DESENVOLVIMENTO DE UM SISTEMA EM LARGA ESCALA PARA O ESTUDO COMPUTACIONAL DE FORMAS DE *SPLICING* ALTERNATIVO DIFERENCIALMENTE EXPRESSAS EM TUMORES E SUA VALIDAÇÃO EXPERIMENTAL

NATANJA SARA KIRSCHBAUM-SLAGER

Tese de Doutorado apresentada a Fundação Antonio Prudente para obtenção de Grau de Doutor em Ciências.

Área de concentração: Oncologia

Orientador: Dr. Sandro José de Souza
Co-Orientadora: Dra. Anamaria Aranha Camargo

São Paulo
2005

FICHA CATALOGRÁFICA

Preparada pela Biblioteca do Centro de Tratamento e Pesquisa

Hospital do Câncer A.C. Camargo

Kirschbaum-Slager, Natanja Sara

Desenvolvimento de um sistema em larga escala para o estudo computacional de formas de *splicing* alternativo diferencialmente expressas em tumores e sua validação experimental. / Natanja Sara Kirschbaum-Slager – São Paulo, 2005.

106p.

Tese(Doutorado)-Fundação Antônio Prudente.

Curso de Pós-Graduação em Ciências - Área de concentração: oncologia.

Orientador: Dr Sandro Jose de Souza

Descritores: 1. PROCESSAMENTO ALTERNATIVO. 2. BIOINFORMÁTICA. EXPRESSÃO GÊNICA. 4. TUMORES/genética. 5. VALIDAÇÃO DE PROGRAMAS DE COMPUTADOR.

AGRADECIMENTOS

Ao meu orientador **Dr. Sandro J. de Souza** por me dar a oportunidade de realizar este trabalho sob sua orientação. Pelas sugestões dadas e pelos momentos de descontração fora do laboratório.

À minha co-orientadora **Dra. Anamaria Aranha Carmargo** por me dar o privilégio de realizar este trabalho sob sua co-orientação. Muito obrigada pela amizade, pelos conselhos em momentos difíceis e pela confiança depositada em mim.

À **Maria D. Vibranovski** pela amizade sincera, por compartilhar todos os momentos deste período em São Paulo, tanto os bons quanto os ruins, e por sempre me apoiar e incentivar. Já estou sentindo saudades! Its your face!!!

Ao **Raphael B. Parmigiani** pela amizade, ajuda e apoio nos momentos mais difíceis. Pelos muitos litros de café que tomamos juntos. Boa sorte no exterior! Você é muito cool!

À **Maria D. Vibranovski, Raphael B. Parmigiani e Noboru Jo Sakabe** pela disponibilidade em ler este trabalho e todos os outros textos em Português, pelas sugestões e conselhos. Sem vocês simplesmente não seria possível!!!

A **Noboru Jo Sakabe, Pedro A.F. Galante, Jorge E.S. de Souza, Robson F. de Souza e Elza H. A. Barbosa** pela amizade, convivência, e pelos bons momentos dentro e fora do laboratório, vou sentir saudades!

À **Ana Cláudia Pereira** pela amizade, ajuda e paciência infinita.

À **Dra. Helena P. Brentani** pela amizade, disponibilidade de sempre ouvir e ajudar, e pelo incentivo constante. Pelos almoços divertidos e por me ensinar o seu jeito de enxergar além.

Ao pessoal do Laboratório de Biologia Computacional: **Arthur, Patrícia, Milton, André Zaiatz e Elisson Osorio** pela assistência técnica.

À **Dra. Dirce M Carraro** pela amizade e colaborações.

À **Fabi, Lílian e Anna Chris** pela amizade, paciência comigo e a ajuda no 'wetlab'.

À toda turma do **laboratório de biologia molecular e genômica** por sempre me ajudar, chamar para os 'journal clubs', festinhas do laboratório, amigos secretos e muito mais. Sempre senti que a 'casinha' era minha segunda casa.

À turma do **laboratório de análise de expressão gênica**, especialmente à Maria Cristina, Elisa e Nádia, pela amizade e colaborações.

À **Márcia Hiratori, Ana Maria Kuninari e Luciana C. Pitombeira** pelos auxílios prestados durante esses anos.

A todos os **funcionários da biblioteca do Hospital do Câncer A.C. Camargo**, em especial à Suely Francisco, pela ajuda prestada na correção da tese.

Ao **Prof. Dr. Ricardo R. Brentani** pela direção do Instituto Ludwig e por permitir a realização deste trabalho.

Ao Dr. **Luís Fernando Lima Reis** pela direção da pós-graduação da Fundação Antônio Prudente e por me dar o privilégio de realizar este trabalho no ILPC.

A todos os **funcionários do ILPC** pelo suporte técnico e administrativo e também pela convivência durante este período.

Aos demais **colegas do ILPC** pela convivência durante esses anos.

Aos **meus pais** pelo amor, confiança e apoio em todas as situações em qualquer lugar do mundo. Fico muito feliz de poder ficar mais perto de vocês em breve.

Ao meu irmão **Joram**. Você me ajudou muito no meu caminho e sempre me apóia, ajuda no que pode e dá sugestões muito importantes. À minha cunhada **Hadassa e aos meus sobrinhos** maravilhosos pelo apoio e por me darem tanta alegria.

À minha sogra **Fruma** e seu marido **Bensi**, às minhas cunhadas. a 'bobe' e toda **minha família aqui no Brasil** pelo apoio em TODOS os momentos e pelo carinho, por me tratarem como filha, irmã, neta, sobrinha, etc. Foi uma experiência inesquecível conviver com vocês e compartilhar todos os momentos do nosso tempo aqui. Esperamos vocês lá....

