positive selection acting on splicing motifs reflects compensatory evolution." Genome Res **18**(4): 533-543.
- Kim, E., A. Goren, et al. (2008). "Alternative splicing: current perspectives." Bioessays **30**(1): 38-47.
- Koren, E., G. Lev-Maor, et al. (2007). "The emergence of alternative 3' and 5' splice site exons from constitutive exons." PLoS Comput Biol **3**(5): e95.

- Krainer, A. R., G. C. Conway, et al. (1990). "Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells." Genes Dev **4**(7): 1158-1171.
- Lavigne, A., H. La Branche, et al. (1993). "A splicing enhancer in the human fibronectin alternate ED1 exon interacts with SR proteins and stimulates U2 snRNP binding." Genes Dev **7**(12A): 2405-2417.
- Letunic, I., R. R. Copley, et al. (2002). "Common exon duplication in animals and its role in alternative splicing." Hum Mol Genet **11**(13): 1561-1567.
- Lev-Maor, G., A. Goren, et al. (2007). "The "alternative" choice of constitutive exons throughout evolution." PLoS Genet **3**(11): e203.
- Lewis, B. P., R. E. Green, et al. (2003). "Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans." Proc Natl Acad Sci U S A **100**(1): 189-192.
- Liu, H. X., M. Zhang, et al. (1998). "Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins." Genes Dev **12**(13): 1998-2012.
- Lynch, M. (2007). The Origins of Genome Architecture, Sinauer Associates, Inc.
- Majewski, J. and J. Ott (2002). "Distribution and characterization of regulatory elements in the human genome." Genome Res **12**(12): 1827-1836.
- Malek, O. and V. Knoop (1998). "Trans-splicing group II introns in plant mitochondria: the complete set of cis-arranged homologs in ferns, fern allies, and a hornwort." RNA **4**(12): 1599-1609.
- McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila." Nature **351**(6328): 652-654.
- Mount, S. M. (1982). "A catalogue of splice junction sequences." Nucleic Acids Res **10**(2): 459-472.
- Nielsen, R. (2005). "Molecular signatures of natural selection." Annu Rev Genet **39**: 197-218.
- Ohno, S., U. Wolf, et al. (1968). "Evolution from fish to mammals by gene duplication." Hereditas **59**(1): 169-187.
- Pan, Q., O. Shai, et al. (2004). "Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform." Mol Cell **16**(6): 929-941.
- Parmley, J., A. Urrutia, et al. (2007). "Splicing and the evolution of proteins in mammals." PLoS Biol **5**(2): e14.
- Parmley, J. L., J. V. Chamary, et al. (2006). "Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers." Mol Biol Evol **23**(2): 301-309.
- Plass, M., E. Agirre, et al. (2008). "Co-evolution of the branch site and SR proteins in eukaryotes." Trends Genet **24**(12): 590-594.
- Plass, M. and E. Eyras (2006). "Differentiated evolutionary rates in alternative exons and the implications for splicing regulation." BMC Evol Biol **6**: 50.
- Ramensky, V., R. Nurtdinov, et al. (2008). "Positive selection in alternatively spliced exons of human genes." Am J Hum Genet **83**(1): 94-98.

- Robart, A. R. and S. Zimmerly (2005). "Group II intron retroelements: function and diversity." Cytogenet Genome Res **110**(1-4): 589-597.
- Robberson, B. L., G. J. Cote, et al. (1990). "Exon definition may facilitate splice site selection in RNAs with multiple exons." Mol Cell Biol **10**(1): 84-94.
- Sabeti, P. C., S. F. Schaffner, et al. (2006). "Positive natural selection in the human lineage." Science **312**(5780): 1614-1620.
- Sanford, J. R., X. Wang, et al. (2009). "Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts." Genome Res **19**(3): 381-394.
- Schwartz, S. H., J. Silva, et al. (2008). "Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes." Genome Res **18**(1): 88-103.
- Solis, A. S., R. Peng, et al. (2008). "Growth hormone deficiency and splicing fidelity: two serine/arginine-rich proteins, ASF/SF2 and SC35, act antagonistically." J Biol Chem **283**(35): 23619-23626.
- Sorek, R., G. Ast, et al. (2002). "Alu-containing exons are alternatively spliced." Genome Res **12**(7): 1060-1067.
- Sterner, D. A., T. Carlo, et al. (1996). "Architectural limits on split genes." Proc Natl Acad Sci U S A **93**(26): 15081-15085.
- Ule, J., G. Stefani, et al. (2006). "An RNA map predicting Nova-dependent splicing regulation." Nature **444**(7119): 580-586.
- van der Veen, R., A. C. Arnberg, et al. (1986). "Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro." Cell **44**(2): 225-234.
- Wallace, J. C. and M. Edmonds (1983). "Polyadenylated nuclear RNA contains branches." Proc Natl Acad Sci U S A **80**(4): 950-954.
- Wang, E. T., R. Sandberg, et al. (2008). "Alternative isoform regulation in human tissue transcriptomes." Nature **456**(7221): 470-476.
- Wang, G. and T. Cooper (2007). "Splicing in disease: disruption of the splicing code and the decoding machinery." Nat Rev Genet **8**(10): 749-761.
- Wang, Z., M. Rolish, et al. (2004). "Systematic identification and analysis of exonic splicing silencers." Cell **119**(6): 831-845.
- Warnecke, T., C. Weber, et al. (2009). "Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence." Biochem Soc Trans **37**(Pt 4): 756-761.
- Watakabe, A., K. Tanaka, et al. (1993). "The role of exon sequences in splice site selection." Genes Dev **7**(3): 407-418.
- Xing, Y. and C. Lee (2005a). "Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences." Proc Natl Acad Sci U S A **102**(38): 13526-13531.
- Xing, Y. and C. Lee (2006). "Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes." Nat Rev Genet **7**(7): 499-509.

- Xing, Y. and C. J. Lee (2005b). "Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing." PLoS Genet **1**(3): e34.
- Yeo, G. W., E. Van Nostrand, et al. (2005). "Identification and analysis of alternative splicing events conserved in human and mouse." Proc Natl Acad Sci U S A **102**(8): 2850-2855.
- Zhang, L. and W. H. Li (2004). "Mammalian housekeeping genes evolve more slowly than tissue-specific genes." Mol Biol Evol **21**(2): 236-239.
- Zhang, P. G., S. Z. Huang, et al. (2010). "Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of Arabidopsis." Mol Biol Evol **27**(7): 1686-1697.
- Zhang, X. and L. Chasin (2004). "Computational definition of sequence motifs governing constitutive exon splicing." Genes Dev **18**(11): 1241-1250.
- Zhang, Z., L. Zhou, et al. (2009). "Divergence of exonic splicing elements after gene duplication and the impact on gene structures." Genome Biol **10**(11): R120.

Capítulo 2

Splicing* alternativo e diversidade genética: SNVs associados à variantes de splicing são mais frequentes em inibidores de *splicing

Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders

Prólogo

O artigo a seguir, publicado na revista *Nucleic Acids Research (Nucl. Acids Res. (2011) 39 (12): 4942-4948)*, apresenta uma análise da densidade dos principais reguladores de *splicing* exônicos em sequências de éxons constitutivos e alternativos do genoma humano.

A análise individual dos três principais tipos de éxons alternativos (*skipped*, borda alternativa e retenção de íntron) representou um avanço na distinção entre os tipos, pois até o momento, a maioria das publicações descreveu a densidade de reguladores apenas em éxons constitutivos e *skipped*.

Primeiramente, nossos resultados corroboram um resultado prévio da literatura que se refere à menor densidade de ESS (*ESS, Exonic Splicing Silencers*) em éxons constitutivos do que em alternativos. Além disso, o nosso artigo demonstrou que os éxons constitutivos apresentam menor densidade de reguladores de *splicing* do que éxons alternativos desafiando resultados prévios da literatura. Atribuímos essa discrepância à utilização de uma base de dados de éxons do genoma humano, mais atual do que as anteriormente utilizadas. Essa atualização é importante, pois a quantidade de sequências de cDNA depositadas nos bancos de dados públicos aumentou rapidamente ao longo dos últimos anos, e com uma maior amostragem ocorrem mudanças na classificação de alguns éxons constitutivos, que passam a pertencer à categoria alternativos.

Finalmente, apresentamos um método novo de identificação de possíveis reguladores de *splicing* envolvidos no *splicing* alelo-específico

(*splicing* alternativo regulado por variações genéticas de um único nucleotídeo). Enquanto outros autores demonstraram experimentalmente a influência dos SNVs na determinação de variantes de *splicing*, no nosso trabalho usamos uma abordagem em larga escala, onde os três principais tipos de éxons alternativos foram analisados. Nosso resultado trouxe evidências de que os *ESS* estão envolvidos mais frequentemente nesse tipo de *splicing* do que os *ESEs*.

Minha participação nesse trabalho foi fundamental para a redefinição da questão central do artigo, para a atualização das referências citadas ao longo do texto (acrescentando novas discussões), na redação de trechos do artigo, nas análises estatísticas e no processo de submissão e re-submissão das respostas aos revisores.

Resumo

Devido a grande quantidade de dados genômicos disponíveis espera-se que a influência das mutações de ponto em diversos fenômenos biológicos seja elucidada. Neste trabalho, abordamos a questão de como as variações de um único nucleotídeo (SNVs, *Single Nucleotide Variants*) afetam o *splicing* alternativo no genoma humano. Nós observamos que os éxons alternativos apresentam maior densidade de SNVs e de reguladores de *splicing* (ESRs, *Exonic Splicing Regulators*) em relação aos éxons constitutivos. Além disso, vimos que os ESRs de alternativos apresentam mais SNVs do que os ESRs de constitutivos. Analisando um subconjunto de SNVs cujos alelos apresentam associação estatística com variantes de *splicing* nós observamos que um tipo específico de ESR, os inibidores de *splicing*, são encontrados em frequência relativamente alta. Nossos resultados estabelecem uma associação entre diversidade genética intrapopulacional e *splicing* alternativo mediado por inibidores de *splicing*.

Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders

Jorge E. S. de Souza^{1#}, Rodrigo F. Ramalho^{2#}, Pedro A. F. Galante¹, Diogo Meyer² & Sandro J. de Souza^{1*}

1- Ludwig Institute for Cancer Research at Hospital Alemão Oswaldo Cruz, São Paulo, Brasil

2- Instituto de Biociências, Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo

# these authors contributed equally to this work

*To whom correspondence should be addressed at:

Rua João Julião

São Paulo, SP, 01509-010, Brasil

Phone: +55-11-33883211

Fax: +55-11-31411325

Email:sandro@compbio.ludwig.org.br

Abstract

With the availability of a large amount of genomic data it is expected that the influence of Single Nucleotide Variations (SNVs) in many biological phenomena will be elucidated. Here, we approached the problem of how SNVs affect alternative splicing. First, we observed that SNVs and Exonic Splicing Regulatory sequences (ESRs) independently show a biased distribution in alternative exons. More importantly, SNVs map more frequently in ESRs located in alternative exons than in ESRs located in constitutive exons. By looking at SNVs associated with alternative exon/intron borders (by their common presence in the same cDNA molecule), we observed that a specific type of ESR, the Exonic Splicing Silencers (ESSs), are more frequently modified by SNVs. Our results establish a clear association between genetic diversity and alternative splicing involving ESSs.

Introduction

The large amount of data on the human transcriptome has allowed several studies that, without exception, show a high prevalence of alternative splicing in the human transcriptome (Mironov, *et al.*, 1999; Xu, *et al.*, 2002; Modrek & Lee, 2003). The fact that most human genes undergo alternative splicing has raised doubts about the biological significance of most of the variants. One possibility is that a significant fraction of all variants are spurious products of the splicing machinery, without any functional relevance. Indeed, there are significant differences (e.g. the preservation of codon reading frame) between splicing variants that are conserved between human and mouse, and

therefore deemed as functional, and those that are not, suggesting that a fraction of the splicing variants are spurious products (Resch, *et al.*, 2004). Some authors have even suggested that these products would have functional implications by down-regulating the expression of functional variants (Lewis, *et al.*, 2003). On the other hand, some have argued that most of the splicing variants are products of a regulated process. For instance, Wang *et al.*, (2008) observed that most of the splicing variants of human genes show differential expression among different tissues whereas variation between individuals was approximately twofold to threefold less common. These results corroborate the hypothesis that alternative transcripts could have tissue specific functionalities. It has also been shown by our group that intron retention events are not randomly distributed regarding several parameters, again suggesting the notion that the majority of these splicing variants are not spurious and their expression is somehow regulated (Sakabe, *et al.*, ; Galante, *et al.*, 2004). Based on the current evidence it is reasonable to speculate that at least one third of all splicing variants are products of regulated expression.

In addition to the importance of 5'/ 3' splicing sites, branch point and polypyrimidine tracts in the control of splicing, there are several known *cis*-regulatory splicing elements that contribute to the splicing process and are located in intronic or exonic regions (for a review see (Pagani & Baralle, 2004; Wang & Burge, 2008). Many lines of evidence suggest that these elements can act by stimulating (enhancing) or inhibiting (silencing) the inclusion of the respective exon, or the neighbor exon, in the mature RNA transcript. These features are taken into account for the nomenclature of splicing regulatory elements. Those present in exons and with the capacity of enhancing splicing

are called exonic splicing enhancers (ESE) and those with the capacity of inhibiting the splicing are the exonic splicing silencers (ESS). Generally, these classes of elements are called exonic splicing regulators (ESRs).

Several studies suggest that ESS have a significant role in the control of alternative splicing. For example, a) using a set of paralogous exons, where one copy showed constitutive splicing and the other alternative splicing, Zhang Z. *et al.*, (2009) found that the alternative copy had significantly lower ESE and higher ESS densities than the constitutive copy; b) using designed exons constructed by random ligation of ESEs, ESSs, and neutral sequences, Zhang X., *et al.*, (2009) showed that negative correlation between ESS density and inclusion rate was stronger than the positive correlation between ESE density and inclusion rate; c) The set of *motifs* which bind the tissue-specific splicing factors Nova1 and Nova2 can act as ESEs or ESS depending on their position in the primary transcript. When located in alternative exons they mainly act as silencers (Ule, *et al.*, 2006).

In the present study we used single nucleotide variation (SNV) and cDNA data to compare the genetic diversity of ESRs located in constitutive and alternative exons. By establishing an association between the SNV alleles and distinct borders of alternative exons our results show that variations in ESSs, and not in ESEs, are more commonly associated with alternative splicing.

Materials and Methods

Public data

We obtained genomic (build 36.1) and cDNA (mRNAs and ESTs) sequence data from UCSC Genome Browser (<http://genome.ucsc.edu/>, files: [mrna.fa.gz](#) and [est.fa.gz](#)). Additional sequences were obtained from NCBI Reference Sequence Project (<http://www.ncbi.nlm.nih.gov/RefSeq>, release 22). We also downloaded EST libraries annotation from eVOC (<http://www.evocontology.org/>).

Identification of splicing events

A catalog of all splicing variants reported by the alignment of cDNAs on the human genome was obtained as previously described (Sakabe, *et al.*, ; Galante, *et al.*, 2004; Kirschbaum-Slager, *et al.*, 2005; Galante, *et al.*, 2007). Briefly, the coordinates of exon/intron borders for all cDNAs mapped onto the human genome were compared against each other to identify all splicing variants for all human genes. We used the software SIM4 (<http://globin.cse.psu.edu/html/docs/sim4.html>) for a more refined definition of the exon/intron borders. To increase the reliability of splicing events identified, we have chosen for further analysis only those events supported by at least two ESTs from two distinct libraries.

Exon classification

The definition of exonic and intronic regions was based on the genomic coordinates of cDNA sequences classified as “mRNA” in GenBank. Regarding alternative splicing, four groups of exons were defined for the current analyses:

exons reporting different donor or acceptor sites formed the Cryptic group (alternative splice site, 35391 exons), exons missing in two or more transcripts formed the Skipping group (46586 exons) and exons reporting an intron retention formed the Retention group (8310 exons). These three groups represent the major forms of alternative splicing. A fourth group, named Alternative (60383 exons), was formed by the union of the three groups of alternative exons mentioned above excluding redundant exons among these groups. Finally, the Constitutive group (70801 exons) was composed by exons for which no alternative exon/intron borders were detected.

Mapping the ESRs in exons

Eight different datasets of putative regulatory elements (6 ESEs and 2 ESS) were obtained from the literature (Liu, *et al.*, 1998; Liu, *et al.*, 2000; Fairbrother, *et al.*, 2002; Wang, *et al.*, 2004; Zhang & Chasin, 2004; Smith, *et al.*, 2006). Four (SF2_IgM, SRP40, SRP55 and SC35) out of six ESE datasets were discovered "in vitro" by using the SELEX methodology while the other two were discovered "in silico". Regarding the SELEX-ESEs, only those oligomers with a score equal or higher than the threshold scores defined by the original study were considered as ESEs. For the remaining datasets of ESEs (RESCUE and PESE), a list of ESE motifs was obtained from the supplementary material associated with the articles of Fairbrother *et al.*, (2002) and Zhang & Chasin, (2004). The PESS dataset of silencers was also obtained from Zhang & Chasin, (2004) and the dataset of ESS reported by Wang *et al.*, (2004) will be called ESS herein.

To identify the ESR motifs in the exons of our in-house database we perform a pairwise alignment between each set of ESRs and the exon sequences. The ESR counts were calculated independently for each group of exons analyzed in this study (Constitutive, Alternative, Cryptic, Skipping and Retention).