A todos os **amigos e à família na Holanda e em Israel** por sempre me apoiarem, se interessarem e estarem próximos mesmo tão longe.

Ao **Beno** por nunca parar de acreditar em mim. por me mostrar o mundo. Obrigada por ser quem você é.

Ao **Hans,** por ter nos acompanhado ao longo de nosso caminho. A sua contribuição para este trabalho foi indispensável.

À Comissão de Aperfeiçoamento de Pessoal de Nível Superior (**CAPES**) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (**FAPESP**) pelas bolsas concedidas.

RESUMO

Kirschbaum-Slager NS. **Desenvolvimento de um sistema em larga escala para o estudo computacional de formas de "*splicing*" alternativo diferencialmente expressas em tumores e sua validação experimental.** São Paulo; 2005. [Tese de Doutorado-Fundação Antônio Prudente].

O *splicing* alternativo é uma das maiores fontes de diversidade genética e a caracterização desta variabilidade é fundamental para que se possa decifrar o *transcriptoma*. Foi demonstrado que certos genes são alternativamente processados em tecidos tumorais e suas variantes são associadas a progressão tumoral e invasão em diferentes tumores. Portanto, o entendimento da associação do *splicing* alternativo ao câncer possui um grande valor diagnóstico e terapêutico e gerará um entendimento mais amplo da regulação e envolvimento do *splicing* alternativo no tumorigênese. Combinamos uma análise computacional de dados de expressão gênica com validações experimentais para desenvolver um sistema capaz de selecionar exons representados predominantemente em amostras tumorais de diferentes tecidos em relação aos respectivos tecidos normais. Uma análise estatística foi desenvolvida para calcular a probabilidade de um exon estar realmente associado a tumor e vários critérios foram adotados para diminuir a probabilidade de selecionar candidatos falso-positivos. Assim, foram selecionados 1295 genes contendo 2878 exons com maior expressão em amostras tumorais que foram denominados exons associados a tumor. A validação de alguns candidatos foi feita por RT-PCR e

demonstrou que nossa lista de candidatos incluía casos de exons pertencentes a genes que eram super expressos em tumores de uma maneira geral e independente da variante de *splicing*. A seleção destes casos não era o objetivo da nossa busca, uma vez que procurávamos variantes que pudessem representar casos em que o *splicing* alternativo fosse diferencialmente regulado em câncer. Para aumentar a probabilidade de encontrar variantes de *splicing* que são realmente associadas a tumor, realizamos uma análise de *SAGE* (Serial Analysis of Gene Expression) para excluir genes que são super expressos em tumores específicos. O grupo final de candidatos inclui 1386 exons pertencendo a 638 genes. Em linhagens celulares tumorais foram confirmados 4 de 10 exons como super-expressos em tumores, cujos protótipos dos mesmos genes (que não possuem os exons) não foram super expressos em tecido tumoral. Em amostras tumorais de pacientes, 5 de 6 exons validados experimentalmente foram confirmados como sendo associados a tumor. Classificações funcionais dos genes candidatos demonstraram que nossa lista final está enriquecida com genes relacionados funcionalmente com câncer. Os candidatos foram validados também através de comparações com trabalhos publicados. O presente trabalho demonstra a importância da combinação de sistemas de seleção computacionais com validações experimentais. Análises experimentais em larga escala validarão até que nível nossos exons candidatos diferencialmente expressos têm um potencial diagnóstico ou terapêutico.

SUMMARY

Kirschbaum-Slager N.S. **Development of a large scale system for the computational study of differentially expressed alternative splicing forms in tumor and its experimental validation.** São Paulo; 2005. [Tese de Doutorado-Fundação Antônio Prudente].

Alternative splicing is one of the major sources of the transcriptional diversity found in human cells and its characterization is fundamental to decipher the human transcriptome. Certain genes have been shown to be alternatively spliced in tumor tissues and their isoforms have been shown to be associated with spreading and progression in several human tumors. Therefore, understanding the association between alternative splicing and cancer is of great diagnostic and therapeutic value and will generate a broader understanding of the involvement and regulation of alternative splicing in tumorigênesis. We combined the use of a transcriptome database for a computational analysis of gene expression data with experimental validations in order to develop a system capable of selecting exons that are predominantly represented in tumor samples of different tissues as compared to their respective normal counterparts. A statistical analysis was developed to calculate the probability that an exon was indeed tumor associated and various criteria were defined to decrease the probability of selecting false-positive candidates. This way we selected 1295 genes containing 2878 exons having an elevated expression level in tumor samples, which we called tumor-associated

exons. The validation of a few candidates was performed by RT-PCR and showed that our list of candidates included cases of exons belonging to genes that are over-expressed in tumors in general, independently of their splicing variants. The selection of such cases was not the object of our study as we were looking for tumor associated variants that could represent cases of differentially regulated alternative splicing in cancer. To increase the probability of finding bona fide regulated splicing variants that were really tumor-associated, we performed a Serial Analysis of Gene Expression (*SAGE*) analysis, excluding those genes that are up-regulated in specific tumors. Our final group of candidates included 1386 exons belonging to 638 genes. In tumor cell lines, 4 of 10 validated exons were confirmed as over expressed in tumor while their prototype variants (that don't include the candidate exons) were not over expressed in tumor. In patient tumor samples, 5 of 6 experimentally validated exons were confirmed to be tumor associated. Functional classification of our candidate genes showed that our final list is slightly inflated with cancer-related genes. We validated our candidates once more by comparing them with published studies. Our work shows the importance of the combination of computational selection systems with experimental validations. Large scale experimental analyses will validate to what extent our candidate exons might be of therapeutic or diagnostic use.