Mapping SNVs in exons

To make sure the SNVs were correctly indexed in our exon database, we mapped all 17804036 SNVs available in the dbSNP (release 130) in the genomic sequences used for our analysis. Relative positions of SNVs regarding exonic, intronic e intergenic regions were defined by comparing SNVs cDNA coordinates.

Mapping exonic SNVs in published ESRs motifs

Sequence tags comprising each SNV were generated by extracting from the reference human genome the corresponding variant nucleotide plus ten nucleotides flanking the SNVs (totaling a 21 nucleotide tag, which we will be called SNP-tag herein). We only extracted tags from the same strand orientation of the Refseq gene which contains the exon where the SNV was mapped. The alignment of these tags with the published ESR motifs defined whether or not the SNV was mapped into a known ESR.

Finding isoform-associated SNVs and defining putative ESR motifs

Alignments between all human mRNAs and the genome were searched for the presence of mismatches. The position of the mismatches was then compared to the genomic position of SNVs. These analyses resulted in 106271 mismatches that that co-occur with SNVs. For 96756 of these mismatches, the

discordant nucleotide reported by the mRNA corresponded to one of the alleles reported in the dbSNP for the respective SNV. Since the mRNA sequences are supposedly of high quality, this last number strongly suggests that the great majority of mismatches reported in the alignments are due to SNVs. Among these mismatches we selected 3533 SNVs where each allele was completely associated with alternative exon/intron borders (we refer to these as isoform-associated SNVs dataset). Considering all the alleles of these SNVs, we obtained 7087 sequence tags (17 tri-allelic and 2 quadri-allelic SNVs). SNVs that presented the same allele in cDNAs reporting different exon/intron borders were not included in the category of isoform-associated SNVs.

To make statistical inferences about a possible enrichment of known regulatory elements in the isoform-associated SNV dataset, we randomly created 1000 control datasets from a pool of 46336 SNVs also mapped in alternative exons but without alleles in complete association with alternative exon/intron borders. A schematic view of our approach is shown in Figure 1.

a) Isoform-associated SNVs

b) Control

Figure 1: Schematic representation of the computational approach used in this study to identify splicing regulatory elements. Figure1A: Schematic view of association between genetic and splicing variation. A *sine qua non* condition is the presence of both the SNV and the alternative exon/intron border of cDNA sequences. Complete association between an alternative exon/intron border and a SNV allele is required. Figure1B: schematic view of SNVs not associated with splicing variation, which defined the control set used in the simulation.

Distinct control datasets were created for the three main groups of alternative exons (Cryptic, Retention and Skipping). Each control dataset is of exactly the same size as the isoform-associated SNV dataset (1385 SNVs for Cryptic, 1780 SNVs for Retention, 958 SNVs for Skipping). Similarly, we create control datasets, with 3533 SNVs each, to use in the comparison with the main

group of alternative exons (which includes all three forms of alternative exons). Next, we used the same strategy described above to generate tags around the SNVs of the control datasets and search for known ESR motifs. For each replicate dataset we counted the number of ESRs that were part of an SNP-tag, and used the distribution of these values to test the null hypothesis that our set of isoform-associated SNVs are not associated with ESRs. The p-value was defined as the fraction of the ranked values observed in the control datasets which were greater than the observed value in the case set.

Defining the ancestral and derived alleles of SNVs

In order to establish the polarity of the ESR modification imposed by SNVs and so define events as gains, losses or maintenance/alteration of ESRs, we compared the alleles of human SNVs to the orthologous alleles of the chimpanzee (*Pan troglodytes*) genome. To perform this analysis we used data from the table “snp130OrthoPt2Pa2Rm2” available at the UCSC Genome Browser (<http://genome.ucsc.edu/>), which contains the orthologs of 11797184 human SNVs in 4 species of primates, including the chimpanzee.

Statistical test

For all comparisons of proportions presented in this manuscript we used the chi-square distribution to evaluate the statistical significance of the difference between the expected and observed values. We used the chi-square test implemented in the function *prop.test* of R statistical software (<http://www.r-project.org/>).

Results and Discussion

a) Alternative exons are enriched in SNVs when compared to constitutive exons

We first compared the density of SNVs between constitutive and alternative exons. Alternative exons, when taken as an unique group, have approximately 10% more SNVs than constitutive exons (5.09 and 4.52 SNVs per 1000 nucleotides for alternative and constitutive exons respectively, $p = 2.53 \cdot 10^{-2}$, chi-square test).

Comparisons among three sub-groups of alternative exons reveals that the Skipping group show a significantly lower SNV density than the other two groups. (Skipping=5.01 versus Cryptic=5.19 and Retention=5.17 SNVs per 1000 nucleotides, $p\text{-value} < 3.81 \cdot 10^{-9}$ for both comparisons). This reduced genetic diversity of skipped exons in relation to other forms of alternative exons may reflect a stronger selective constraint. This result is in accordance with the findings from Wang, Sandberg *et al.*, (2008), who showed that skipped exons are more conserved among four mammalian genomes and seem to be most important in tissue-specific alternative splicing.

b) Alternative exons are enriched in ESRs when compared to constitutive exons

Next, the density of ESRs was compared between constitutive and alternative exons. Table 1 shows that ESR motifs are enriched in the group of alternative exons. RESCUE-ESEs and PESEs were the exceptions, presenting an opposite trend. These exceptions are, in fact, expected since RESCUE-ESEs are identified from a set of constitutive exons (Fairbrother, *et al.*, 2002)

and PESEs are identified from a set of exons with high inclusion levels (Zhang & Chasin, 2004).

Table 1: Comparison of ESR density between constitutive and alternative exons.

ESR	Constitutive (10,650,372)	Alternative (22,622,280)	% change[#]	p-value*
RESCUE	0.11305	0.08943	-20.8	0
SF2_IgM	0.05466	0.06135	12.2	0
SC35	0.04156	0.04534	9.0	0
SRP40	0.04275	0.04400	2.9	⁻⁶¹ 6.97
SRP55	0.02514	0.02530	0.6	0.01
PESE	0.07030	0.06398	-8.9	0
PESS	0.01430	0.01916	33.9	0
ESS	0.00006	0.00013	116.6	⁻⁵⁹ 2.76

*p-value for bitailed chi-square test. Between brackets is the total number of nucleotides analyzed in each exon group. #Approximate fold change of Alternative compared with Constitutive exons. Positive and negative values represents excess and depletion respectively.

Exons belonging to the Skipping group showed a significant depletion in the density of SELEX-ESEs and ESSs when compared to the other groups of alternative exons (Figure 2).

Figure 2: Comparison of ESR density between the 3 main types of alternative exons. Values represent the $\log_{10}$ of the ratio between each individual group of alternative exons and the Alternative group (pool of the three types).

These results corroborates the results of Kurmangaliyev & Gelfand, (2008), who observed similar results in a comparison between skipped exons and exons with alternative splicing sites with mutations in their splice sites. However, they contradict the findings from Wang J. *et al.*, (2005), who performed a similar analysis and found a significantly lower density of SELEX-ESEs in skipped exons when compared to constitutive exons. We believe that the discrepancy between our result and those of Wang J. *et al.*, (2005) may be due to differences in the group of constitutive exons used in both studies.

Currently, the coverage of the human transcriptome is deeper when compared to 2005, and probably a significant proportion of exons defined as constitutive in his work is currently defined as alternative.

c) Alternative exons show higher proportion of ESRs modified by a SNV than constitutive exons

We next compared the proportion of ESRs modified by a SNV in both constitutive and alternative exons. Generally, ESRs in alternative exons are proportionally more modified by a SNV than those ESRs in constitutive exons (p-value < 1.26^{-16} , chi-square test, highest significant p-value from Table 2).

Table 2: Proportion of ESRs affected by SNVs in both Constitutive and Alternative exons.

ESR	Constitutive	Alternative	% change [#]	p-value*
RESCUE	0.03900	0.04614	18.3	⁻²⁰³ 1.72
SF2_IgM	0.07658	0.08064	5.3	⁻²² 7.82
SC35	0.07341	0.08319	13.3	⁻⁸⁹ 3.95
SRP40	0.06326	0.06889	8.9	⁻³⁶ 2.69
SRP55	0.06240	0.06721	7.7	⁻¹⁶ 1.26
PESE	0.05819	0.06800	16.8	⁻¹⁷³ 6.82
PESS	0.05858	0.06724	14.7	⁻³² 4.87
ESS	0.12006	0.11569	-3.6	0.8

Proportions were obtained by dividing the number of ESRs affected by a SNV by the total of ESRs, within each group. *p-value for bitailed chi-square test. #Approximate fold change of Alternative compared with Constitutive exons. Positive and negative values represent excess and depletion respectively.

The only exception was for ESS. Despite the fact that PESE and RESCUE sets are enriched in constitutive exons, we also observed a higher proportion of these motifs mapped in SNVs of alternative exons than in SNVs of constitutive exons. Moreover, these two sets of ESRs showed the most significant differences in comparisons between constitutive and alternative exons (Table 2). Consistent with our previous observations, the Skipping group showed the lowest proportion of ESRs affected by SNVs among the groups of alternative exons (Table S5).

The observations that the alternative exons show a higher density of SNVs, ESRs and also a higher proportion of ESRs modified by SNVs, suggest that this genetic variation could to some extent be one of the causal factors distinguishing alternative and constitutive splicing. In fact several studies analyzed the impact of single nucleotide polymorphism in the regulation of transcript isoform expression in tissue-specific and non-specific manners (Stranger, *et al.*, 2007; Ge, *et al.*, 2009; Kwan, *et al.*, 2009) and validated some causative SNPs occurring in splicing regulators (Coulombe-Huntington, *et al.*, 2009).

d) ESS associated with alternative splicing are more modified by SNVs

We decided to further explore this putative association between the SNVs and alternative splicing by examining those cDNAs that reported both an alternative exon/intron border and known SNV. Two categories of SNVs were used. The first category contains SNVs with alleles in complete association with different exon/intron borders (isoform-associated SNVs). The second category, the control set, contains the remaining SNVs mapped to alternative exons, i.e.,

those SNVs without a complete association with alternative exon/intron borders (Figure 1).

Among the set of sequence tags derived from the isoform-associated SNVs we found that approximately 86% contained at least one ESR already described. Is there an enrichment of any particular type of ESR in this set of sequence tags in comparison to tags derived from SNVs not associated with alternative borders? To answer this question, resampling was performed comparing the isoform-associated SNV dataset to 1000 control datasets, each comprised of the same number of SNVs but not associated with alternative exon/intron borders. The analysis was performed independently for each set of published ESRs using alternative exons either as a unique group or divided into the three categories previously discussed.

Table 3 shows that isoform-associated tags are enriched in ESSs in comparison with the control dataset when the alternative exons are analyzed as a unique group. This is true for both sets of ESS analyzed (PESS and ESS). Moreover, the isoform-associated SNVs is significantly depleted in ESEs when compared to the replicate datasets.

Table 3: Difference in ESR densities between experimental and control sets of SNVs.

ESR	exon group	p-value
Enhancers		
RESCUE	Skipping	0.74
	Retention	0.87
	Cryptic	0.98 ^a
	Alternative	1 ^a
SC35	Skipping	0.95 ^a

	Retention	1 ^a
	Cryptic	0.85
	Alternative	1 ^a
SRP40	Skipping	0.86
	Retention	0.99 ^a
	Cryptic	0.87
	Alternative	0.9
SRP55	Skipping	0.19
	Retention	0.86
	Cryptic	0.44
	Alternative	0.46
PESE	Skipping	0.11
	Retention	1 ^a
	Cryptic	1 ^a
	Alternative	1 ^a
SF2_IgM	Skipping	0.09
	Retention	1 ^a
	Cryptic	0.93
	Alternative	1 ^a
Silencers		
PESS	Skipping	0.92
	Retention	0 ^b
	Cryptic	0 ^b
	Alternative	0 ^b
ESS	Skipping	0.1
	Retention	0.09
	Cryptic	0.01 ^b
	Alternative	0 ^b

^a: significantly lower than control, ^b: significantly higher than control.

Interestingly, Zhang X. *et al.*, (2009), showed that the absolute number of ESS correlates significantly ($R^2=0.78$, $p < 5^{-47}$) with the non-inclusion rate (negative correlation with inclusion rate) of exons, when other splicing signals are constant. Moreover they found a significant positive correlation ($R^2=0.53$, P

$<3e-6$) between inclusion level and the ratio ESE/ESE+ESS. Consistent with this, we found this ratio to be significantly lower in the SNP-tags of our experimental dataset when compared to the SNP-tags of control datasets (data not shown). Together, these results suggest that the influence of SNVs on some types of alternative splicing occur predominantly through their effects on ESSs.

A recent study by Woolfe *et al.*, (2010), analyzed a small set of well curated SNPs (a total of 87) associated to exon skipping, which they compared to a large set HapMap SNPs which were putatively neutral with respect to splicing. Using an approach different from ours, they also found that alterations of ESSs were significantly overrepresented when compared to alterations which are putatively neutral with respect to splicing. Moreover, they also found that the degree of ESS alterations was even greater for events of alternative splice site than that for exon skipping. The concordance between these two studies, which used different approaches to define the association between SNVs and splicing variants, corroborates the important role played by ESSs in the splicing regulation.

Analyzing the polarity of the ESRs changes imposed by SNVs

Can we further discriminate the effect of SNVs in ESSs? If we assume that the derived allele increases transcriptome variability by allowing the use of alternative exon/intron borders, we can try to better understand the effect of SNVs on ESS by defining a pattern of ESS gain or loss with the emergence of a derived allele. To this end we defined the polarity of change by assuming the reference chimpanzee genome as the ancestral allele, as done by others (Fairbrother, *et al.*, 2004).

Results in Table 4 confirm that exon skipping does not seem to be primarily regulated by SNVs that create ESSs. When we independently analyzed the events of ESS gain from non-ESR motifs, the difference between case and control sets does not exist (11 SNVs in the isoform-associated set against 12 SNVs in the control Set, p-value=0.56). The difference is restricted to those events of ESS gain from an ancestral ESE (29 SNVs in the case set against 44 SNVs in the control set, p-value=0.99). This suggests that the significant depletion of SNVs involved in ESS gain observed for this type of alternative splicing reflects depletion in the number of SNVs that affect ESEs.

Table 4: SNP counts of ESS loss, gain and maintenance found in isoform-associated SNV data set and control data set.

		Isoform-associated SNV set	# Control	p-value
Skipping	ESS loss	38	23-63	0.67
	ESS gain	40	40-57	0.99 ^a
	ESS maintenance	22	(10-45)	0.6
Cryptic	ESS loss	81	39-87	0.005 ^b
	ESS gain	68	51-115	0.97 ^a
	ESS maintenance	54	19-60	0.005 ^b
Intron Retention	ESS loss	111	63-115	0.01 ^b
	ESS gain	108	75-140	0.53
	ESS maintenance	90	43-93	0 ^b

range for 1000 replicate datasets, a: significantly lower than the control, b: significantly higher than the control.

These findings differ from those of Woolfe *et al.*, (2010), and show that the mechanism of splicing regulation among the skipped exons is more complex than just an increase in the proportion of ESS gains. We note, however, that

these authors compared SNVs alleles associated to skipped exons to a group of SNVs alleles belonging to a heterogeneous set of exons. This differs from our approach, in which the isoform-associated and control SNVs alleles were all from skipped exons, and may explain the differences between the studies.

For cryptic and intron retention, the predominant pattern involves loss and alteration/maintenance of ESS. Based on the significant frequency of ESS loss, we predict that the derived allele could be generating a decrease in ESS strength in those cases where ESS is maintained.

Final Remarks

The results reported here support the view that ESRs have a higher genetic diversity in alternative exons when compared to constitutive exons. We believe that this genetic variation could to some extent be one of the major features distinguishing alternative from constitutive splicing. Furthermore, we provide evidence that this effect is mainly due through SNVs acting on ESS.

A caveat of our approach is that we cannot directly distinguish between causal and associated SNVs since an isoform-associated SNV may be in linkage disequilibrium with a different causal variant. However, our re-sampling analysis addresses this issue by examining if the isoform associated SNVs are associated to ESRs as frequently as non-isoform associated SNVs (used as a “control”). Using this approach we were able to show that ESS are significantly overrepresented among isoform-associated SNVs, supporting their functional role in splicing regulation.

The emergence of next generation sequencing is beginning to provide a huge amount of both genomic and expressed sequence data. We believe that

the strategy used in this manuscript will be very useful in the next few years to further explore the role of SNVs in alternative splicing.

Funding

This work was supported by a research grant from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) [2007/55790-5 to SJS] and by a Ph.D fellowship [2007/59721-8 to RFR].

Acknowledgements

We thank Daniel Ohara for the technical support on computers.

References

- Coulombe-Huntington, J., K. Lam, et al. (2009). "Fine-scale variation and genetic determinants of alternative splicing across individuals." *PLoS Genet* **5**(12): e1000766.
- Fairbrother, W., R. Yeh, et al. (2002). "Predictive identification of exonic splicing enhancers in human genes." *Science* **297**(5583): 1007-1013.
- Fairbrother, W. G., D. Holste, et al. (2004). "Single nucleotide polymorphism-based validation of exonic splicing enhancers." *PLoS Biol* **2**(9): E268.
- Galante, P., N. Sakabe, et al. (2004). "Detection and evaluation of intron retention events in the human transcriptome." *RNA* **10**(5): 757-765.
- Galante, P., D. Vidal, et al. (2007). "Sense-antisense pairs in mammals: functional and evolutionary considerations." *Genome Biol* **8**(3): R40.
- Ge, B., D. Pokholok, et al. (2009). "Global patterns of cis variation in human cells revealed by high-density allelic expression analysis." *Nat Genet* **41**(11): 1216-1222.
- Kirschbaum-Slager, N., R. Parmigiani, et al. (2005). "Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data." *Physiol Genomics* **21**(3): 423-432.
- Kurmangaliyev, Y. and M. Gelfand (2008). "Computational analysis of splicing errors and mutations in human transcripts." *BMC Genomics* **9**: 13.
- Kwan, T., E. Grundberg, et al. (2009). "Tissue effect on genetic control of transcript isoform variation." *PLoS Genet* **5**(8): e1000608.
- Lewis, B. P., R. E. Green, et al. (2003). "Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans." *Proc Natl Acad Sci U S A* **100**(1): 189-192.
- Liu, H. X., S. L. Chew, et al. (2000). "Exonic splicing enhancer motif recognized by human SC35 under splicing conditions." *Mol Cell Biol* **20**(3): 1063-1071.

- Liu, H. X., M. Zhang, et al. (1998). "Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins." *Genes Dev* **12**(13): 1998-2012.
- Mironov, A. A., J. W. Fickett, et al. (1999). "Frequent alternative splicing of human genes." *Genome Res* **9**(12): 1288-1293.
- Modrek, B. and C. J. Lee (2003). "Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss." *Nat Genet* **34**(2): 177-180.
- Pagani, F. and F. Baralle (2004). "Genomic variants in exons and introns: identifying the splicing spoilers." *Nat Rev Genet* **5**(5): 389-396.
- Resch, A., Y. Xing, et al. (2004). "Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation." *Nucleic Acids Res* **32**(4): 1261-1269.
- Sakabe, N., J. de Souza, et al. "ORESTES are enriched in rare exon usage variants affecting the encoded proteins." *C R Biol* **326**(10-11): 979-985.
- Smith, P., C. Zhang, et al. (2006). "An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers." *Hum Mol Genet* **15**(16): 2490-2508.
- Stranger, B., A. Nica, et al. (2007). "Population genomics of human gene expression." *Nat Genet* **39**(10): 1217-1224.
- Ule, J., G. Stefani, et al. (2006). "An RNA map predicting Nova-dependent splicing regulation." *Nature* **444**(7119): 580-586.
- Wang, E., R. Sandberg, et al. (2008). "Alternative isoform regulation in human tissue transcriptomes." *Nature* **456**(7221): 470-476.
- Wang, J., P. Smith, et al. (2005). "Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes." *Nucleic Acids Res* **33**(16): 5053-5062.
- Wang, Z. and C. Burge (2008). "Splicing regulation: from a parts list of regulatory elements to an integrated splicing code." *RNA* **14**(5): 802-813.
- Wang, Z., M. Rolish, et al. (2004). "Systematic identification and analysis of exonic splicing silencers." *Cell* **119**(6): 831-845.
- Woolfe, A., J. Mullikin, et al. (2010). "Genomic features defining exonic variants that modulate splicing." *Genome Biol* **11**(2): R20.
- Xu, Q., B. Modrek, et al. (2002). "Genome-wide detection of tissue-specific alternative splicing in the human transcriptome." *Nucleic Acids Res* **30**(17): 3754-3766.
- Zhang, X., M. Arias, et al. (2009). "Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing." *RNA* **15**(3): 367-376.
- Zhang, X. H. and L. A. Chasin (2004). "Computational definition of sequence motifs governing constitutive exon splicing." *Genes Dev* **18**(11): 1241-1250.
- Zhang, Z., L. Zhou, et al. (2009). "Divergence of exonic splicing elements after gene duplication and the impact on gene structures." *Genome Biol* **10**(11): R120.

Capítulo 3

Sobre as origens do exon *skipping*: Polimorfismos fracamente deletérios criam inibidores de *splicing* em éxons constitutivos

On the origin of exon skipping: slightly deleterious polymorphisms create silencers in constitutive exons

Prólogo

No manuscrito, “On the origin of exon skipping: slightly deleterious polymorphisms create silencers in constitutive exons”, testamos a hipótese de que os éxons alternativos do tipo *skipped* se originam através de um relaxamento de seleção natural em éxons constitutivos. Embora existam evidências de que, em uma escala interespecífica, os inibidores de *splicing* (ESS) são evitados em éxons constitutivos, pouco se sabe sob o efeito de mutações fracamente deletérias nestes reguladores. Por apresentar um tamanho efetivo pequeno, a espécie humana está mais sujeita aos efeitos aleatórios da deriva genética na segregação das variações genéticas do que espécies com tamanhos efetivos maiores. Sob esse cenário espera-se que mutações fracamente deletérias possam segregar na população aparecendo como polimorfismos, porém não possam se fixar na espécie e aparecer como divergências.

Resumo

Através da abordagem filogenética, diversos estudos reportaram que os inibidores de *splicing* exônicos (*Exonic Splicing Silencers*, ESS) são evitados em éxons constitutivos por seleção natural. Entretanto, pouco se sabe a respeito dos polimorfismos fracamente deletérios que ocorrem nesses reguladores e seus efeitos no tipo de *splicing* dos éxons (constitutivo ou alternativo). Através da aplicação de uma modificação do teste McDonald-Kreitman (teste MK) nós comparamos a quantidade de polimorfismos humanos normalizada pela quantidade de substituições entre humano e rhesus que altera ou não altera os reguladores de *splicing* exônicos (ESRs) de éxons humanos. Para evitar o efeito de seleção natural que atuam sobre as proteínas usamos apenas mudanças sinônimas. Nossos resultados mostram que a categoria de mudança associada ao ganho de ESS em éxons constitutivos é fracamente deletéria. A existência de um excesso de SNPs desse tipo em relação às substituições sustenta esta hipótese. Além disso, nós demonstramos que esse padrão evolutivo também está presente em ESRs envolvidos na alteração do tipo de *splicing* de certos éxons de vertebrados que ao longo da evolução se tornaram alternativos na linhagem dos mamíferos. Devido à similaridade do padrão evolutivo entre estes dois conjuntos de ESRs propomos que a transição de éxons constitutivos para alternativos na linhagem dos mamíferos é mais frequentemente associada com a inibição do que com o fortalecimento dos sinais de *splicing*. Essa hipótese está de acordo com a origem constitutiva para os éxons alternativos do tipo *skipped* e corrobora resultados anteriores sobre a função antagônica de certos promotores de *splicing* exônicos (*Exonic Splicing Enhancers*, ESEs).

Title:

On the origin of exon skipping: slightly deleterious polymorphisms create silencers in constitutive exons with weak splice sites.

Publication type:

Research Article

Authors:

Rodrigo F. Ramalho¹, Sahar Gelfman², Jorge S. de Souza³, Gil Ast², Sandro J. de Souza³, Diogo Meyer¹

Affiliations:

¹Instituto de Biociências, Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, São Paulo, SP 05508-900, Brasil

²Department of Human Molecular Genetics & Biochemistry, Sackler Faculty of Medicine, Tel-Aviv University, Ramat Aviv 69978, Israel.

³Instituto de Bioinformática e Biotecnologia

⁴ Instituto do Cérebro da Universidade Federal do Rio Grande do Norte (UFRN)

Corresponding author:

Name: Rodrigo F. Ramalho

Address: Instituto de Biociências, Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, São Paulo, SP 05508-900, Brasil

Telephone number: (55-11-30918757)

e-mail: rfrusp@gmail.com

Keywords:

alternative splicing, human polymorphism, exonic splicing regulators, MK-test

Abstract

Despite some evidence that, at the interspecific scale, Exonic Splicing Silencers (ESS) are under negative selection in constitutive exons, little is known about the effects of slightly deleterious polymorphisms on these splicing regulators. Through the application of a modified version of the McDonald-Kreitman test (MK test), we compared the normalized proportions of human polymorphisms and human/rhesus substitutions which affect Exonic Splicing Regulators (ESRs) on sequences of constitutive and alternative exons. Our results show a depletion of substitutions and an enrichment of SNPs associated with ESS gain in constitutive exons. Moreover, we show that this evolutionary pattern is also present in a set of ESRs previously involved in the transition from constitutive to skipped exons in the mammalian lineage. This similarity between these two sets of ESRs suggests that the transition from constitutive to skipped exons in mammals is more frequently associated with the inhibition than with the promotion of splicing signals. This is in accordance with the constitutive origin of exon skipping hypothesis and corroborates previous findings about the antagonistic role of certain Exonic Splicing Enhancers (ESEs).

Introduction

Splicing is the process by which introns are removed from a mRNA precursor and exons are ligated to form a mature mRNA. During this process several *cis* and *trans* factors are involved. Besides the canonical *cis* factors (e.g., splicing sites, branch point and polypyrimidine tract), Splicing Regulators – short sequences located in exons and introns – have an important role in assisting the spliceosome to correctly recognize exon/intron boundaries. Exonic splicing regulators (ESRs) can be divided in two groups, according to their function in splicing: exonic splicing enhancers (ESEs) promote, and exonic splicing silencers (ESSs) inhibit the inclusion of exons in mRNA, respectively.

Through alternative splicing, great transcript diversity is generated, and currently it is known that more than 80% of human genes present splicing variants (Wang *et al.*, 2008). There are three main hypotheses for the origin of alternative exons: (a) “exonization” of Alu elements (Sorek *et al.*, 2002), (b) gene/exon duplication (Letunic *et al.*, 2002) and (c) the weakening of purifying selection in the splicing *cis*-elements of constitutive exons (Lev-Maor *et al.*, 2007). Among these, only the latter is able to explain the high frequency of alternative exons observed in the genomes of various eukaryotes, especially in mammals (Kim *et al.*, 2007), while the other two can only explain a minority of alternative exons in the human genome because less than 5% of these contains Alus (Sorek *et al.*, 2002) and less than 20% are associated with duplicated exons (Letunic *et al.*, 2002).

The evolutionary mechanism underlying the hypothesis of a constitutive origin of alternative exons was inferred combining two sources of information:

differences in the splicing pattern of orthologous genes in mammals and non-mammals, and the phylogenetic position of modification in cis-splicing signals (Lev-Maor *et al.*, 2007). These authors observed that the transition from constitutive splicing in non-mammals to exon skipping in mammals is based on the weakening of splice site strength, associated with the fixation of certain Exonic Splicing Regulatory (ESR) motifs. Based on this result, a model was proposed, predicting that the weakening of the 5' splice site is the main driving force in the transition from constitutive to alternative splicing and that following this weakening, there is a gain of ESR functionality to properly regulate the exon inclusion level (Lev-Maor *et al.*, 2007). Although this hypothesis attributes to the fixation of ESRs an important functional role in the switching of splicing pattern, it does not clearly define the major function of these ESRs. Do they mainly act as splicing silencers, inhibiting the exon inclusion, or as splicing enhancers, promoting exon inclusion?

Despite this uncertainty, there is a strong pattern that gives a clue to the answer: several studies suggest that the lack of Exonic Splicing Silencers (ESS) is characteristic of constitutive exons (Ke *et al.*, 2008; Wang *et al.*, 2004; Xiao *et al.*, 2007; Zhang *et al.*, 2009). Wang *et al.*, (2004) showed that ESS are significantly depleted in constitutive exons relative to introns and also depleted in constitutive exons with weaker splice sites relative to constitutive exons with strong splicing sites. Similar conclusions were reported by Xiao *et al.*, (2007).

A second line of evidence, supporting the importance of ESS density in defining whether an exon will be constitutive or alternative was provided by the analysis of duplicated genes (Zhang *et al.*, 2009). Using pairs of paralogous exons that exhibit a constitutive and an alternative copy, these authors showed

that the density of ESS is lower in the constitutive copy than in the alternative one.

Further evidence of purifying selection on ESSs in constitutive exons was described by Ke *et al.*, (2008). Using synonymous divergence between human and macaque, these authors showed that ESS creation occurs at a much lower rate in constitutive exons than in introns. Additionally, the rate of ESS creation in constitutive exons was lower than the rate of creation of non-ESS motifs for this same set of exons. Overall, this evidence suggests that a gain in ESSs could explain the transition from constitutive to alternative.

Assuming that the presence of ESSs in constitutive exons is essential for promoting exon skipping, the above evidence suggests that the ESSs fixed in skipped exons after the split between mammals and non-mammals should act as silencers, downregulating exon inclusion level.

In the present study, we investigate the evolution of hexamers supposedly involved with constitutive/alternative transition throughout vertebrate evolution (Gelfman *et al.*, 2012; Lev-Maor *et al.*, 2007). Our approach differs from other studies because we directly contrasted the evolutionary rate of these splicing regulators at populational and inter-species levels. The use of SNP data compared to substitutions, provides an opportunity to investigate how evolution occurs in a relative recent time scale (within human population). The nearly-neutral model of molecular evolution (Ohta, 1973) predicts that slightly deleterious mutations are more prone to accumulate within populations, but over long spans of time are removed by natural selection with a high probability. Therefore, the comparison of DNA polymorphisms with substitutions allows us

to test the hypothesis that skipped exons are accumulating slightly deleterious mutations and thus weakening the cis splicing signals of constitutive exons. Specifically, we test if the hexamers identified in human skipped exons and fixed in mammals present a significant excess of polymorphism relative to substitutions when located in human constitutive exons. Assuming that the mammalian fixed hexamers from skipped exons act as silencers, we expected to detect stronger signal of purifying selection on mutations that create these ESRs in constitutive exons. Such a pattern would indicate that genetic changes create favorable conditions for the transition from constitutive to alternative splicing.

By using a modified version of the McDonald-Kreitman test (McDonald and Kreitman 1991) we find that the hexamers associated with the origin of exon skipping in mammals (called rate-shift motifs herein) present a depletion of substitutions and an excess of polymorphisms when located in constitutive exons. This evolutionary pattern, typical of loci under purifying selection, reaches a higher statistical significance in constitutive exons with weak splice sites.

Based on the similar pattern found here with that observed for previously known ESS, we suggest that these hexamers might have inhibitory activity on exon inclusion, corroborating previous results about the context-dependent function of ESRs (Goren *et al.*, 2006; Jumaa and Nielsen 1997; Kanopka *et al.*, 1996; Solis *et al.*, 2008; Ule *et al.*, 2006).

Materials and Methods

Primary exon database

For all the data mining described below we started the procedures using an in-house exon database consisting of 60,383 internal alternative exons (including exons with alternative splice site and skipped) and 70,801 internal constitutive exons (de Souza *et al.*, 2011; Galante *et al.*, 2004; Galante *et al.*, 2007). Briefly, the definition of exonic and intronic regions was based on the genomic coordinates of cDNA sequences classified as 'mRNA' in GenBank. All the analyzed genes have at least one RefSeq identification code. To increase the reliability of splicing events identified, we chose for further analysis only those events supported by at least two ESTs from two distinct libraries using information from Evoc database (<http://www.evocontology.org/>) (see Supplemental material).

Polymorphism dataset

We obtained the SNP data from the low coverage pilot phase of the 1000 Genomes project (Consortium 2010; Wang *et al.*, 2010). We chose the African sample since it has more SNPs than other samples increasing the power of our tests. The SNP data can be downloaded from ftp://ftptrace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2010_07/low_coverage/snps/. We use the software Annovar (Wang *et al.*, 2010) to annotate all SNPs and filter the synonymous ones for further analysis. In total, 37,080 synonymous SNPs were identified. We identified 2,277 internal skipped exons and 10,232 internal constitutive exons with at least one synonymous SNP from the 1000 Genomes Project data.

Substitution dataset

We downloaded the genomic alignments of human and rhesus (axtNet format) from the UCSC genome browser (<http://hgdownload.cse.ucsc.edu/>). This file can be obtained at <http://hgdownload.cse.ucsc.edu/goldenPath/hg18/vsRheMac2/>).

Using a local Perl script we parsed this file to find the chromosome, the genomic coordinate and the nucleotide for each divergent site between the two species. Approximately 180 million substitutions were found for the whole genomes and approximately 650,000 for the human exome.

We used the Annovar software to annotate the substitutions, entering the divergent nucleotides found at each site as alleles from a biallelic SNP.

We identified 1,581 orthologous internal skipped exons and 7,575 orthologous internal constitutive exons containing, at least, one synonymous substitution (between human and rhesus) and one synonymous SNP.

Prediction of splicing cis-regulatory elements

Starting from 28,970 sequences from skipped exons we apply the strategy described in Gelfman *et al.*, (2012) to identify functionally important hexamers for splicing regulation.

Briefly, the method uses a multiple alignment with sequences from twelve vertebrates to identify hexamers enriched around orthologous nucleotide sites which experienced fixation in one group (mammals) but vary in four non mammalian species. The assumption of the method is that a significant

decrease in evolutionary rate reflects a functional constraint in a specific clade (Pupko and Galtier 2002).

We applied the same strategy to identify rate-shift hexamers present in constitutive exons. The resulting set was also tested for signals of natural selection and serves as control in comparisons against hexamers identified from skipped exons.

All the hexamers identified by this phylogenetic approach (from both constitutive and skipped exons) will be called rate-shifted motifs herein.