PREFÁCIO

Em meados do ano 2000, um grupo de pesquisa denominado Consórcio Internacional de Seqüenciamento do Genoma Humano (The International Human Genome Sequencing Consortium), juntamente com uma empresa do setor privado chamada Celera Genomics anunciou oficialmente ter desvendado a maior parte da seqüência do genoma humano (LANDER et al. 2001; VENTER et al. 2001). Esta conquista científica marcou o fim de um trabalho que começou no início da década de 90. A seqüência anunciada foi chamada de esboço (*draft)* e foi a primeira de várias versões até a publicação da versão final no ano de 2004 (International Human Genome Sequencing Consortium 2004).

Embora o seqüenciamento do "livro da vida" tenha fornecido respostas a inúmeras perguntas que não eram possíveis de serem respondidas na era pré-genômica, a seqüência completa gerou espaço para muitas novas questões. Foram identificados 3,2 bilhões de nucleotídeos que correspondem às "letras" deste livro. No entanto, ainda não se sabia quantas "palavras", os genes, eram codificadas por estas letras. Para responder a esta pergunta, foram utilizados programas computacionais de predição gênica que usam seqüências conservadas e características dos genes humanos já conhecidos para predizer onde outros genes estão localizados. Os dois grupos de pesquisa realizaram uma análise da seqüência obtida e estimaram que todo o genoma humano contém aproximadamente 35.000 genes (LANDER et al. 2001; VENTER et al. 2001). A partir daquele momento surgiram novos desafios: a necessidade de validar experimentalmente a existência dos 35.000 genes e entender quais destes eram expressos em determinadas condições

celulares. Por exemplo, quais fatores são capazes de diferenciar uma célula tronco em uma célula hepática ou uma célula cerebral? Por que uma célula passa por um processo celular específico e a outra não?

Para responder a estas e outras perguntas, foram iniciados outros campos de pesquisa: do *Transcriptoma* – com o intuitivo de decifrar quais genes estão expressos em diferentes condições e tipos celulares; do *Reguloma* – para entender como a expressão gênica é regulada e quais fatores são responsáveis por esta regulação; e por fim, do *Proteoma* – para entender quais proteínas são realmente geradas a partir de genes específicos.

O genoma contém, além de regiões codificantes para proteínas, regiões com funções desconhecidas. Uma pergunta importante que ainda persiste é: quais partes do genoma representam genes que codificam proteínas, e quais regiões são intergênicas? Originalmente, as regiões intergênicas foram chamadas de "DNA lixo" traduzido do inglês *junk DNA*. Porém, hoje em dia sabe-se que estas regiões representam até 97% de todo o genoma (VENTER et al. 2001) e provavelmente não representam "lixo" genômico, mas têm um papel importante na regulação da transcrição de regiões gênicas. Existem também regiões não codificadoras, localizadas dentro de genes chamadas de introns. As regiões dos genes que são transcritos em mRNA e das quais uma parte codifica proteínas são chamadas de exons (LANDER et al. 2001).

Para gerar uma proteína a partir de uma seqüência de DNA, este tem que ser transcrito para gerar o pré-mRNA (RNA imaturo). O mRNA imaturo é então processado para formar o mRNA maduro (mRNA). Este processamento (do inglês *splicing*) envolve a retirada dos introns e a ligação dos exons um ao outro.

Além deste processo, o RNA sofre dois outros tipos de alteração: a modificação da extremidade 5' dos transcritos pela adição de uma seqüência *cap*, que consiste em um nucleotídeo de guanina modificado e um grupo metil (m^7Gppp). O terceiro evento envolve a clivagem do RNA em um ponto específico da extremidade 3' da molécula de RNA e a adição de aproximadamente 200 nucleotídeos de Adenina (cauda poli-A). Por fim, o RNA maduro é transportado do núcleo para o citoplasma, onde serão geradas proteínas a partir deste mRNA nos ribossomos (ALBERTS et al. 2002).

Durante os estudos do genoma mencionados anteriormente (LANDER et al. 2001; VENTER et al. 2001) foi verificado que na realidade existe uma quantidade menor de genes no genoma humano (cerca de 35.000 genes) do que havia sido inicialmente estimado a partir do agrupamentos de seqüências parciais de cDNA que indicaram a existência de aproximadamente 80.000 genes {Liang, 2000 305 /id}. Além disso, o número de genes identificados a partir da seqüência genômica humana era similar ao número de genes encontrados em organismos menos complexos como *Arabidopsis thaliana* {2000 306 /id} e *C. elegans* {1998 286 /id}. Uma das possíveis explicações para esta discrepância pode ser a existência de mecanismos de *splicing alternativo*. Através desses mecanismos uma grande parte dos genes humanos pode dar origem a vários transcritos de mRNA diferentes, chamados variantes de *splicing*. Desta forma, a partir de um determinado número inicial de genes, é possível obter uma diversidade muito maior de seqüências de mRNA e conseqüentemente de proteínas o que explicaria o número relativamente baixo de genes encontrados no genoma humano.

O processo de *splicing* alternativo é considerado uma das maiores fontes de diversidade genética (BLACK 2003), uma vez que permite a produção de diversos transcritos a partir de um gene único. A caracterização desta variabilidade é fundamental para que se possa decifrar todos os transcritos expressos numa determinada condição e permitir um minucioso entendimento do genoma humano. Além disso, tem sido mostrado que o *splicing* alternativo pode estar envolvido no desenvolvimento de várias doenças. Especificamente, já foi demonstrado que certos genes são alternativamente processados em tecidos tumorais. Assim, a caracterização destas variantes de *splicing* possui um grande valor diagnóstico e terapêutico.

O foco do meu projeto foi a identificação de variantes de *splicing* alternativo associadas ao câncer, utilizando uma combinação entre a biologia computacional e a validação experimental. O presente trabalho relata os principais resultados obtidos neste projeto. Uma vez que a maioria dos resultados gerados diretamente a partir do meu projeto encontra-se publicada, decidimos integrar o artigo publicado na sessão "Resultados" para facilitar a leitura da tese. Além disso, apresentaremos também os resultados e os respectivos métodos que não se encontram em formato de manuscrito. Por fim, os resultados obtidos serão discutidos na sessão de "Discussão".

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

µg	Micrograma
µl	Microlitro
µM	Micromolar
ASF/SF2	*Alternative splicing factor/splicing factor 2*
ATCC	*American Type Culture Collection*
Bcl-2	*B-cell CLL/lymphoma 2*
BLAST	*Basic Local Alignment Search Tool*
Brca1 e 2	*Breast Cancer 1 e 2*
Bin1	*Bridging integrator 1*
cDNA	DNA complementar
dbEST	Banco de dados de *ESTs*
DNA	Ácido desoxirribonucléico
dNTP	Deoxinucleotídeo
domínio SR	Domínio serina arginina
dscam	*Down syndrom cell adhesion molecule*
DTT	Ditiotreitol
EJC	Complexo da junção de exons (do inglês *Exon Junction Complex*)
ESE	*Exonic splicing enhancer*
ESS	*Exonic splicing silencer*
EST	Etiqueta de seqüência expressa (do inglês *Expressed Seqüence Tags*)
g	Grama

GAPDH	Gliceraldeído 3-fosfato desidrogenase
GO	Ontologia Gênica (do inglês *Gene ontology*)
hMLH-1	*Human mut-L homologue-1*
hnRNP	*Heterogeneous nuclear ribonucleoprotein*
hTERT	*Human telomerase reverse transcriptase*
hTid-1	*Human tumorous imaginal discs homolog*
ISE	*Intronic splicing enhancer*
ISS	*Intronic splicing silencer*
M	*Molar*
ml	Mililitro
mM	Milimolar
mRNA	RNA mensageiro
NCBI	Centro Nacional para Informação de Biotecnologia
ng	Nanograma
nm	Nanômetro
nM	Nanomolar
NMD	*Nonsense mediated decay*
nr	Não redudante
ORESTES	*Open Reading Frame ESTs*
pb	Pares de base
PCR	Reação em cadeia da polimerase (do inglês *Polimerase Chain Reaction*)
PTB	*Polypyrimidine tract binding protein*
PTPN6	*Protein tyrosine phosphatase, non-receptor type 6*

RNA	Ácido ribonucléico
RNAi	RNA de interferência
RT	Transcrição reversa (do inglês *Reverse Transcriptase*)
SAFB	*Scaffold attachment factor B*
SAGE	*Serial Analysis of Gene Expression*
SELEX	*Systematic evolution of ligands by exponential enrichment*
SNP	*Single nucleotide polimorphism*
snRNP	*Small nuclear riboneucleoproteins*
syk	*Spleen tyrosine kinase*
TAE	Tris Acetato EDTA (do inglês *Tris Acetate EDTA*)
TIGR	O instituto para pesquisa genômica (do inglês *The institute for genomics research)*
Tsg101 delta	*Tumor susceptibility gene 101*
tubd1	*Delta Tubulin 1*
U	Unidade
U2AF	*U2 auxiliary factor*
V	Volume
wt1	*Wilms Tumor 1*

ÍNDICE

INTRODUÇÃO

1 INTRODUÇÃO

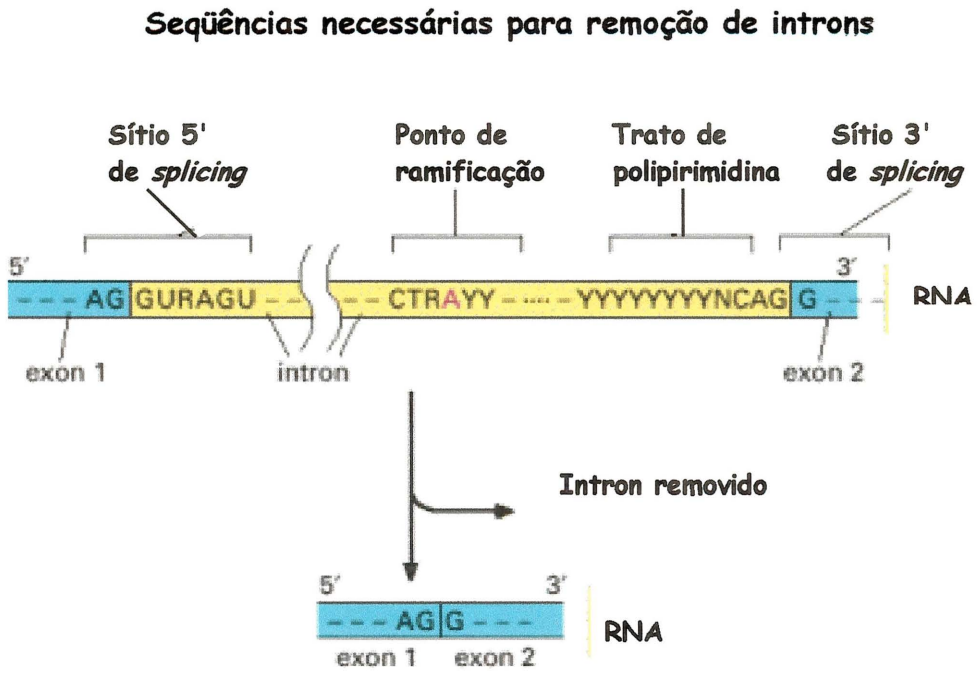
1.1 *SPLICING*: O PROCESSAMENTO DE PRÉ-mRNA