ESR datasets

We used a Perl script to search for exact matches between the rate-shifted hexamers, six ESE sets (SF2_IgM, SRP40, SRP55 and SC35, RESCUE, PESE) and one ESS set (FAS-hex2). The ESE sets SF2_IgM, SRP40, SRP55 and SC35 were discovered *in vitro* by SELEX procedure (Liu *et al.*, 1998; Liu *et al.*, 2000) and each contains binding sites for four distinct SR-proteins. Regarding the SELEX-ESEs, only those oligomers with a score equal or higher than the threshold scores defined by the original study were considered as ESEs. The other two ESE sets (RESCUE and PESE), were predicted *in silico*. Motifs that make up the RESCUE set were computationally identified based on their enrichment in a set of constitutive exons with weak splice site relative to constitutive exons with strong splice site, and also to introns (Fairbrother *et al.*, 2002). The PESE set contains motifs enriched in constitutive exons relative to pseudoexons and 5' UTR of intronless genes (Zhang and Chasin 2004).

The ESS set was discovered by an experimental procedure based on splicing in cultivated cells and a fluorescent system which reports the silencer role of random motifs. For ESSs, we used the list of 176 Fas-hex2 hexamers with silencer function (Wang *et al.*, 2004).

Splice site scoring

We scored the splice site of constitutive exons using the MaxEntScan webserver:http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html according to the method described by Yeo and Burge (2004). The 3' splice site (3' ss) was defined as the twenty intronic nucleotides upstream to the exon and the first three exonic nucleotides, whereas the 5' splice site (5' ss) was defined as the three terminal exonic nucleotides and the first six downstream intronic nucleotides. We use the median of the splice-site score distribution for constitutive exons (8.5) as a threshold to define weak and strong splice-sites.

Mapping SNPs and divergent sites onto ESR motifs

For each of the polymorphic or divergent sites, we created sequence tags by extracting the sequence of the site and ten flanking nucleotides from each side from human genome (assembly hg18). The tags were extracted with respect to the same strand orientation of the RefSeq gene which contains the SNP. Next, each tag was tested for its status as containing ESRs by aligning it with known ESR sets.

Defining the ancestral and derived alleles of SNPs

In order to distinguish the SNPs that create or disrupt rate-shifted hexamers, we compared the SNP alleles to the orthologous sites of the chimpanzee (*Pan troglodytes*) and rhesus macaque (*M. mulatta*) genomes. To perform this, we used a Perl script that reads the genomic alignment between human and rhesus to retrieve the rhesus orthologous nucleotides for each SNP given their genomic coordinates. The SNP data which we use (provided by the 1000 genome project) already contains the annotation of ancestral allele related to the chimpanzee genome. Only SNPs with ancestral alleles identical to orthologous positions of rhesus and chimpanzee genomes were further analyzed. In this way, ancestral and derived tags (around the SNPs) were defined, and the polarity of the change (i.e., creation or disruption of rate-shift hexamers) was determined.

Modified McDonald-Kreitman test (MK test)

The usual implementation of the MK test uses data from SNPs and substitutions to compare, through a 2X2 contingency table, the proportion of variants from two distinct functional categories (for example, synonymous and non-synonymous changes). One of these category is assumed to be evolving neutrally, allowing a formal test with respect to the other. In our modified version of this test we compare mutations that create ESRs (which are supposedly under purifying selection) with those that do not alter the ESRs' status (which are putatively neutral). We avoid the confounding effects of selection at the aminoacid level by only using synonymous variants (for substitutions and polymorphisms).

The null hypothesis of homogeneity for contingency tables was tested using the chi-square distribution, implemented in the function 'chisq.test' of R statistical software (<http://www.r-project.org/>). For each cell of contingency table we calculated the normalized deviation, defined as (observed-expected)/(expected), from the expected proportions assuming independence among categories. The normalized deviation was used to graphically summarize the results of MK test.

Results and Discussion

The rate-shift analysis applied to sequences of human skipped exons revealed 145 hexamers (called rate-shifted_{skipped}, herein). Although our method probably captures motifs under purifying selection, it did not reveal their function as splicing regulators. Therefore, our first analysis was the alignment of these rate-shifted hexamers with known ESRs (supplementary Table S1). The great majority (116 hexamers; 80%) were found to be known ESRs and nine (6%) were identical to known ESSs. This result suggests that the rate-shifted_{skipped} motifs represent potential ESRs.

When applied to constitutive exons, the rate-shift analysis revealed 198 significant motifs (called rate-shifted_{constitutive}). Again, most of them (173 hexamers; 87%) fully aligned with known ESEs and four (2%) were identical to ESSs. Interestingly, a very low overlap was observed between the two sets of rate-shifted hexamers (constitutive vs. skipped exons); only 14 rate-shifted hexamers were identical between them (Online Resource 1).

We observed that the rate-shifted_{skipped} hexamers have a higher proportion of unknown motifs (those absent in all sets considered herein) than the rate-shifted_{constitutive} hexamers (29 (~20%) and 25 (~12%) unknown hexamers respectively). Moreover, the proportion of known ESSs among the rate-shifted_{skipped} hexamers was higher (~6%) than for rate-shifted_{constitutive} (~2%) (chi-square p-value ~ 0.04 for both tests; 20% vs. 12% and 6% vs. 2%; one-tailed test). These differences suggest that for skipped exons the evolutionary rate shifts affect more often ESRs with inhibitory than enhancing activity.

To further explore the tendency of silencer functionality for the rate shifted hexamers found in skipped exons we applied two approaches. First, we checked if our modified version of McDonald-Kreitman test could detect signals of selection on ESRs. To this end, we apply this test to previously known ESS, which were previously described as targets of negative selection when located in constitutive exons. Second, we applied the same test to rate-shifted_{skipped} motifs and compared the results of these two tests.

Using a modification of the McDonald-Kreitman test, we found a significant excess of SNPs that create ESSs relative to ESS creating substitutions, in constitutive exons (p-value < 0.05, Figure 1).

Figure 1 Normalized difference between the observed and expected counts of substitutions and polymorphisms which create ESS within sequences of human constitutive exons. Bars represent the enrichment (positive values) or depletion (negative values) of slightly deleterious related to the expectation based on supposedly neutral mutations. * Chi-square p-value <0.05

This result confirmed previous evidence that ESSs are under purifying selection on constitutive exons (Ke *et al.*, 2008; Wang *et al.*, 2004; Xiao *et al.*, 2007; Zhang *et al.*, 2009). In total, we tested 3,501 synonymous substitutions and 1,425 synonymous SNPs from constitutive exons (Table 1a).

In contrast, when we applied the MK test to substitutions and SNPs that create ESSs in skipped exons, the proportions of deleterious SNPs and substitutions were not significantly different (p-value=0.48, table 1b).

Table 1: Modified MK test applied for Exonic Splicing Silencers (ESS) located within (a) Constitutive exons and (b) Skipped exons. Only synonymous mutations that create or maintain Exonic Splicing Silencers (ESS) were used.

Table 1a: Constitutive exons

	CREATE-ESS	MAINTAIN-ESS
Substitution	2535	966
Polymorphism	1073	352
Chi-square p-value = 0.04		

Table 1b: Skipped exons

	CREATE-ESS	MAINTAIN-ESS
Substitution	462	190
Polymorphism	238	87
Chi-square p-value = 0.48		

This result is in agreement with a scenario in which ESSs from skipped exons are evolving according to the neutral model.

Consistent with the abovementioned result --that silencers are weakly deleterious in constitutive exons sequences-- we observed that the contingency table for ESS disruptions presented an opposite trend of deviation relative to ESS creations, i.e., substitutions are enriched relative to polymorphisms (supplementary Table S2). This result suggests that changes that maintain ESSs in constitutive exons are more deleterious than those that disrupt them.

Next, we applied the modified version of the MK test to look for signatures of natural selection on the 145 rate-shifted_{skipped} hexamers. Note that these putative ESRs are expected to be involved in the regulation of alternative

splicing in mammals (Lev-Maor *et al.*, 2007). For comparison purposes, we analyzed these ESRs in two distinct contexts; a) when located in skipped exons, b) when located in constitutive exons.

Similarly to what was observed for previously known ESSs, the MK test applied to SNPs and substitutions associated with creation of these putative ESRs showed a significant excess of polymorphisms with respect to substitutions (Figure 2) but not in skipped exons (Table 2).

Figure 2 Normalized difference between the observed and expected counts of which creates rate-shifted_{skipped} ESRs within sequences of human constitutive exons. * Chi-square p-value <0.05

Table 2: Modified MK test applied for rate-shifted ESRs identified from sequences of human skipped exons (rate-shifted_{skipped} ESRs) located within (a)

Constitutive exons and (b) Skipped exons. Only synonymous mutations that create or maintain rate-shifted_{skipped} ESRs were used.

Table 2a: Constitutive exons

	CREATE-rate-shifted _{skipped}	MAINTAIN- rate-shifted _{skipped}
Substitutions	4779	1327
Polymorphism	1759	424
Chi-square p-value = 0.02		

Table 2b: Skipped exons

	CREATE- rate-shifted _{skipped}	MAINTAIN- rate-shifted _{skipped}
Substitutions	934	246
Polymorphism	387	93
Chi-square p-value = 0.5		

This result is consistent with our hypothesis that ESRs important for regulation of human skipped exons act mainly as silencers and hence are under purifying selection in human constitutive exons. We next investigated if the above signature of purifying selection for constitutive exons reflects differences in nucleotide composition between constitutive and skipped exons, and is not necessarily associated with splicing regulation. We checked this by analyzing the set of 198 rate-shifted_{constitutive} hexamers. Assuming the existence of a nucleotide composition bias, we expect that this set of motifs (identified from constitutive exons) should present a signal of purifying selection when located in skipped exons. However, this was not the case, since the MK tests applied to synonymous variant sites located in this set did not show significant signals of natural selection either in constitutive or skipped exons (Tables S3). This result

suggests that for constitutive exons the events of evolutionary rate-shifts captured by our method are probably caused by selection at the protein level. This could explain the lack of evidence of natural selection using synonymous sites. Therefore, we conclude that our method of motif identification could not, by itself, explain the observed signal of purifying selection.

Notably, the deviation from neutral expectation observed for rate-shifted_{skipped} hexamers located in constitutive exons, is very small (Figure 2). In fact, this observation is not surprising given that we analyzed only synonymous single nucleotide variations (SNPs and substitutions), which evolve in a predominantly neutral manner. We next examined if the fact of a significant p-value for constitutive exons (and not for skipped) was a consequence of a higher absolute number of SNPs and substitutions within this category, leading to increased power in the statistical test. If this were true, a reduction in the absolute number of single nucleotide variants from this exon category would diminish the statistical significance.

To this end, we subdivided the set of constitutive exons according to an important biological feature which governs the splicing process - the strength of the splice sites. Our intent was to investigate if the accumulation of slightly deleterious polymorphisms observed for the whole set of constitutive exons was in any way associated with splice site strength.

Generally, constitutive exons have stronger splice sites than alternative exons (Clark and Thanaraj 2002; Shepard *et al.*, 2011). This implies that ESRs are less necessary in constitutive than alternative exons. However, among the constitutive exons the splice site score is variable. Assuming that constitutive

exons with stronger splice sites depend less on ESRs to be correctly included in mRNA than those with weaker splice sites, is predictable that the constitutive exons with weaker splice sites would be more sensitive to the inhibitory effects of ESS on exon inclusion, than constitutive exons with strong splice sites. We tested this hypothesis using the modified MK approach to ESS creating changes within constitutive exons with strong or weak splice sites. We used the method described by Yeo and Burge (2004) to estimate the 5' and 3' splice-site strengths for all constitutive exons containing substitutions and polymorphisms. We then used the median of the splice-site score distribution (splice site score median = 8.5) as a threshold to define the categories of exons with weak and strong splice-sites. The results show an increase in the enrichment of polymorphisms (with respect to substitutions) with the weakening of the splice site strength (Figure 3, see also supplementary tables S4).

Figure 3 Normalized difference between the observed and expected counts of substitutions and polymorphisms which create rate-shifted_{skipped} ESRs within sequences of human constitutive exons. Constitutive exon sequences were divided accordingly to the splice site score and tested separately. Only constitutive exons with weak splice site showed significant results for the modified version of MK test. ** Chi-square p-value <0.01. * Marginally significant (Chi-square p-value = 0.05)

This suggests that purifying selection is stronger on changes that occur in exons with weaker splice sites and refutes the abovementioned hypothesis that the significant signal of purifying selection on constitutive exons was only due to higher absolute number of variant sites analyzed.

Consistently, similar results were observed by Xiao *et al.*, (2007), who analyzed ESS motifs and found that the purifying selection against ESS is stronger in constitutive exons with weaker splice sites. Moreover, it is important to note that among the exons with weak splice sites, those with weak 5' splice

site show stronger signal of natural selection against the rate-shifted hexamers than those with weak 3' splice site (Figure 3). This result is in accordance with the major role of 5' splice site in stabilizing the spliceosome complex (Robberson *et al.*, 1990), especially in mammals where exon definition predominates (Sterner *et al.*, 1996; Xiao *et al.*, 2007). These results are again concordant with the findings of Xiao *et al.*, (2007), who showed that changes in 5' splice site strength better predicts the evolution of ESS than changes in the 3' splice site strength.

Concluding remarks

Certain evolutionary rate shift events may be caused by selective pressure acting at amino-acid level (Pupko and Galtier, 2002). However, this argumentation is not sufficient to explain our results, which show that synonymous mutations located in rate-shifted ESRs display significant deviations from neutral expectations. Moreover, the fact that the purifying selection signal varies depending on the kind of ESR alteration, i.e., associated with ESRs creation or loss, suggests that selection is acting at the RNA level.

The majority of rate-shifted hexamers that we identified in this study are identical (or completely aligned) to known ESEs, including several binding sites for SR-proteins. However, we believe that those identified from skipped exons act as silencers. This hypothesis is supported by our finding of a higher proportion of known ESS and lower proportion of ESEs among them (relative to ESRs identified in constitutive exons). Moreover, these hexamers present an evolutionary pattern similar to known ESS, i.e., both is under purifying selection in constitutive exons. These findings are in accordance with previous reports

about the context-dependent function of ESRs (Goren *et al.*, 2006; Jumaa and Nielsen 1997; Kanopka *et al.*, 1996; Solis *et al.*, 2008; Ule *et al.*, 2006) but further experimental analysis will be necessary to prove the silencer function of these rate-shifted hexamers.

Finally, this study provides the first attempt to demonstrate consistent evidence that, at the populational level, slightly deleterious mutations occurs in constitutive exons which reduce their potential to be recognized during the splicing. Given that the MK test revealed an excess of SNPs creating ESRs with putative silencer function, relative to substitutions within constitutive exons, we concluded that these SNPs should reduce the strength of splicing signals and exon recognition process.

This result strengthens the hypothesis of a constitutive origin for skipped exons and predicts that many human exons defined as constitutive should present some splicing variants at low level. The new technology of deep RNA sequencing applied to large populational samples will probably detect those cases.

References

- Clark F, Thanaraj TA (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet* 11:451
- Consortium GP (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061
- de Souza JE, Ramalho RF, Galante PA, Meyer D, de Souza SJ (2011) Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders. *Nucleic Acids Res* 39:4942
- Fairbrother W, Yeh R, Sharp P, Burge C (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* 297:1007
- Galante P, Sakabe N, Kirschbaum-Slager N, de Souza S (2004) Detection and evaluation of intron retention events in the human transcriptome. *RNA* 10:757
- Galante P, Vidal D, de Souza J, Camargo A, de Souza S (2007) Sense-antisense pairs in mammals: functional and evolutionary considerations. *Genome Biol* 8:R40
- Gelfman S, Burstein D, Penn O, Savchenko A, Amit M, Schwartz S, Pupko T, Ast G (2012) Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Res* 22:35
- Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G (2006) Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. *Mol Cell* 22:769
- Jumaa H, Nielsen PJ (1997) The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. *EMBO J* 16:5077
- Kanopka A, Mühlemann O, Akusjärvi G (1996) Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature* 381:535
- Ke S, Zhang XH, Chasin LA (2008) Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Res* 18:533
- Kim E, Magen A, Ast G (2007) Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res* 35:125
- Letunic I, Copley RR, Bork P (2002) Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet* 11:1561
- Lev-Maor G, Goren A, Sela N, Kim E, Keren H, Doron-Faigenboim A, Leibman-Barak S, Pupko T, Ast G (2007) The "alternative" choice of constitutive exons throughout evolution. *PLoS Genet* 3:e203
- Liu H, Zhang M, Krainer A (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* 12:1998
- Liu HX, Chew SL, Cartegni L, Zhang MQ, Krainer AR (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol* 20:1063
- McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652
- Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246:96

- Pupko T, Galtier N (2002) A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. *Proc Biol Sci* 269:1313
- Robberson BL, Cote GJ, Berget SM (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol* 10:84
- Shepard PJ, Choi EA, Busch A, Hertel KJ (2011) Efficient internal exon recognition depends on near equal contributions from the 3' and 5' splice sites. *Nucleic Acids Res* 39:8928
- Solis AS, Peng R, Crawford JB, Phillips JA, Patton JG (2008) Growth hormone deficiency and splicing fidelity: two serine/arginine-rich proteins, ASF/SF2 and SC35, act antagonistically. *J Biol Chem* 283:23619
- Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. *Genome Res* 12:1060
- Sternier DA, Carlo T, Berget SM (1996) Architectural limits on split genes. *Proc Natl Acad Sci U S A* 93:15081
- Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ, Darnell RB (2006) An RNA map predicting Nova-dependent splicing regulation. *Nature* 444:580
- Wang ET, Sandberg R, Luo S, Khrebukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470
- Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164
- Wang Z, Rolish M, Yeo G, Tung V, Mawson M, Burge C (2004) Systematic identification and analysis of exonic splicing silencers. *Cell* 119:831
- Xiao X, Wang Z, Jang M, Burge CB (2007) Coevolutionary networks of splicing cis-regulatory elements. *Proc Natl Acad Sci U S A* 104:18583
- Yeo G, Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 11:377
- Zhang XH, Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* 18:1241
- Zhang Z, Zhou L, Wang P, Liu Y, Chen X, Hu L, Kong X (2009) Divergence of exonic splicing elements after gene duplication and the impact on gene structures. *Genome Biol* 10:R120

Supplementary Material

Exon database

Our exon database was created with public genomic (build 36.1) and cDNA (mRNAs and ESTs) sequence data from UCSC Genome Browser (<http://genome.ucsc.edu/>, files: mrna.fa.gz and est.fa.gz). Additional sequences were obtained from NCBI Reference Sequence Project (<http://www.ncbi.nlm.nih.gov/RefSeq>, release 22). We also downloaded EST libraries annotation from eVOC (<http://www.evocontology.org/>).