1.1.1 O processo de *splicing*

Até o ano de 1977, acreditava-se que todos os genomas eram compostos apenas por seqüências transcritas, como o observado nos genomas mais conhecidos naquela época: o das bactérias. Posteriormente, foi descoberto que em genomas mais complexos (como o genoma humano), as seqüências intragênicas não codificantes, os introns, são removidas da molécula de RNA mensageiro (mRNA) imaturo para formar a molécula de mRNA madura (SAMBROOK 1977). Este fenômeno responsável pelo processamento do mRNA é chamado de "*splicing*".

Durante cada evento de *splicing*, um intron é removido através de duas reações de trans-esterificação (substituição de uma ligação fosfodiéster por outra), unindo dois exons. Existem seqüências específicas nas junções intron/exon que são chamadas de sítios de *splicing*. Na extremidade 5' do intron geralmente (em 98,71% dos casos, (BURSET et al. 2000) são encontrados os nucleotídeos conservados GU. Já na porção 3' existem três seqüências conservadas: o ponto de ramificação *(branch point*), o trato de polipirimidinas (*polypyrimidine tract*) e finalmente, na extremidade 3' do intron estão localizados os nucleotídeos AG na maioria dos casos (Figura 1). As outras posições (também o *ponto de ramificação*) podem ser ocupadas por uma variedade de nucleotídeos, embora os nucleotídeos indicados na Figura 1 sejam os mais freqüentes. A distância ao longo da molécula de RNA entre o ponto de

ramificação e o trato de polipirimidina é bastante variável. No entanto, a distância entre o ponto de ramificação e a junção 3' de *splicing* é normalmente muito mais curta do que a distância entre a junção 5' e o ponto de ramificação.



Fonte: Adaptada de ALBERTS et al. (2002).
Legenda: Quatro seqüências de nucleotídeos são necessárias para a remoção de uma seqüência intrônica. R refere-se aos nucleotídeos A ou G; Y refere-se aos nucleotídeos C ou U. O nucleotídeo A marcado em vermelho representa o *ponto de ramificação* onde a extremidade 5' do intron é esterificada, formando um laço.

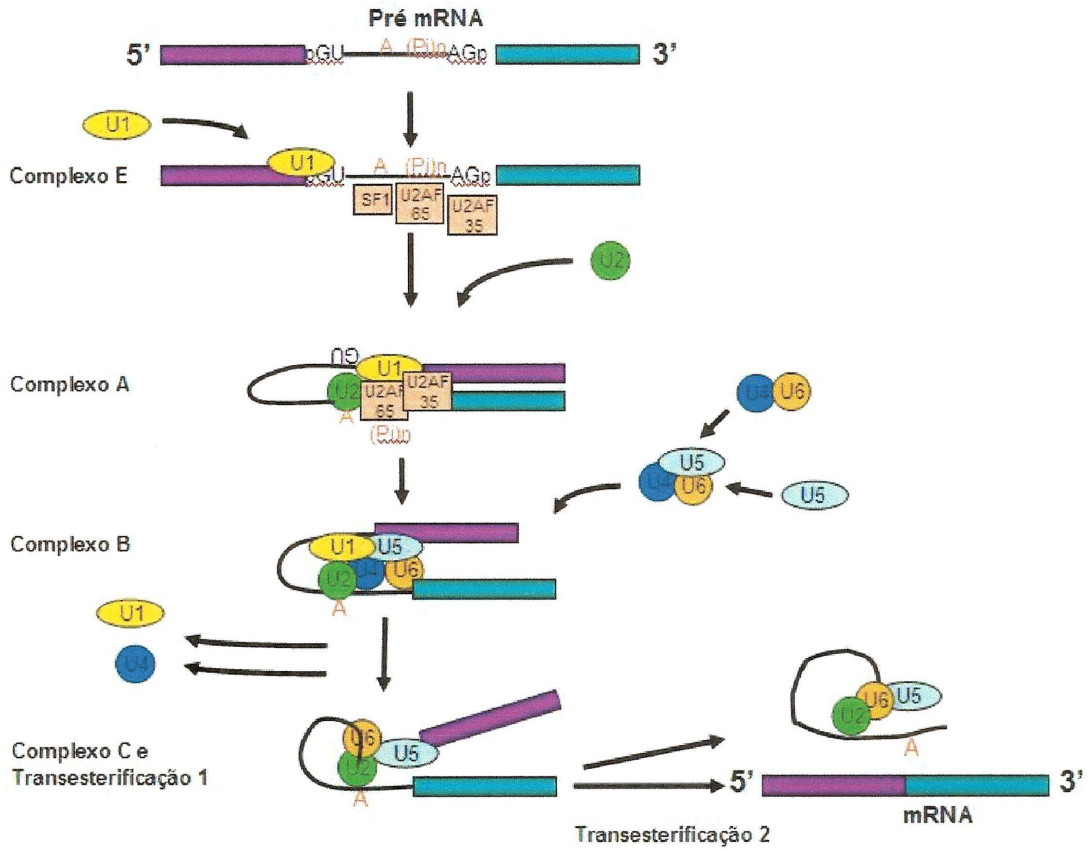
Figura 1 - Seqüências consenso encontradas na maioria dos introns humanos.

Depois da transcrição do DNA em pré-mRNA, diferentes proteínas chamadas de partículas ribonucleoprotéicas heterogêneas (*heterogeneous ribonucleoprotein particles - hnRNPs*) se associam à molécula de RNA formando o denominado complexo H. Após esta associação, outro complexo macromolecular chamado "spliceossomo" se liga em etapas às seqüências conservadas do pré-mRNA

mencionadas acima. Este complexo é formado por basicamente cinco moléculas de pequenas proteínas ribonucléicas nucleares (*small nuclear riboneucleoproteins (snRNPs)*). As snRNPs são constituídas por moléculas curtas de RNA com menos de 200 nucleotídeos cada uma (snRNAs – *small nuclear RNAs*) e associadas a pelo menos sete subunidades de proteínas. Além destas cinco snRNPs, se ligam ao spliceossomo outras proteínas que parecem ter influência sobre a regulação do processo de *splicing*, chamadas de fatores de *splicing*.

O processo de *splicing* ocorre em 4 fases (ver Figura 2) seguidas de duas reações de trans-esterificação. –Na primeira fase o snRNP U1 se liga à extremidade 5' do intron a ser retirado, através de pareamento de bases com as seqüências conservadas nesta região. –Na segunda fase, na região 3' do intron, a proteína SF1 se liga ao ponto de ramificação, a subunidade 65-kDa do fator auxiliar U2AF (*U2 auxiliary factor*) se liga ao trato de polipirimidina e a subunidade 35-kDa do mesmo fator se liga aos nucleotídeos AG da junção intron/exon. Este complexo do pré-mRNA junto com as proteínas associadas é chamado de complexo E (do inglês "*Early*"). –Na terceira fase, forma-se o complexo A no qual a snRNP U2 se liga ao ponto de ramificação. –Na quarta e última fase, ocorre formação do complexo B, no qual o tri-snRNP U4/U5/U6 se liga ao complexo de snRNPs U1 e U2 através de interações proteína-proteína, formando assim o spliceossomo. Para formar o complexo C, que catalisa a primeira e segunda fases da clivagem por reações de trans-esterificação, os snRNPs U1 e U4 se desligam do complexo e o snRNP U6 se liga no lugar do snRNP U1 na extremidade 5' do intron. Os snRNAs U2 e U6 formam uma estrutura tridimensional na qual o sítio 5' de *splicing* do pré-mRNA é colocado próximo ao sítio do ponto de ramificação, o que leva à primeira reação de trans-

esterificação. Da mesma maneira, as junções 5' e 3' são aproximadas pelo snRNA U5, o que possibilita a segunda trans-esterificação (ALBERTS et al. 2002).



Legenda: O spliceossomo contém cinco snRNPs que se associam ao intron a ser removido. O complexo E contém o snRNP U1 ligado ao sítio de *splicing* 5' e as proteínas SF1 ligadas ao ponto de ramificação, o $U2AF^{65}$ ligado ao trato de polipirimidina ((Pi)n) e o $U2AF^{35}$ ligado ao dinucleotídeo AG na extremidade 3'. O complexo A é formado quando o snRNP U2 se liga ao ponto de ramificação no lugar do SF1. Quando o Tri-snRNP U4/5/6 é incluído, o complexo B é formado e passa por uma complexa mudança, excluindo os snRNPs U1 e U4, para formar o complexo C catalítico. O *splicing* acontece por duas reações de trans-esterificacão. A primeira reação resulta em duas estruturas: o exon 5' livre e o fragmento intron/3'-exon formando uma estrutura de laço. A segunda reação é responsável pela ligação entre os dois exons e libera o laço intrônico.

Figura 2 - O mecanismo de *splicing* do pré-mRNA

A troca do snRNP U1 por snRNP U6 permite uma grande precisão na determinação do sítio de *splicing* 5' pois estas duas proteínas precisam reconhecê-lo independentemente. O sítio do ponto de ramificação é reconhecido primeiro pela

proteína SF1 e subseqüentemente pela snRNP U2. Novamente, isto permite um nível alto de precisão do processo (ALBERTS et al. 2002; NILSEN 2003). Desta maneira, os snRNPs são responsáveis pelo reconhecimento preciso dos nucleotídeos e pela configuração do sítio catalítico necessário para as duas reações de trans-esterificação.