The identification of splicing events was obtained by the alignment of cDNAs on the human genome as previously described (Galante *et al.*, 2004; Galante *et al.*, 2007). In summary, the coordinates of exon/intron borders for all cDNAs mapped onto the human genome were compared against each other to identify all splicing variants for all human genes. We used the software SIM4 (<http://globin.cse.psu.edu/html/docs/sim4.html>) for a more refined definition of the exon/intron borders. To increase the reliability of splicing events identified, we have chosen for further analysis only those events supported by at least two ESTs from two distinct libraries.

Supplementary Tables

Table S1: Number of identical (or completely aligned) motifs between rate-shifted hexamers and known ESRs sets.

Motif set	SRP55 (133)	SRP40 (671)	SF2_IgM (924)	SC35 (2585)	RESCUE (238)	PESE (2069)	ESS (176)
Rate-shifted_{Skipped} (145)	1	25	32	85	12	103	9
Rate-shifted_{Constitutive} (198)	4	30	50	113	21	161	4

Table S2: Modified MK test applied for Exonic Splicing Silencers (ESS) located within (a) Constitutive exons and (b) Skipped exons.

Table S2a: Constitutive exons

	DISRUPT-ESS	MAINTAIN-ESS
Substitutions	2525	966
Polymorphism	803	352
Chi-square p-value = 0.07		

Table S2b: Skipped exons

	DISRUPT-ESS	MAINTAIN-ESS
Substitutions	508	190
Polymorphism	189	87
Chi-square p-value = 0.48		

Table S3: Modified MK test applied for rate-shifted ESRs identified from sequences of human constitutive exons (rate-shifted_{constitutive} ESRs) located within (a) Constitutive exons and (b) Skipped exons.

Table S3a: Constitutive exons

	CREATE- rate-shifted_{constitutive}	MAINTAIN- rate-shifted_{constitutive}
Substitutions	5829	2725
Polymorphism	2138	967
Chi-square p-value = 0.47		

Table S3b: Skipped exons

	CREATE- rate-shifted_{constitutive}	MAINTAIN- rate-shifted_{constitutive}
Substitutions	1142	498
Polymorphism	473	197
Chi-square p-value = 0.68		

Table S4: Modified MK test applied for rate-shifted_{skipped} ESRs located in Constitutive exons with (a) strong 3' splice site, (b) strong 5' splice site, (c) weak 3' splice site and (d) weak 5' splice site.

Table S4a: Strong 3'ss

	CREATE- rate-shifted_{skipped}	MAINTAIN- rate-shifted_{skipped}
Substitutions	2450	701
Polymorphism	927	238
Chi-square p-value =0.21		

Table S4b: Strong 5'ss

	CREATE- rate-shifted_{skipped}	MAINTAIN- rate-shifted_{skipped}
Substitutions	2842	736
Polymorphism	1002	249
Chi-square p-value = 0.6		

Table S4c: weak 3'ss

	CREATE- rate-shifted_{skipped}	MAINTAIN- rate-shifted_{skipped}
Substitutions	2329	626
Polymorphism	832	186
Chi-square p-value = 0.05		

Table S4d: weak 5'ss

	CREATE- rate-shifted_{skipped}	MAINTAIN- rate-shifted_{skipped}
Substitutions	1937	591
Polymorphism	757	175
Chi-square p-value = 0.004		

Capítulo 4

**Seleção natural em éxons humanos com diferentes
níveis de inclusão**

**Natural selection on human exons with distinct
inclusion levels**

Prólogo

No artigo a seguir revisitamos um tema bastante debatido na literatura de evolução molecular, que é a compreensão das diferenças nos regimes e na intensidade de seleção natural que atuam sobre éxons constitutivos e alternativos.

Existem diversas evidências de que os éxons constitutivos estão sob maior restrição evolutiva do que os alternativos (ver citações no artigo abaixo). Além disso, existem estudos que defendem que éxons *skipped* de baixa inclusão evoluem sob seleção positiva. Entretanto, um dos mais contundentes trabalhos que defenderam essa hipótese (Ramensky *et al.*, 2008) apresentou resultados inconsistentes. Uma das ressalvas é que o próprio trabalho demonstrou que os resultados defendidos dependem do conjunto de SNPs utilizados. Além disso, esse trabalho utilizou apenas um tipo de método para estimar a estatística que resume o teste de seleção natural usado no trabalho (teste McDonald-Kreitman, MK). Esse teste, que faz uso de tabela de contingência 2x2, gera como estatística sumária a razão de chances (*Odds ratio*). Essa estatística sumariza os possíveis desvios entre as proporções comparadas na tabela. O estudo em questão usou uma tabela para cada categoria de éxons de modo que essas tabelas foram criadas através da soma dos dados (SNPs e substituições sinônimas e não-sinônimas) dos éxons de cada categoria. Embora válida, esta abordagem é passível de críticas dado que o teste MK foi originalmente proposto para analisar genes individuais e a soma de diversas tabelas de contingência pode produzir razões de chances não confiáveis.

Motivados por essas questões, no artigo a seguir decidimos reavaliar as evidências de diferença nos regimes seleção natural em éxons constitutivos e alternativos de humanos. Nós usamos um conjunto de SNPs descobertos através de sequenciamento e disponibilizados pelo projeto 1000 genomas. Esses SNPs apresentam frequências alélicas menos susceptíveis aos vieses de averiguação existentes em outros projetos públicos (por exemplo, HAPMAP). Além disso, estimamos a razão de chances do teste MK usando três métodos distintos, sendo um deles igual ao usado por Ramensky *et al.*, (2008) e outros dois que calculam, para uma dada categoria de éxon, a média ponderada das razões de chances calculadas para os gene individuais. Além disso, essas duas últimas abordagens nos permitiu estimar o intervalo de confiança para a estatística sumária do teste de seleção, aumentando a precisão dos nossos resultados.

Resumo

Análises evolutivas demonstraram que os éxons constitutivos e *skipped* de alta inclusão são filogeneticamente mais conservados que os *skipped* de baixa inclusão. Além disso, essa última categoria de éxon apresenta maior taxa de substituição não-sinônima do que as outras duas, sugerindo seleção positiva ou um relaxamento de seleção negativa. Recentemente essas análises foram ampliadas e incluíram SNPs humanos. Aplicando o teste McDonald-Kreitman (MK), foi reportado que éxons com baixo nível de inclusão apresentam um excesso de substituições não-sinônimas em relação ao polimorfismo, levando à conclusão de que esses éxons estão sob seleção positiva. Esse resultado é questionável pois certos conjuntos de SNPs humanos, contém viés para SNPs mais comuns (com frequências de alelos raros mais altas), implicando na sub-representação de SNPs não-sinônimos (que geralmente são mais raros). Tal viés pode mimetizar o sinal de seleção positiva no teste MK. Outra ressalva refere-se à aplicação do teste de MK, para uma dada categoria de éxon, através da somatória de tabelas de contingência contendo dados de diversos genes. Sabe-se que a razão de chances de uma tabela de contingência criada a partir da soma das células de diversas tabelas de contingência pode apresentar viés e diferir substancialmente da média ponderada dos valores de razão de chances observados para cada gene individual. No presente trabalho usamos três métodos distintos para obter a estimativa da razão de chances para uma dada categoria de éxon. Além disso, usamos SNPs descobertos através de sequenciamento de última geração. Nossos resultados confirmam resultados prévios que indicam que os éxons constitutivos e *skipped* de alta inclusão apresentam fortes sinais de seleção

negativa contra mudanças não-sinônimas fracamente deletérias. Entretanto, diferentemente do que foi descrito por outros autores, encontramos evidências de que a categoria de éxons skipped de baixa inclusão apresenta padrão evolutivo consistente modelo neutro e não indicativo de seleção positiva. Finalmente, nossos resultados corroboram a hipótese de que a seleção negativa contra mutações fracamente deletérias está diretamente correlacionada com o nível de inclusão dos éxons.

Title:

Natural selection on human exons with distinct inclusion levels.

Publication type:

Research Article

Authors:

Rodrigo F. Ramalho¹, Diogo Meyer¹

Affiliations:

¹Instituto de Biociências, Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, São Paulo, SP 05508-900, Brasil

Corresponding author:

Name: Diogo Meyer

Address: Instituto de Biociências, Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, São Paulo, SP 05508-900, Brasil

Telephone number: (55-11-30918757)

e-mail: diogo@ib.usp.br

Keywords:

alternative splicing, natural selection, MK-test

Abstract

Evolutionary analyses have revealed that constitutive and skipped exons with high inclusion levels are much more conserved (between human and mouse) than skipped exons with low inclusion level. Moreover, this latter category presented higher rate of non-synonymous substitutions than the other two, suggesting a relaxation of purifying selection and/or positive selection. Recently these analyses have been extended to human SNP variation. By applying the McDonald-Kreitman (MK) test; exons with low inclusion level were found to have an excess of non-synonymous substitutions relative to polymorphisms, revealing a pattern that has been interpreted as due to positive selection. One caveat of such analyses refers to a possible dependence on the chosen SNP dataset. It is well known that the excess of intermediate frequency alleles in public SNP datasets can bias the analysis, due to an underrepresentation of non-synonymous polymorphism (which are overrepresented in the category of rare variants). This bias for common SNPs could mimic the signals of positive selection. Another caveat refers to the application of the MK test by using a single contingency table created with data pooled from several loci which represents a given category. The Odds ratio obtained with this approach could be substantially different from the corrected mean for each locus of a given category. Here we investigate signatures of natural selection on human exons with distinct inclusion levels using the gene-based MK tests with SNPs discovered by NGS sequencing, which although not free from bias are less affected than public SNP databases. Our results confirm previous results showing that constitutive and skipped exons with high inclusion

level present strong signals of purifying selection against slightly deleterious non-synonymous variations. However, different from others, the category of skipped exons with low inclusion level showed an evolutionary pattern well-fitted to the strictly neutral model. Finally, our findings support the notion that exon inclusion level is inversely correlated with the accumulation of nearly neutral variation.

Introduction

In eukaryotes, the alternative splicing of genes is undoubtedly involved with creation of biological diversity at molecular level. Through this phenomenon, the number of gene products is increased and the gene expression breath can be modulated (Black, 2003). In the human genome, the frequency of alternative splicing is extremely high. In 2008, Wang *et al.*, (2008) estimated that 95% of human genes presents a least one splicing variant.

The alternative splicing mechanism was described as the cause of some rare human diseases and several lines of evidence corroborate its role in modulating the susceptibility to diseases (Cartegni & Krainer, 2002; Cáceres & Kornblihtt, 2002). In all these cases, genetic alterations of *cis* or *trans* splicing factors are the main determinants of splicing alterations. Although yet uncertain, the fraction of genetic mutations causing splicing alterations associated with disease is not negligible, estimated in the range of 15% to 60% (Wang & Cooper, 2007).

Given the essential role of splicing on eukaryotic gene expression, several studies have compared the evolutionary rates and selective regimes of constitutive and alternative exons (Iida & Akashi, 2000; Fairbrother, *et al.*, 2004;

Xing & Lee, 2005a; Xing & Lee, 2005b; Parmley, *et al.*, 2006; Plass & Eyras, 2006; Lev-Maor, *et al.*, 2007; Parmley, *et al.*, 2007; Ke, *et al.*, 2008; Gelfman, *et al.*, 2012). Most of these, analyzed phylogenetically conserved exons between human and mouse and concluded that constitutive exons evolve under a higher constraint than alternative exons, especially due to their impact at the protein level (Xing & Lee, 2006). Moreover, constitutive exons show the following features that corroborate a higher selective constraint than that of alternative exons: higher proportion of orthologs (Haerty & Golding, 2009; Gelfman, *et al.*, 2012), lower non-synonymous evolutionary rate (Iida & Akashi, 2000; Xing & Lee, 2005a; Plass & Eyras, 2006), higher codon bias (Iida & Akashi, 2000; Haerty & Golding, 2009), higher occurrence inside protein domains (Haerty & Golding, 2009) and lower proportion inside non-structured regions of proteins (Haerty & Golding, 2010). All of these results corroborate the hypothesis that the exon inclusion level is positively correlated with evolutionary constraint because constitutive exons, by definition, have the maximum inclusion level (Carlini & Genut, 2006; Xing & Lee, 2006; Haerty & Golding, 2009; Haerty & Golding, 2010).

However, among alternative exons there is substantial heterogeneity of inclusion levels which, in turn, are correlated with distinct evolutionary rates and selective regimes. This perspective was supported by the finding that the non-synonymous rate is four-fold higher in cassette exons with low inclusion than in high inclusion levels suggesting that the former are under positive selection (Xing & Lee, 2005a). Recently, Ramensky *et al.*, (2008), directly tested this hypothesis by applying the McDonald-Kreitman test (MK test) (McDonald & Kreitman, 1991) to human and chimpanzee orthologous exons. They found

evidence that exons with low inclusion levels present an excess of non-synonymous substitutions relative to polymorphisms, while constitutive and high inclusion exons categories showed the opposite trend. The authors concluded that the exons with low inclusion level are under positive selection while the other two exon categories evolve under purifying selection.

In the present work we propose to study the effect of exon inclusion level on the signals of natural selection. Differently from others (Ramensky, *et al.*, 2008), we use a low coverage NGS dataset which is less affected by ascertainment bias than SNP databases (although not free from bias, since rare variants are still underrepresented). The usage of a SNP dataset with reduced ascertainment bias is particularly important for studies which use the SNP allelic frequency (Clark, *et al.*, 2005).

Our study also differs from previous ones with respect to the analytical methods used. Rather than pooling all SNPs from alternative and constitutive exons in a single group, we apply two distinct methods that allow the patterns of polymorphism and divergence of individual genes to be averaged, so as to obtain an overall estimate the Neutrality Index (NI) (the summary statistic for the MK test). Stoletzki & Eyre-Walker, (2011) have shown that the pooling approach can produce biased estimates of NI. Moreover, the methods we employ allow us to calculate the confidence intervals for this estimate. The confidence interval for the neutrality index is useful to evaluate the significance of the difference among distinct categories, in this case, the categories of exons with distinct inclusion levels.

Materials and Methods

Primary exon database

Our exon database was created with public genomic (build 36.1) and cDNA (mRNAs and ESTs) sequence data from the UCSC Genome Browser (<http://genome.ucsc.edu/>, files: `mrna.fa.gz` and `est.fa.gz`). Additional sequences were obtained from NCBI Reference Sequence Project (<http://www.ncbi.nlm.nih.gov/RefSeq>, release 22). We also downloaded EST libraries annotation from eVOC (<http://www.evocontology.org/>).

Briefly, the definition of exonic and intronic regions was based on the genomic coordinates of cDNA sequences classified as 'mRNA' in GenBank. All the analyzed genes have at least one RefSeq identification code. To increase the reliability of splicing events identified, we chose for further analysis only those events supported by at least two ESTs from two distinct libraries using information from Evoc database (<http://www.evocontology.org/>). Further details about the creation of this database are described in (Galante, *et al.*, 2004; Galante, *et al.*, 2007; de Souza, *et al.*, 2011).

For all the analyses described below we started the procedures using an exon database consisting of 60,383 internal alternative exons (including all types) and 70,801 internal constitutive exons.

Polymorphism dataset

We obtained the SNP data, including the derived allele frequency (DAF), from the low coverage pilot phase of the 1000 Genomes project. We chose to

work with the African sample because it has more SNPs than the other samples (Consortium 2010) allowing us to reach more power in the statistical test. The SNP data was downloaded from ftp://ftptrace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2010_07/low_coverage/snps/. We used the software Annovar (Wang, *et al.*, 2010) to annotate all SNPs and filter the synonymous or non-synonymous variants for further analysis.

The analyses reported here are based on 3,395 skipped exons containing a total of 4,527 SNPs and 14,860 constitutive exons containing a total of 20,604 SNPs. A total of 7,074 distinct genes were analyzed.

Substitution dataset

We downloaded the genomic alignments of human and rhesus (axtNet format) from the UCSC genome browser (<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/vsRheMac2/>). Using a local Perl script we parsed this file to find the chromosome, the genomic coordinate and the nucleotide for each divergent site between the two species. Approximately 180 million substitutions were found for the whole genomes and approximately 650,000 for the human exome. We used the Annovar software to annotate the substitutions. The divergent nucleotides of each site were entered in the software as being alleles from a biallelic SNP.