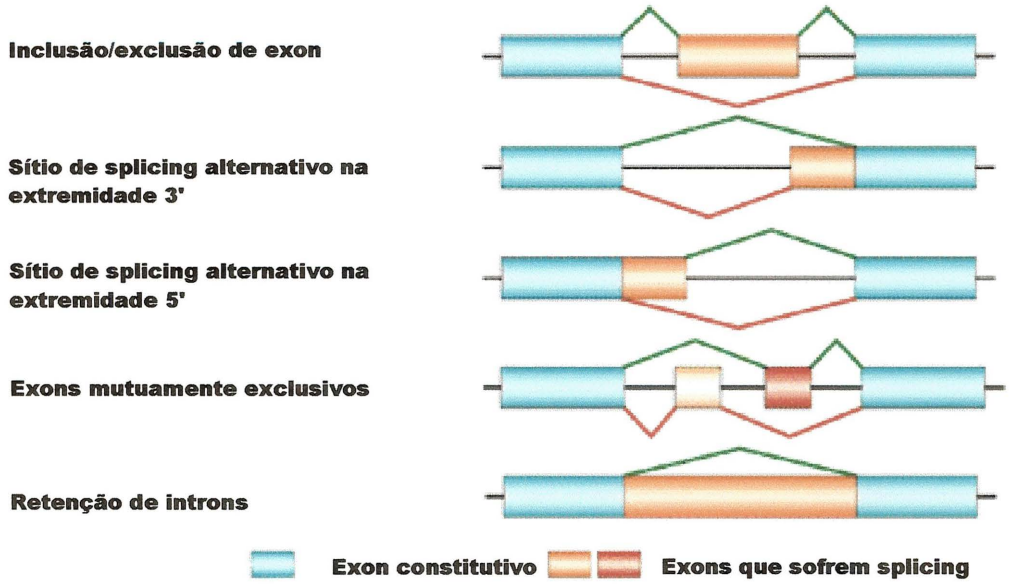
Já foi demonstrado que as proteínas de processamento de mRNA se ligam à molécula de RNA polimerase II no início da transcrição (MISTELI e SPECTOR 1999). O movimento da polimerase II ao longo da molécula de pré-mRNA ajuda a definir os introns e exons: os snRNPs no lado 5' do sítio de *splicing* podem se ligar somente a um sítio único de *splicing* 3', uma vez que outros sítios de *splicing*, mais *downstream*, só serão transcritos posteriormente (ALBERTS et al. 2002; KORNBLIHTT et al. 2004). Assim, componentes do spliceossomo se ligam às seqüências conservadas localizadas nas extremidades de um exon, estimulando a excisão dos introns flanqueadores. A ligação das proteínas SR (proteínas contendo vários dipeptídeos de arginina e serina, formando o domínio RS) aos exons, auxilia na ligação dos componentes do spliceossomo e permite a ligação na ordem correta dos exons um ao outro (IBRAHIM et al. 2005). Este processo é chamado *definição do exon* (BERGET 1995) e ocorre antes da formação de interações que definem as extremidades do intron a ser excluído. A marcação do sítio de *splicing* e a definição dos exons e introns começam então já na fase da transcrição, enquanto que as reações de trans-esterificação podem ocorrer muito mais tarde.

Além do *splicing* do mRNA, durante a transcrição ocorre também a adição de uma seqüência *cap* na extremidade 5' que permite o começo de transcrição pela RNA polimerase II e a posterior adição da cauda de poli A na extremidade 3' do transcrito. Após todos estes eventos, a molécula final formada é chamada de mRNA maduro.

1.1.2 *Splicing* Alternativo

Após a descoberta da existência de introns e exons, foi proposta a possibilidade de que combinações diferentes de exons poderiam levar à geração de diferentes isoformas do mesmo gene (GILBERT 1978). Posteriormente foi demonstrado que isto acontece em muitos casos e diferentes mRNAs maduros podem ser produzidos a partir de um único gene, devido ao uso de sítios de *splicing* alternativos. Um exemplo extraordinário de *splicing* alternativo é o que ocorre com o gene *dscam* em *Drosophila Melanogaster*, a partir do qual podem ser gerados mais de 38.000 variantes (SCHMUCKER et al. 2000). Através do uso diferencial de exons, é possível gerar isoformas distintas de mRNA e conseqüentemente é possível a geração de proteínas com estruturas e até funções diferentes.

Existem quatro tipos de *splicing* alternativo (Figura 3): 1) exons podem ser excluídos de um determinado transcrito (uso alternativo de exons); 2) sítios doadores e/ou aceptores crípticos podem ser usados ao invés dos sítios de *splicing* originais, o que gera exons mais curtos ou longos; 3) exons podem sofrer *splicing* mutuamente exclusivo (os dois exons nunca serão encontrados simultaneamente no mesmo transcrito); 4) seqüências intrônicas podem ser mantidas em alguns transcritos (retenção de intron); O uso alternativo de exons parece ser o tipo de *splicing* alternativo mais freqüente. Diferentes estudos demonstraram que este tipo de *splicing* alternativo ocorre em 35-59% dos genes (BRETT et al. 2000; HIDE et al. 2001; MODREK et al. 2001) e a retenção de introns ocorre em 5-14,8% dos genes (KAN et al. 2002; GALANTE et al. 2004).



Fonte: Adaptada do site http://med.stanford.edu/sgtc/research/images/altern_*splicing*.gif
Legenda: Os retângulos coloridos representam os exons que são alternativamente processados. As linhas vermelhas que ligam retângulos (exons) mostram quais exons serão combinados depois dos diferentes tipos de *splicing* alternativo. As linhas verdes mostram a ligação de exons depois de *splicing* constitutivo.

Figura 3 - Tipos de *splicing* alternativo.

Vários fenômenos biológicos complexos parecem depender de *splicing* alternativo. O dimorfismo sexual em *Drosophila melanogaster* envolve um padrão bastante complexo de *splicing* alternativo em genes-chave, como o gene *Transformer* (BOGGS et al. 1987). Os reguladores da determinação sexual em *Drosophila* são proteínas de ligação ao RNA que alteram o padrão de *splicing* de transcritos específicos.

O *splicing* alternativo também está envolvido na detecção de diferentes freqüências de som em mamíferos. Existem 576 formas possíveis de *splicing* alternativo do mRNA que codificam um canal de potássio localizado no ouvido interno de aves (BLACK 1998). Estas variantes são expressas em um gradiente ao

longo de 10.000 células receptoras, o que permite a percepção de freqüências diferentes de som.

O *splicing* alternativo também está envolvido na regulação da morte celular programada – a apoptose. Foi mostrado que diferentes isoformas do gene *asap* possuem um papel importante na ligação dos processos de apoptose e *splicing* (SCHWERK e SCHULZE-OSTHOFF 2005). Desta maneira, se torna evidente que o *splicing* alternativo pode exercer um papel fundamental em processos que necessitam de um controle preciso da diferenciação celular e da ativação de vias específicas de desenvolvimento.