For further analyses described here, 11,238 skipped exons containing a total of 40,342 substitutions and 46,054 constitutive exons containing a total of 187,319 substitutions were selected. A total of 11,000 genes belong to the dataset of substitutions.

Exon inclusion

We further explore our exon database and used the alignments of cDNA sequences to estimate the exon inclusion level for each alternatively spliced exon. Similarly to Xing & Lee, (2005), we defined the inclusion level as the ratio between the total number of cDNA sequences which report the corresponding exon divided by the sum of cDNAs which include and skip the corresponding exon. These exons were then divided into two major classes: minor (inclusion level below 1/3) and major (inclusion level above 2/3).

McDonald-Kreitman test

The MK test uses a 2x2 contingency table to compare the ratio between the counts of substitutions and polymorphisms among synonymous and non-synonymous categories of variation. Non-synonymous variants have a higher phenotypic effect than synonymous variants and are therefore under stronger influence of natural selection. Under strict neutrality the ratio of polymorphism to divergence will be the same for synonymous and non-synonymous variation. An excess of non-synonymous substitutions compared to polymorphism indicates positive selection, while the opposite -- an excess of non-synonymous SNPs compared to substitutions- indicates the presence of slightly deleterious variants.

The MK test assumes that highly deleterious mutations do not segregate in a population as polymorphisms (because they are rapidly removed by natural selection) and therefore they also do not appear as substitutions. On the other hand, slightly deleterious variants generate an excess of non-synonymous

polymorphism because they are not sufficiently deleterious to be immediately removed by negative selection, and thus contribute to polymorphism. However, over longer spans of time they rarely go to fixation and are frequently ultimately removed, thus failing to contribute to divergence. The excess of non-synonymous polymorphism in the human genome is well documented (Bustamante, *et al.*, 2005) and is strong evidence that many mutations segregating in our species are weakly deleterious.

The result of MK test can be summarized using the Odds ratio $(P_n/P_s)/(D_n/D_s)$, also called Neutrality Index (NI), where P_n and P_s are the counts of non-synonymous and synonymous polymorphic sites respectively, and D_n and D_s are the counts of non-synonymous and synonymous substitutions respectively. When the Odds ratio is greater than one (i.e., the ratio P_n/P_s is greater than D_n/D_s) there is evidence for a depletion of non-synonymous substitutions relative to polymorphisms (or an excess of non synonymous polymorphism). This suggests negative selection against the accumulation non-synonymous weakly deleterious substitutions. A neutrality index less than one means that D_n/D_s is greater than P_n/P_s , and such a result is often interpreted as a signal of positive selection.

Demographic effects are known to influence the NI (Hughes, 2007) and suggests that for non-equilibrium populations, as is the case for our species, the NI should be interpreted accounting for this. In the present study our use of the NI is focused on comparing the values for different regions of the genome, and we thus expected a cancelling out effect of any demographic influence, since it should affect all autosomes similarly.

Estimating the neutrality index

We estimate neutrality index using two distinct approaches. The first, called “pooled” herein, we use all variant sites of a given exon category and construct a contingency table where each cell contains the total number of SNPs or divergent sites in a mutational category (synonymous or non-synonymous). The second approach uses the statistical methods of Mantel-Haenszel (MH) and Tarone-Greenland (TG) -- implemented by the software Dofe (http://www.lifesci.sussex.ac.uk/home/Adam_Eyre-Walker/Website/Software.html) -- to produce an estimate of NI from a large number of contingency tables, which one containing variant sites from a single gene. These methods will be indicated by NI_{pool} , NI_{MH} and NI_{TG} during the text. The MH and TG methods are well defined statistical methods that allow values for multiple contingency tables to be combined, avoiding biases of the “pooled approach”. Stoletzki & Eyre-Walker, (2011) used simulations and resampling methods to show that NI_{TG} is an essentially unbiased estimator of NI.

Statistical test

For all comparisons of proportions presented in this article we used the chi-square distribution to evaluate the statistical significance of the difference between the expected and observed values. We used the chi-square test implemented in the function ‘chisq.test’ of R statistical software (<http://www.r-project.org/>) to estimate the p-value for contingency tables and to calculate the normalized deviation to the expected proportions ((observed-expercted)/expected) for each cell of a contingency table.

Results and Discussion

First, we performed the MK test for constitutive and skipped exons from the pools of 5835 and 2033 genes, respectively. The Neutrality Index (NI) was $NI_{pool}=2,12$ for constitutive exons and $NI_{pool}=2,09$ for skipped exons (see Tables S1A and S1B for the total amount of variant sites analyzed). Similar results were obtained by using the gene-based approach implemented through the Mantel-Haenszel (MH) and Tarone-Greenland (TG) methods. For constitutive the estimates were $NI_{MH}=2,16$ and $NI_{TG}=1,72$ and for skipped $NI_{MH}=2,03$ and $NI_{TG}=1,47$.

These results show that both categories of exons present significant signatures of purifying selection acting against slightly deleterious mutations—depletion of non-synonymous substitutions relative to polymorphisms (chi-square $p\text{-value} < 2.2e^{-16}$ for both exon categories). Additionally, the confidence interval (C.I) obtained for the NI estimates showed little overlap for values in constitutive and skipped exons (Table 1). The values of NI_{TG} found in the present study are very similar to those reported by Stoletzki & Eyre-Walker, (2011) in their reanalysis of 11,624 genes from the dataset of Bustamante *et al.*, (2005) where $NI_{TG}=1.59$, which is contained within the CI for both sets of exons which we analyzed.

Table 1: Neutrality index (Odds ratio) estimated for constitutive and alternative (skipped) exons using the pooled data and two distinct gene-based approaches.

	Pool	Mantel-Haenszel (C.I)	Tarone-Greenland (C.I)
Constitutive	2,12	2,16 (2,04-2,28)	1,72 (1,62-1,82)
Skipped	2,09	2,03 (1,85-2,24)	1,47 (1,33-1,62)

Previous studies have proposed that skipped exons with low inclusion level (low I.L.) are under positive selection (Xing & Lee, 2005a; Ramensky, *et al.*, 2008). Therefore, we tested if the inclusion level of skipped exons could alter the signatures of purifying selection described above. The results showed that the signature of purifying selection remains present for both categories of exons but is higher for exons with high inclusion level (Figure 1). Although the NI values obtained by distinct method differ, they all reflect the abovementioned trend. For high inclusion category the neutrality indexes were $NI_{pool}=2,37$, $NI_{MH}=2,14$, $NI_{TG}=1,58$, all higher than NI observed for low inclusion ($NI_{pool}=1,52$, $NI_{MH}=1,46$, $NI_{TG}=0,99$). Moreover the chi-square p-values for contingency tables containing pooled data from high or low inclusion exons remain significant ($p < 2.2e^{-16}$ and $p=1,72e^{-9}$, respectively).

Figure 1: Normalized difference between the observed and expected counts of non-synonymous substitutions and polymorphisms. Skipped exons with high inclusion (black bars) and with low inclusion level (grey bars) were analyzed separately.

Our results do not corroborate the previous evidence of positive selection in skipped exons with low inclusion level. However, consistent with previous studies we found that the Odds ratio observed for this category was lowest relative to exons with high inclusion level or constitutive (Figure 2).

A)

B)

Figure 2: Exons with low inclusion level show the lowest neutrality index relative to other exon categories. Two distinct methods were applied for estimate Odds ratio, (A) Mantel-Haenszel and (B) Tarone-Greenland. Lines represent the confidence interval for the Odds ratio (point).

Note that, for both methods, the confidence intervals for the low inclusion category does not overlap with the other exon categories, suggesting that they are significantly lower. Moreover, the neutrality index estimated by the method of Tarone-Greenland was almost equal to one ($O.R=0.99$) reflecting an evolutionary pattern that fits the neutral model. If slightly deleterious polymorphisms are present in the sample, it is reasonable to assume that they will probably be non-synonymous rather than synonymous, because the former have greater phenotypical effect. Slightly deleterious polymorphisms should be enriched in the fraction of SNPs with low heterozygosity (specifically, with low derived allele frequency, DAF). If this is the case, the inclusion of SNPs with low DAF in the analysis could inflate the category of non-synonymous polymorphism relative to substitutions and mask a possible signal of positive selection in the MK test (an enrichment of non-synonymous substitutions relative to polymorphisms characterizes positive selection in MK test).

Therefore, we applied the MK test using only common SNPs (DAF greater than 5%). Our results show that without the rare SNPs -a category probably enriched in slightly deleterious mutations- the signal of purifying selection remains present and significant for both constitutive and skipped categories, although reduced to what is seen in the analysis with the entire dataset.

For exons with low inclusion level, we again observed that the estimate of NI reached values of less than one (Figure 3).

A)

B)

Figure 3: Neutrality index using only common SNPs (DAF > 5%). Exons with low inclusion level shows the lowest neutrality index relative to other exon categories. The estimate by using the Tarone-Greenland method (B) resulted in Odds ratio <1. On the other hand, the method of Mantel-Haenszel (A) results in NI value near from 1. Lines represent the confidence interval for the Odds ratio (point).

The chi-square test applied to the category of exons with low inclusion level with only common SNPs was only marginally significant ($p=0,05$, contingency table not shown). Therefore, our result does not indicate strong evidence of positive selection in exons with low inclusion level.

Next, we summarize all results obtained by us. As illustrated in Figure 4, there is a clear trend toward a reduction of selective constraints according to a reduction in inclusion levels independently of the method used for estimate the NI.

A)

B)

Figure 4: Signature of purifying selection against slightly deleterious non-synonymous changes is positive correlated with exon inclusion level. Both statistical methods Mantel-Haenszel (A) Tarone-Greenland (B) reveal this pattern.

Conclusions

Here we confirm previous results showing that constitutive and skipped exons with high inclusion level present an excess of non-synonymous polymorphism, a pattern consistent with selection acting against weakly deleterious mutations. We also found that the neutrality index differs between constitutive and skipped exons, with less evidence of selection against slightly deleterious mutations in the skipped category. These results corroborate the hypothesis that the exon inclusion level on mRNA is positively correlated with evolutionary constraint. It is important to note that these two exon categories represent at least 75% of exons from human coding genes (Kim, *et al.*, 2007; de Souza, *et al.*, 2011).

The category of skipped exons with low inclusion level showed an evolutionary pattern well-fitted to the strictly neutral model (supposedly deleterious substitutions were proportional to supposedly deleterious polymorphisms). Finally, our results do not corroborate that all exons with low inclusion level are under positive selection.

References

- Black, D. (2003). "Mechanisms of alternative pre-messenger RNA splicing." Annu Rev Biochem **72**: 291-336.
- Bustamante, C., A. Fledel-Alon, et al. (2005). "Natural selection on protein-coding genes in the human genome." Nature **437**(7062): 1153-1157.
- Carlini, D. B. and J. E. Genut (2006). "Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers." J Mol Evol **62**(1): 89-98.
- Cartegni, L. and A. Krainer (2002). "Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1." Nat Genet **30**(4): 377-384.
- Clark, A. G., M. J. Hubisz, et al. (2005). "Ascertainment bias in studies of human genome-wide polymorphism." Genome Res **15**(11): 1496-1502.

- Consortium, G. P. (2010). "A map of human genome variation from population-scale sequencing." *Nature* **467**(7319): 1061-1073.
- Cáceres, J. F. and A. R. Kornblihtt (2002). "Alternative splicing: multiple control mechanisms and involvement in human disease." *Trends Genet* **18**(4): 186-193.
- de Souza, J. E., R. F. Ramalho, et al. (2011). "Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders." *Nucleic Acids Res* **39**(12): 4942-4948.
- Fairbrother, W. G., D. Holste, et al. (2004). "Single nucleotide polymorphism-based validation of exonic splicing enhancers." *PLoS Biol* **2**(9): E268.
- Galante, P., N. Sakabe, et al. (2004). "Detection and evaluation of intron retention events in the human transcriptome." *RNA* **10**(5): 757-765.
- Galante, P., D. Vidal, et al. (2007). "Sense-antisense pairs in mammals: functional and evolutionary considerations." *Genome Biol* **8**(3): R40.
- Gelfman, S., D. Burstein, et al. (2012). "Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons." *Genome Res* **22**(1): 35-50.
- Haerty, W. and B. Golding (2009). "Similar selective factors affect both between-gene and between-exon divergence in *Drosophila*." *Mol Biol Evol* **26**(4): 859-866.
- Haerty, W. and G. B. Golding (2010). "Genome-wide evidence for selection acting on single amino acid repeats." *Genome Res* **20**(6): 755-760.
- Hughes, A. L. (2007). "Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level." *Heredity (Edinb)* **99**(4): 364-373.
- Iida, K. and H. Akashi (2000). "A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes." *Gene* **261**(1): 93-105.
- Ke, S., X. H. Zhang, et al. (2008). "Positive selection acting on splicing motifs reflects compensatory evolution." *Genome Res* **18**(4): 533-543.
- Kim, E., A. Magen, et al. (2007). "Different levels of alternative splicing among eukaryotes." *Nucleic Acids Res* **35**(1): 125-131.
- Lev-Maor, G., A. Goren, et al. (2007). "The 'alternative' choice of constitutive exons throughout evolution." *PLoS Genet* **3**(11): e203.
- McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the *Adh* locus in *Drosophila*." *Nature* **351**(6328): 652-654.
- Parmley, J., A. Urrutia, et al. (2007). "Splicing and the evolution of proteins in mammals." *PLoS Biol* **5**(2): e14.
- Parmley, J. L., J. V. Chamary, et al. (2006). "Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers." *Mol Biol Evol* **23**(2): 301-309.
- Plass, M. and E. Eyraas (2006). "Differentiated evolutionary rates in alternative exons and the implications for splicing regulation." *BMC Evol Biol* **6**: 50.
- Ramensky, V., R. Nurtdinov, et al. (2008). "Positive selection in alternatively spliced exons of human genes." *Am J Hum Genet* **83**(1): 94-98.
- Stoletzki, N. and A. Eyre-Walker (2011). "Estimation of the neutrality index." *Mol Biol Evol* **28**(1): 63-70.
- Wang, E. T., R. Sandberg, et al. (2008). "Alternative isoform regulation in human tissue transcriptomes." *Nature* **456**(7221): 470-476.
- Wang, G. S. and T. A. Cooper (2007). "Splicing in disease: disruption of the splicing code and the decoding machinery." *Nat Rev Genet* **8**(10): 749-761.
- Wang, K., M. Li, et al. (2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." *Nucleic Acids Res* **38**(16): e164.
- Xing, Y. and C. Lee (2005). "Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences." *Proc Natl Acad Sci U S A* **102**(38): 13526-13531.

- Xing, Y. and C. Lee (2006). "Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes." Nat Rev Genet **7**(7): 499-509.
- Xing, Y. and C. J. Lee (2005). "Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing." PLoS Genet **1**(3): e34.

Supplementary Material

Tables S1A and S1B:

A) Skipped exons

Skipped	Substitutions	SNPs
Synonymous	22625	2648
Non synonymous	6899	1695
p-value<2.2e-16 Odds ratio = 2.09		

B) Constitutive exons

Constitutive	Substitutions	SNPs
Synonymous	126700	12616
Non synonymous	36878	7803
p-value<2.2e-16 Odds ratio = 2,12		

Capítulo 5

Discussão geral e Conclusões

Os resultados obtidos nos capítulos anteriores permitem algumas conclusões gerais sobre os processos evolutivos que atuam sobre os elementos reguladores de *splicing* (*Exonic Splicing Regulators, ESRs*), e também sobre as diferenças no modo e na intensidade de seleção natural que atua sobre éxons constitutivos e alternativos. Neste capítulo faço uma síntese dos principais resultados obtidos e discuto as implicações e perspectivas geradas por eles.

A) Seleção natural sobre reguladores de splicing é relativamente fraca

Comparando o déficit de substituições fracamente deletérias em reguladores de *splicing* (resultados do capítulo 3) com o déficit de substituições fracamente deletérias que afetam as proteínas (resultados do capítulo 4), concluímos que o sinal de seleção natural sobre mudanças sinônimas que afetam reguladores de *splicing* é aproximadamente 5 vezes menor que o observado para seleção sobre mudanças não-sinônimas (redução de 2% vs. 10%). Essa é uma estimativa conservadora, pois escolhemos o caso em que o sinal de seleção natural sobre os reguladores de *splicing* foi mais significativo (mutações sinônimas de ganho de *Exonic Splicing Silencers* em éxons constitutivos de sítio de *splicing* fraco).

Em nossa análise sobre seleção natural em reguladores de *splicing* (capítulo 3), observamos apenas 2% de redução nas substituições fracamente deletérias em ESRs em relação ao esperado por nosso controle neutro. É possível que a existência de um alto nível de degeneração dos ESRs, explicada pela baixa especificidade de seus ligantes (Fu, 1995), resulte em um

cenário de evolução quase neutra, onde as mutações têm baixo impacto no valor adaptativo (*fitness*).

Em concordância com esta hipótese, outros autores relataram grande dificuldade em se prever o efeito de mutações dirigidas aos ESRs na alteração do nível de inclusão dos éxons (Zhang, *et al.*, 2009). Supondo o cenário em que os elementos reguladores são altamente degenerados, a mutação deletéria dirigida a um dado regulador pode ao mesmo tempo criar um novo regulador adjacente.

As abordagens filogenéticas acessam os sinais de seleção natural acumulados ao longo de milhões de anos e, por esse motivo, são capazes de detectar evidências de seleção mesmo sobre mudanças com baixo efeito no *fitness*. Por outro lado, as abordagens populacionais acessam sinais de seleção acumulados durante um período de tempo menor (dezenas de milhares de anos) e portanto não tem o mesmo poder estatístico. Sendo assim, as abordagens filogenéticas, ainda que combinadas com abordagens populacionais são fundamentais para a obtenção de evidências empíricas de seleção natural em reguladores de *splicing*.

B) ESRs de éxons constitutivos e alternativos evoluem de modos distintos

Os resultados obtidos nos capítulos 2 e 3 indicam que os reguladores de *splicing* evoluem de forma distinta em éxons constitutivos e alternativos, corroborando resultados prévios (Ke, *et al.*, 2008). No capítulo 3, apresentamos evidências adicionais de seleção purificadora sobre mutações associadas à criação de ESS (*Exonic Splicing Silencers, ESS*) em éxons constitutivos. Por outro lado, em éxons alternativos do tipo *skipped* esse sinal não foi observado.

Esse resultado sugere um possível relaxamento de seleção sobre mudanças que criam ESSs em éxons alternativos do tipo *skipped* e é consistente com a densidade duas vezes menor de ESSs em éxons constitutivos do que alternativos, conforme descrito no capítulo 2.

Embora indiretamente, esse resultado também é consistente com os resultados observados no capítulo 4, que corroboraram uma maior restrição evolutiva ao nível proteico em éxons constitutivos do que em alternativos.

C) Inibição dos sinais de splicing como o principal mecanismo de origem de éxons skipped

No capítulo 3, usando um teste específico para detecção de sinais de seleção natural em ESRs (teste McDonald-Kreitman), vimos que mudanças associadas ao ganho de ESSs evoluem de modo neutro (sem restrição evolutiva) em éxons alternativos, mas, em éxons constitutivos, apresentam certa restrição. Além disso, vimos no capítulo 2 que as variações de um único nucleotídeo que apresentam associação significativa com eventos de *splicing* alelo-específicos (uma forma plausível de regulação do *splicing* alternativo) (Stranger, *et al.*, 2007; Coulombe-Huntington, *et al.*, 2009; Ge, *et al.*, 2009; Kwan, *et al.*, 2009) estão localizadas em ESSs e não em ESEs (*Exonic Splicing Enhancers, ESEs*).

Esses dois resultados, obtidos através de abordagens independentes, sugerem que o mecanismo de regulação do *splicing* alternativo se dá principalmente através da função inibitória, normalmente associada aos *Exonic Splicing Silencers* (ESS), e não através da função promotora, normalmente associada aos *Exonic Splicing Enhancers* (ESEs).

D) Acúmulo de ESSs em éxons constitutivos é um mecanismo de criação de éxons alternativos do tipo skipped

A existência de um excesso de polimorfismos (em relação às substituições de mesmo tipo) associados ao ganho de ESSs em éxons constitutivos atesta que essas mudanças são fracamente deletérias. Se elas fossem altamente deletérias elas não segregariam como polimorfismos na população e por isso não seriam amostradas como SNPs.

Curiosamente, demonstramos no capítulo 3 que outra categoria de ESRs, envolvidos com a origem de éxons *skipped* na linhagem dos mamíferos (Lev-Maor, *et al.*, 2007), também apresenta esse mesmo padrão evolutivo.

Se o ganho desses ESRs (cujas maioria é de promotores de inclusão) é fracamente deletério em éxons constitutivos é porque eles provavelmente têm função de inibir a inclusão dos éxons. Portanto, primeiramente concluímos que, conforme descrito por outros autores, certos reguladores descritos como ESEs podem apresentar função ambígua. Assumindo que esses ESRs atuam principalmente como ESS, podemos inferir que os ESRs envolvidos com a origem de éxons *skipped* na linhagem dos mamíferos tem função inibitória sobre a inclusão dos éxons durante o processamento do RNAm.

À primeira vista, esse modelo pode parecer contraditório com o fato de que, conforme demonstramos no capítulo 2, os éxons constitutivos apresentam significativa redução na densidade de inibidores em relação aos éxons *skipped*. Entretanto, a união desses dois resultados (menor densidade de inibidores em éxons constitutivos em relação aos alternativos e significativo excesso de polimorfismo associado ao ganho de inibidores apenas em éxons constitutivos)

sugere que deve haver um limiar tolerável de acúmulo de ESSs nas sequências dos éxons constitutivos. Quando esse limiar é ultrapassado ocorre diminuição no nível de inclusão dos éxons, originando assim o exon alternativo. A partir desse momento os inibidores passam a evoluir de modo neutro, sem restrição evolutiva.

A observação de que esse padrão evolutivo é mais evidente em éxons constitutivos cujos sítios de *splicing* são mais fracos corrobora a hipótese de que a origem de éxons *skipped* se dá a partir de éxons constitutivos. Os éxons constitutivos com sítios fracos estariam mais próximos de se tornarem alternativos devido às reduções em seus sinais de *splicing* (enfraquecimento de sítios de *splicing* e acúmulo de ESRs com função inibitória).

E) O nível de inclusão dos éxons está diretamente relacionado à intensidade de seleção negativa sobre mudanças não-sinônimas

Usando dados públicos de SNPs descobertos por sequenciamento de última geração pudemos corroborar resultados prévios que evidenciam que o nível de inclusão dos éxons está diretamente correlacionado com o sinal de seleção negativa que atua sobre mudanças não-sinônimas. Usando uma abordagem que contrasta os níveis de polimorfismo e divergência, demonstramos que éxons com máxima inclusão, isto é, os éxons constitutivos, sofrem a ação da seleção purificadora contra mutações não-sinônimas com maior intensidade que éxons *skipped*. Entre os éxons *skipped*, os de alta inclusão (aqueles incluídos em mais do que 2/3 dos cDNAs analisados) estão sob seleção negativa maior que os éxons de baixa inclusão. Considerando as

mudanças que afetam a proteína, podemos afirmar que os éxons *skipped* de baixa inclusão estão sob uma restrição evolutiva menor.

Bibliografia

- Coulombe-Huntington, J., K. Lam, et al. (2009). "Fine-scale variation and genetic determinants of alternative splicing across individuals." PLoS Genet **5**(12): e1000766.
- Fu, X. D. (1995). "The superfamily of arginine/serine-rich splicing factors." RNA **1**(7): 663-680.
- Ge, B., D. Pokholok, et al. (2009). "Global patterns of cis variation in human cells revealed by high-density allelic expression analysis." Nat Genet **41**(11): 1216-1222.
- Ke, S., X. H. Zhang, et al. (2008). "Positive selection acting on splicing motifs reflects compensatory evolution." Genome Res **18**(4): 533-543.
- Kwan, T., E. Grundberg, et al. (2009). "Tissue effect on genetic control of transcript isoform variation." PLoS Genet **5**(8): e1000608.
- Lev-Maor, G., A. Goren, et al. (2007). "The "alternative" choice of constitutive exons throughout evolution." PLoS Genet **3**(11): e203.
- Stranger, B., A. Nica, et al. (2007). "Population genomics of human gene expression." Nat Genet **39**(10): 1217-1224.
- Zhang, X., M. Arias, et al. (2009). "Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing." RNA **15**(3): 367-376.

Apêndice 1

Cópia do artigo “Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders” publicado no periódico Nucleic Acids Research (*Nucl. Acids Res.* (2011) 39 (12): 4942-4948)

Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders

Jorge E. S. de Souza¹, Rodrigo F. Ramalho², Pedro A. F. Galante¹, Diogo Meyer² and Sandro J. de Souza^{1,*}

¹Ludwig Institute for Cancer Research at Hospital Alemão Oswaldo Cruz, São Paulo, SP, 01509-010 and

²Instituto de Biociências, Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, São Paulo, SP 05508-900, Brasil

Received June 15, 2010; Revised January 24, 2011; Accepted February 1, 2011

ABSTRACT

With the availability of a large amount of genomic data it is expected that the influence of single nucleotide variations (SNVs) in many biological phenomena will be elucidated. Here, we approached the problem of how SNVs affect alternative splicing. First, we observed that SNVs and exonic splicing regulators (ESRs) independently show a biased distribution in alternative exons. More importantly, SNVs map more frequently in ESRs located in alternative exons than in ESRs located in constitutive exons. By looking at SNVs associated with alternative exon/intron borders (by their common presence in the same cDNA molecule), we observed that a specific type of ESR, the exonic splicing silencers (ESSs), are more frequently modified by SNVs. Our results establish a clear association between genetic diversity and alternative splicing involving ESSs.

INTRODUCTION

The large amount of data on the human transcriptome has allowed several studies that, without exception, show a high prevalence of alternative splicing in the human transcriptome (1–3). The fact that most human genes undergo alternative splicing has raised doubts about the biological significance of most of the variants. One possibility is that a significant fraction of all variants are spurious products of the splicing machinery, without any functional relevance. Indeed, there are significant differences (e.g. the preservation of codon reading frame) between splicing

variants that are conserved between human and mouse, and therefore deemed as functional, and those that are not, suggesting that a fraction of the splicing variants are spurious products (4). Some authors have even suggested that these products would have functional implications by down-regulating the expression of functional variants (5). On the other hand, some have argued that most of the splicing variants are products of a regulated process. For instance, Wang, E. *et al.* (6) observed that most of the splicing variants of human genes show differential expression among different tissues whereas variation between individuals was ~2 to 3-fold less common. These results corroborate the hypothesis that alternative transcripts could have tissue specific functionalities. It has also been shown by our group that intron retention events are not randomly distributed regarding several parameters, again suggesting the notion that the majority of these splicing variants are not spurious and their expression is somehow regulated (7,8). Based on the current evidence it is reasonable to speculate that at least one third of all splicing variants are products of regulated expression.

In addition to the importance of 5'/3'-splicing sites, branch point and polypyrimidine tracts in the control of splicing, there are several known *cis*-regulatory splicing elements that contribute to the splicing process and are located in intronic or exonic regions [for a review see (9,10)]. Many lines of evidence suggest that these elements can act by stimulating (enhancing) or inhibiting (silencing) the inclusion of the respective exon, or the neighbor exon, in the mature RNA transcript. These features are taken into account for the nomenclature of splicing regulatory elements. Those present in exons and with the capacity of enhancing splicing are called exonic

*To whom correspondence should be addressed. Tel: +55-11-33883211; Fax: +55-11-31411325; Email: sandro@compbio.ludwig.org.br

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2011. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.5>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

splicing enhancers (ESE) and those with the capacity of inhibiting the splicing are the exonic splicing silencers (ESS). Generally, these classes of elements are called exonic splicing regulators (ESRs).

Several studies suggest that ESS have a significant role in the control of alternative splicing. For example, (i) using a set of paralogous exons, where one copy showed constitutive splicing and the other alternative splicing, Zhang, Z. *et al.* (11) found that the alternative copy had significantly lower ESE and higher ESS densities than the constitutive copy; (ii) using designed exons constructed by random ligation of ESEs, ESSs and neutral sequences, Zhang, X. *et al.* (12) showed that negative correlation between ESS density and inclusion rate was stronger than the positive correlation between ESE density and inclusion rate; (iii) The set of motifs which bind the tissue-specific splicing factors Nova1 and Nova2 can act as ESEs or ESS depending on their position in the primary transcript. When located in alternative exons they mainly act as silencers (13).

In the present study we used single nucleotide variation (SNV) and cDNA data to compare the genetic diversity of ESRs located in constitutive and alternative exons. By establishing an association between the SNV alleles and distinct borders of alternative exons our results show that variations in ESSs, and not in ESEs, are more commonly associated with alternative splicing.

MATERIALS AND METHODS

Public data

We obtained genomic (build 36.1) and cDNA (mRNAs and ESTs) sequence data from UCSC Genome Browser (<http://genome.ucsc.edu/>, files: *mrna.fa.gz* and *est.fa.gz*). Additional sequences were obtained from NCBI Reference Sequence Project (<http://www.ncbi.nlm.nih.gov/RefSeq>, release 22). We also downloaded EST libraries annotation from eVOC (<http://www.evoontology.org/>).

Identification of splicing events

A catalog of all splicing variants reported by the alignment of cDNAs on the human genome was obtained as previously described (7,8,14,15). Briefly, the coordinates of exon/intron borders for all cDNAs mapped onto the human genome were compared against each other to identify all splicing variants for all human genes. We used the software SIM4 (<http://globin.cse.psu.edu/html/docs/sim4.html>) for a more refined definition of the exon/intron borders. To increase the reliability of splicing events identified, we have chosen for further analysis only those events supported by at least two ESTs from two distinct libraries.

Exon classification

The definition of exonic and intronic regions was based on the genomic coordinates of cDNA sequences classified as 'mRNA' in GenBank. Regarding alternative splicing, four groups of exons were defined for the current analyses:

exons reporting different donor or acceptor sites formed the Cryptic group (alternative splice site, 35 391 exons), exons missing in two or more transcripts formed the Skipping group (46 586 exons) and exons reporting an intron retention formed the Retention group (8310 exons). These three groups represent the major forms of alternative splicing. A fourth group, named Alternative (60 383 exons), was formed by the union of the three groups of alternative exons mentioned above excluding redundant exons among these groups. Finally, the Constitutive group (70 801 exons) was composed by exons for which no alternative exon/intron borders were detected.

Mapping the ESRs in exons

Eight different data sets of putative regulatory elements (six ESEs and two ESSs) were obtained from the literature (16–21). Four (SF2_IgM, SRP40, SRP55 and SC35) out of six ESE data sets were discovered *in vitro* by using the SELEX methodology while the other two were discovered *in silico*. Regarding the SELEX-ESEs, only those oligomers with a score equal or higher than the threshold scores defined by the original study were considered as ESEs. For the remaining data sets of ESEs (RESCUE and PESE), a list of ESE motifs was obtained from the supplementary material associated with the articles of Fairbrother *et al.* (18) and Zhang and Chasin (19). The PESS data set of silencers was also obtained from Zhang and Chasin (19) and the data set of ESS reported by Wang, Z. *et al.* (20) will be called ESS herein.

To identify the ESR motifs in the exons of our in-house database we perform a pairwise alignment between each set of ESRs and the exon sequences. The ESR counts were calculated independently for each group of exons analyzed in this study (Constitutive, Alternative, Cryptic, Skipping and Retention).

Mapping SNVs in exons

To make sure the SNVs were correctly indexed in our exon database, we mapped all 17 804 036 SNVs available in the dbSNP (release 130) in the genomic sequences used for our analysis. Relative positions of SNVs regarding exonic, intronic and intergenic regions were defined by comparing SNV cDNA coordinates.

Mapping exonic SNVs in published ESRs motifs

Sequence tags comprising each SNV were generated by extracting from the reference human genome the corresponding variant nucleotide plus ten nucleotides flanking the SNVs (totaling a 21 nucleotide tag, which we will be called SNP-tag herein). We only extracted tags from the same strand orientation of the Refseq gene which contains the exon where the SNV was mapped. The alignment of these tags with the published ESR motifs defined whether or not the SNV was mapped into a known ESR.

Finding isoform-associated SNVs and defining putative ESR motifs

Alignments between all human mRNAs and the genome were searched for the presence of mismatches.

The position of the mismatches was then compared to the genomic position of SNVs. These analyses resulted in 106 271 mismatches that co-occur with SNVs. For 96 756 of these mismatches, the discordant nucleotide reported by the mRNA corresponded to one of the alleles reported in the dbSNP for the respective SNV. Since the mRNA sequences are supposedly of high quality, this last number strongly suggests that the great majority of mismatches reported in the alignments are due to SNVs. Among these mismatches we selected 3533 SNVs where each allele was completely associated with alternative exon/intron borders (we refer to these as isoform-associated SNVs). Considering all the alleles of these SNVs, we obtained 7087 sequence tags (17 tri-allelic and 2 quadri-allelic SNVs). SNVs that presented the same allele in cDNAs reporting different exon/intron borders were not included in the category of isoform-associated SNVs.

To make statistical inferences about a possible enrichment of known regulatory elements in the isoform-associated SNV data set, we randomly created 1000 control data sets from a pool of 46 336 SNVs also mapped in alternative exons but without alleles in complete association with alternative exon/intron borders. A schematic view of our approach is shown in Figure 1. Distinct control data sets were created for the three main groups of alternative exons (Cryptic, Retention and Skipping). Each control data set is of exactly the same size as the isoform-associated SNV data set (1385 SNVs for Cryptic, 1780 SNVs for Retention and 958 SNVs for Skipping). Similarly, we create control data sets, with 3533

SNVs each, to use in the comparison with the main group of alternative exons (which includes all three forms of alternative exons). Next, we used the same strategy described above to generate tags around the SNVs of the control data sets and search for known ESR motifs. For each replicate data set we counted the number of ESRs that were part of an SNP-tag, and used the distribution of these values to test the null hypothesis that our set of isoform-associated SNVs are not associated with ESRs. The *P*-value was defined as the fraction of the ranked values observed in the control data sets which were greater than the observed value in the case set.

Defining the ancestral and derived alleles of SNVs

In order to establish the polarity of the ESR modification imposed by SNVs and so define events as gains, losses or maintenance/alteration of ESRs, we compared the alleles of human SNVs to the orthologous alleles of the chimpanzee (*Pan troglodytes*) genome. To perform this analysis we used data from the table ‘snp130OrthoPt2Pa2Rm2’ available at the UCSC Genome Browser (<http://genome.ucsc.edu/>), which contains the othologs of 11 797 184 human SNVs in four species of primates, including the chimpanzee.

Statistical test

For all comparisons of proportions presented in this article we used the chi-square distribution to evaluate the statistical significance of the difference between the expected and observed values. We used the chi-square

Figure 1. Schematic representation of the computational approach used in this study to identify splicing regulatory elements. (a) Isoform-associated SNVs. Schematic view of association between genetic and splicing variation. A *sine qua non* condition is the presence of both the SNV and the alternative exon/intron border in the same cDNA molecule allowing the direct association between the alternative exon/intron border and the presence of the SNV. (b) Control. Schematic view of SNVs not associated with splicing variation, which defined the control set used in the simulation.

test implemented in the function *prop.test* of R statistical software (<http://www.r-project.org/>).

RESULTS AND DISCUSSION

Alternative exons are enriched in SNVs when compared to constitutive exons

We first compared the density of SNVs between constitutive and alternative exons. Alternative exons, when taken as an unique group, have ~10% more SNVs than constitutive exons (5.09 and 4.52 SNVs per 1000 nucleotides for alternative and constitutive exons, respectively, P -value = 2.53×10^{-102} , chi-square test).

Comparisons among three sub-groups of alternative exons reveals that the Skipping group shows a significantly lower SNV density than the other two groups.

Table 1. Comparison of ESR density between constitutive and alternative exons

ESR	Constitutive (10 650 372)	Alternative (22 622 280)	Percent change [#]	P -value*
RESCUE	0.11305	0.08943	-20.8	0
SF2_IgM	0.05466	0.06135	12.2	0
SC35	0.04156	0.04534	9.0	0
SRP40	0.04275	0.04400	2.8	6.97×10^{-61}
SRP55	0.02514	0.02530	0.06	0.01
PESE	0.07030	0.06398	-8.9	0
PESS	0.01430	0.01916	33.9	0
ESS	0.00006	0.00013	116.6	2.76×10^{-59}

* P -value for two-tailed chi-square test. Between brackets is the total number of nucleotides analyzed in each exon group. [#]Approximate percent change of Alternative compared with Constitutive exons. Positive and negative values represents excess and depletion, respectively.

(Skipping = 5.01 versus Cryptic = 5.19 and Retention = 5.17 SNVs per 1000 nucleotides, P -value $< 3.81 \times 10^{-9}$ for both comparisons). This reduced genetic diversity of skipped exons in relation to other forms of alternative exons may reflect a stronger selective constraint. This result is in accordance with the findings from Wang, E. *et al.* (6), who showed that skipped exons are more conserved among four mammalian genomes and seem to be most important in tissue-specific alternative splicing.

Alternative exons are enriched in ESRs when compared to constitutive exons

Next, the density of ESRs was compared between constitutive and alternative exons. Table 1 shows that ESR motifs are enriched in the group of alternative exons. RESCUE-ESEs and PESEs were the exceptions, presenting an opposite trend. These exceptions are, in fact, expected since RESCUE-ESEs are identified from a set of constitutive exons and PESEs are identified from a set of exons with high inclusion levels (18,19).

Exons belonging to the Skipping group showed a significant depletion in the density of SELEX-ESEs and ESSs when compared to the other groups of alternative exons (Figure 2). These results corroborate the results of Kurmangaliyev and Gelfand (22), who observed similar results in a comparison between skipped exons and exons with alternative splicing sites with mutations in their splice sites. However, they contradict the findings from Wang, J. *et al.* (23), who performed a similar analysis and found a significantly lower density of SELEX-ESEs in skipped exons when compared to constitutive exons. We believe that the discrepancy between our result and those of Wang, J. *et al.* (23) may be due to

Figure 2. Comparison of ESR density between the three main types of alternative exons. Values represent the $\log_{10}$ of the ratio between each individual group of alternative exons and the Alternative group (pool of the three types).

Table 2. Proportion of ESRs affected by SNVs in both Constitutive and Alternative exons

ESR	Constitutive	Alternative	Percent change [#]	P-value*
RESCUE	0.03900	0.04614	18.3	1.72 ^{−203}
SF2_IgM	0.07658	0.08064	5.3	7.82 ^{−22}
SC35	0.07341	0.08319	13.3	3.95 ^{−89}
SRP40	0.06326	0.06889	8.9	2.69 ^{−36}
SRP55	0.06240	0.06721	7.7	1.26 ^{−16}
PESE	0.05819	0.06800	16.8	6.82 ^{−173}
PESS	0.05858	0.06724	14.7	4.87 ^{−32}
ESS	0.12006	0.11569	−3.6	0.8

Proportions were obtained by dividing the number of ESRs affected by a SNV by the total of ESRs, within each group.
*P-value for two-tailed chi-square test. [#]Approximate fold change of Alternative compared with Constitutive exons. Positive and negative values represents excess and depletion, respectively.

differences in the group of constitutive exons used in both studies. Currently, the coverage of the human transcriptome is deeper when compared to 2005, and probably a significant proportion of exons defined as constitutive in their work is currently defined as alternative.

Alternative exons show higher proportion of ESRs modified by a SNV than constitutive exons

We next compared the proportion of ESRs modified by a SNV in both constitutive and alternative exons. Generally, ESRs in alternative exons are proportionally more modified by a SNV than those ESRs in constitutive exons (P-value <1.26^{−16}, chi-square test, highest significant P-value from Table 2). The only exception was for ESS. Despite the fact that PESE and RESCUE sets are enriched in constitutive exons, we also observed a higher proportion of these motifs mapped in SNVs of alternative exons than in SNVs of constitutive exons. Moreover, these two sets of ESRs showed the most significant differences in comparisons between constitutive and alternative exons (Table 2). Consistent with our previous observations, the Skipping group showed the lowest proportion of ESRs affected by SNVs among the groups of alternative exons (Supplementary Table S5).

The observations that the alternative exons show a higher density of SNVs, ESRs and also a higher proportion of ESRs modified by SNVs, suggest that this genetic variation could to some extent be one of the causal factors distinguishing alternative and constitutive splicing. In fact several studies analyzed the impact of single nucleotide polymorphism in the regulation of transcript isoform expression in tissue-specific and non-specific manners (24–26) and validated some causative SNPs occurring in splicing regulators (27).

ESS associated with alternative splicing are more modified by SNVs

We decided to further explore this putative association between the SNVs and alternative splicing by examining those cDNAs that reported both an alternative exon/intron border and known SNV. Two categories of SNVs were used. The first category contains SNVs with alleles in

complete association with different exon/intron borders (isoform-associated SNVs). The second category, the control set, contains the remaining SNVs mapped to alternative exons, i.e. those SNVs without a complete association with alternative exon/intron borders (Figure 1).

Among the set of sequence tags derived from the isoform-associated SNVs we found that ~86% contained at least one ESR already described. Is there an enrichment of any particular type of ESR in this set of sequence tags in comparison to tags derived from SNVs not associated with alternative borders? To answer this question, re-sampling was performed comparing the isoform-associated SNV data set to 1000 control data sets, each comprised of the same number of SNVs but not associated with alternative exon/intron borders. The analysis was performed independently for each set of published ESRs using alternative exons either as an unique group or divided into the three categories previously discussed.

Table 3 shows that isoform-associated tags are enriched in ESSs in comparison with the control data set when the alternative exons are analyzed as an unique group. This is true for both sets of ESS analyzed (PESS and ESS). Moreover, the isoform-associated SNVs is significantly depleted in ESEs when compared to the replicate data sets (Table 3). Interestingly, Zhang,X. *et al.* (12), showed that the absolute number of ESS correlates significantly ($R^2 = 0.78$, P-value < 5^{−47}) with the non-inclusion rate (negative correlation with inclusion rate) of exons, when other splicing signals are constant. Moreover, they found a significant positive correlation ($R^2 = 0.53$, P-value < 3e^{−6}) between inclusion level and the ratio ESE/ESE+ESS. Consistent with this, we found this ratio to be significantly lower in the SNP-tags of our experimental data set when compared to the SNP-tags of control data sets (data not shown). Together, these results suggest that the influence of SNVs on some types of alternative splicing occur predominantly through their effects on ESSs.

A recent study by Woolfe *et al.* (28), analyzed a small set of well curated SNPs (a total of 87) associated to exon skipping, which they compared to a large set of HapMap SNPs which were putatively neutral with respect to splicing. Using an approach different from ours, they also found that alterations of ESSs were significantly overrepresented when compared to alterations which are putatively neutral with respect to splicing. Moreover, they also found that the degree of ESS alterations was even greater for events of alternative splice site than that for exon skipping. The concordance between these two studies, which used different approaches to define the association between SNVs and splicing variants, corroborates the important role played by ESSs in the splicing regulation.

Analyzing the polarity of the ESRs changes imposed by SNVs

Can we further discriminate the effect of SNVs in ESSs? If we assume that the derived allele increases transcriptome variability by allowing the use of alternative exon/intron

Table 3. Difference in ESR densities between experimental and control sets of SNVs

ESR	Exon group	P-value
Enhancers RESCUE	Skipping	0.74
	Retention	0.87
	Cryptic	0.98 ^a
	Alternative	1 ^a
SC35	Skipping	0.95 ^a
	Retention	1 ^a
	Cryptic	0.85
	Alternative	1 ^a
SRP40	Skipping	0.86
	Retention	0.99 ^a
	Cryptic	0.87
	Alternative	0.9
SRP55	Skipping	0.19
	Retention	0.86
	Cryptic	0.44
	Alternative	0.46
PESE	Skipping	0.11
	Retention	1 ^a
	Cryptic	1 ^a
	Alternative	1 ^a
SF2_IgM	Skipping	0.09
	Retention	1 ^a
	Cryptic	0.93
	Alternative	1 ^a
Silencers PESS	Skipping	0.92
	Retention	0 ^b
	Cryptic	0 ^b
	Alternative	0 ^b
ESS	Skipping	0.1
	Retention	0.09
	Cryptic	0.01 ^b
	Alternative	0 ^b

^aSignificantly lower than control.^bSignificantly higher than control.

borders, we can try to better understand the effect of SNVs on ESS by defining a pattern of ESS gain or loss with the emergence of a derived allele. To this end we defined the polarity of change by assuming the reference chimpanzee genome as the ancestral allele, as done by others (29).

Results in Table 4 confirm that exon skipping does not seem to be primarily regulated by SNVs that create ESSs. When we independently analyzed the events of ESS gain from non-ESS motifs, the difference between case and control sets does not exist (11 SNVs in the isoform-associated set against 12 SNVs in the control set, P -value = 0.56). The difference is restricted to those events of ESS gain from an ancestral ESE (29 SNVs in the case set against 44 SNVs in the control set, P -value = 0.99). This suggests that the significant depletion of SNVs involved in ESS gain observed for this type of alternative splicing reflects depletion in the number of SNVs that affect ESEs.

These findings differ from those of Woolfe *et al.* (28), and show that the mechanism of splicing regulation among the skipped exons is more complex than just an increase in the proportion of ESS gains. We note,

Table 4. SNP counts of ESS loss, gain and maintenance found in isoform-associated SNV data set and control data set

	Isoform-associated SNV set	Control [#]	P-value
Skipping	ESS loss	23–63	0.67
	ESS gain	40–57	0.99 ^a
	ESS maintenance	10–45	0.6
	Cryptic		
Cryptic	ESS loss	39–87	0.005 ^b
	ESS gain	51–115	0.97 ^a
	ESS maintenance	19–60	0.005 ^b
Intron retention	ESS loss	63–115	0.01 ^b
	ESS gain	75–140	0.53
	ESS maintenance	43–93	0 ^b

[#]Range for 1000 replicate data sets.^aSignificantly lower than the control.^bSignificantly higher than the control.

however, that these authors compared SNVs alleles associated to skipped exons to a group of SNVs alleles belonging to a heterogeneous set of exons. This differs from our approach, in which the isoform-associated and control SNVs alleles were all from skipped exons, and may explain the differences between the studies.

For cryptic and intron retention, the predominant pattern involves loss and alteration/maintenance of ESS. Based on the significant frequency of ESS loss, we predict that the derived allele could be generating a decrease in ESS strength in those cases where ESS is maintained.

Final remarks

The results reported here support the view that ESRs have a higher genetic diversity in alternative exons when compared to constitutive exons. We believe that this genetic variation could to some extent be one of the major features distinguishing alternative from constitutive splicing. Furthermore, we provide evidence that this effect is mainly due through SNVs acting on ESS.

A possible caveat of our approach is that we cannot directly distinguish between causal and associated SNVs since an isoform-associated SNV may be in linkage disequilibrium with a different causal variant. However, our re-sampling analysis addresses this issue by examining if the isoform associated SNVs are associated to ESRs as frequently as non-isoform associated SNVs (used as a 'control'). Using this approach we were able to show that ESS are significantly overrepresented among isoform-associated SNVs, supporting their functional role in splicing regulation.

The emergence of next generation sequencing is beginning to provide a huge amount of both genomic and expressed sequence data. We believe that the strategy used in this manuscript will be very useful in the next few years to further explore the role of SNVs in alternative splicing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Daniel Ohara for technical support.

FUNDING

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (2007/55790-5 to S.J.S.); Ph.D. fellowship (2007/59721-8 to R.F.R.). Funding for open access charge: Ludwig Institute for Cancer Research.

Conflict of interest statement. None declared.

REFERENCES

- Mironov, A., Fickett, J. and Gelfand, M. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
- Xu, Q., Modrek, B. and Lee, C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, **30**, 3754–3766.
- Modrek, B. and Lee, C. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, **34**, 177–180.
- Resch, A., Xing, Y., Alekseyenko, A., Modrek, B. and Lee, C. (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.*, **32**, 1261–1269.
- Lewis, B., Green, R. and Brenner, S. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189–192.
- Wang, E., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S., Schroth, G. and Burge, C. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
- Galante, P., Sakabe, N., Kirschbaum-Slager, N. and de Souza, S. (2004) Detection and evaluation of intron retention events in the human transcriptome. *RNA*, **10**, 757–765.
- Sakabe, N., de Souza, J., Galante, P., de Oliveira, P., Passetti, F., Brentani, H., Osório, E., Zaiats, A., Leerkes, M., Kitajima, J. *et al.* (2003) ORESTES are enriched in rare exon usage variants affecting the encoded proteins. *C R Biol.*, **326**, 979–985.
- Pagani, F. and Baralle, F. (2004) Genomic variants in exons and introns: identifying the splicing spoilers. *Nat. Rev. Genet.*, **5**, 389–396.
- Wang, Z. and Burge, C. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, **14**, 802–813.
- Zhang, Z., Zhou, L., Wang, P., Liu, Y., Chen, X., Hu, L. and Kong, X. (2009) Divergence of exonic splicing elements after gene duplication and the impact on gene structures. *Genome Biol.*, **10**, R120.
- Zhang, X., Arias, M., Ke, S. and Chasin, L. (2009) Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing. *RNA*, **15**, 367–376.
- Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. and Darnell, R. (2006) An RNA map predicting Nova-dependent splicing regulation. *Nature*, **444**, 580–586.
- Kirschbaum-Slager, N., Parmigiani, R., Camargo, A. and de Souza, S. (2005) Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. *Physiol. Genomics*, **21**, 423–432.
- Galante, P., Vidal, D., de Souza, J., Camargo, A. and de Souza, S. (2007) Sense-antisense pairs in mammals: functional and evolutionary considerations. *Genome Biol.*, **8**, R40.
- Liu, H., Zhang, M. and Krainer, A. (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev.*, **12**, 1998–2012.
- Liu, H., Chew, S., Cartegni, L., Zhang, M. and Krainer, A. (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol. Cell. Biol.*, **20**, 1063–1071.
- Fairbrother, W., Yeh, R., Sharp, P. and Burge, C. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.
- Zhang, X. and Chasin, L. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.*, **18**, 1241–1250.
- Wang, Z., Rolish, M., Yeo, G., Tung, V., Mawson, M. and Burge, C. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.
- Smith, P., Zhang, C., Wang, J., Chew, S., Zhang, M. and Krainer, A. (2006) An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum. Mol. Genet.*, **15**, 2490–2508.
- Kurmangaliyev, Y. and Gelfand, M. (2008) Computational analysis of splicing errors and mutations in human transcripts. *BMC Genomics*, **9**, 13.
- Wang, J., Smith, P., Krainer, A. and Zhang, M. (2005) Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res.*, **33**, 5053–5062.
- Stranger, B., Nica, A., Forrest, M., Dimas, A., Bird, C., Beazley, C., Ingle, C., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.
- Kwan, T., Grundberg, E., Koka, V., Ge, B., Lam, K., Dias, C., Kindmark, A., Mallmin, H., Ljunggren, O., Rivadeneira, F. *et al.* (2009) Tissue effect on genetic control of transcript isoform variation. *PLoS Genet.*, **5**, e1000608.
- Ge, B., Pokholok, D., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D., Le, J., Koka, V., Lam, K., Gagné, V. *et al.* (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.*, **41**, 1216–1222.
- Coulombe-Huntington, J., Lam, K., Dias, C. and Majewski, J. (2009) Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet.*, **5**, e1000766.
- Woolfe, A., Mullikin, J. and Elnitski, L. (2010) Genomic features defining exonic variants that modulate splicing. *Genome Biol.*, **11**, R20.
- Fairbrother, W., Holste, D., Burge, C. and Sharp, P. (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.*, **2**, E268.