Os primeiros estudos sobre o genoma humano, baseados essencialmente em programas de predição gênica, mostraram um número relativamente pequeno de genes (30.000-40.000 genes, (LANDER et al. 2001; VENTER et al. 2001)) comparado ao número de genes encontrados em organismos menos complexos como *Drosophila* (14.000 genes) (ADAMS et al. 2000) e *Caenorhabditis elegans* (19.000 genes) (C. Elegans Sequencing Consortium 1998). Diante destas observações, foi sugerido que o *splicing* alternativo seria responsável pela complexidade encontrada no genoma humano e pela geração da diversidade transcricional a partir de um número relativamente baixo de genes (MODREK et al. 2002; BLACK 2003). Além disso, o *splicing* alternativo é um importante mecanismo de modulação da função gênica, podendo alterar a sua expressão, função (através de mudanças da posição do códon de terminação e da remoção de domínios estruturais) e localização em diferentes tecidos e estágios de desenvolvimento (BLACK 2003).

1.1.3 A regulação do processo de *splicing*

Como outros processos de controle de expressão gênica, a regulação do *splicing* requer componentes que atuam em *cis*, localizados dentro do pré-mRNA e componentes que atuam em *trans*, formados por fatores celulares (RNA ou proteínas) que não estão localizados na seqüência do pré-mRNA a ser processado, como por exemplo, os componentes do spliceossomo e os fatores de *splicing* (SMITH e VALCARCEL 2000). A presença das seqüências conservadas nos sítios de *splicing* geralmente não é suficiente para a determinação da ocorrência de *splicing* (LIM e BURGE 2001). Dentro de exons e introns estão localizadas outras seqüências reguladoras *cis* que podem ajudar no reconhecimento dos sítios de *splicing* pelo spliceossomo e pelos fatores de *splicing*. O exon pode conter dois tipos de seqüências: aquelas que favorecem o *splicing* (ESE – *Exonic Splicing Enhancers*) e aquelas que inibem o *splicing* (ESS – *Exonic Splicing Silencer*); os mesmos dois tipos de seqüências também são encontrados nos introns (ISE – *Intronic Splicing Enhancer*, ISS – *Intronic Splicing Silencer*). As seqüências reguladoras localizadas nos introns (ISE e ISS) foram menos caracterizadas do que as dos exons. Elas podem estar situadas próximas ao sítio de *splicing* ou podem atuar a centenas de nucleotídeos de distância do exon que sofrerá *splicing* (BLACK 2003).

O tipo de fator de *splicing* juntamente com a seqüência à qual ele se liga determinam se o *splicing* de dois exons será induzido ou inibido. Deste modo, os fatores de *splicing* podem ativar ou inibir o processamento de um determinado exon, levando à sua inclusão ou exclusão no transcrito. Geralmente, diferentes fatores são encontrados juntos, formando complexos de regulação, que podem resultar em regulação positiva ou negativa.

Para analisar quais são estes fatores de *splicing* são feitas purificações de spliceossomo através de ensaios de afinidade e espectrometria de massa. Diferentes estudos encontraram até 300 proteínas associadas ao spliceossomo (ZHOU et al. 2002; JURICA e MOORE 2003; NILSEN 2003). Os fatores de *splicing* podem ser divididos em grupos funcionais, baseado em suas funções e em seus domínios protéicos. Os dois grupos funcionais mais conhecidos são a família de proteínas SR, que em geral se ligam a seqüências *ESE* e aumentam a taxa de *splicing,* e as proteínas hnRNP cuja maioria se liga às seqüências *ESS* e portanto inibe a excisão de introns (MATLIN et al. 2005).

O efeito final da ligação de um fator de *splicing* é dependente do contexto celular e do lugar onde ele se ligará. (BLACK 2003). Por exemplo, além da ativação do processo de *splicing*, as proteínas SR podem se ligar às seqüências ISS e inibir o processo de *splicing*.

1.2 *SPLICING* E CÂNCER

1.2.1 A genética do câncer

A transformação de uma célula normal em uma célula cancerosa normalmente ocorre através de seis mudanças principais (HANAHAN et al. 2000). São elas: 1) capacidade de crescimento autônomo; 2) insensibilidade a sinais inibitórios de crescimento; 3) evasão de sinais apoptóticos intrínsecos; 4) potencial proliferativo ilimitado; 5) capacidade de promover angiogênese; 6) competência para invasão tecidual e formação de metástases.

Estas mudanças estão relacionadas a alterações genéticas e epigenéticas que ocorrem em genes associados a processos de proliferação, diferenciação e regulação do ciclo celular. Essas alterações podem ocorrer em proto-oncogenes levando à formação de um oncogene com função tumorigênica. Estas mudanças também podem causar a perda de função de um gene supressor de tumor, que deixa de exercer a função que prevenia a formação de tumores. O resultado destas mudanças é uma perda do equilíbrio entre proliferação e morte celular, permitindo o desenvolvimento do câncer. Além disso, as células podem ganhar um potencial de invasão de outros tecidos. Desta forma, a quantidade de células tumorais aumenta e ao mesmo tempo as células começam a invadir a camada basal do órgão no qual elas se encontram. Posteriormente, as células podem cair na circulação sanguínea e fixar-se em um outro órgão, configurando uma metástase.

A pesquisa genômica vem desenvolvendo estratégias para investigar os diferentes processos que ocorrem durante a tumorigênese a fim de desenvolver ferramentas para o diagnóstico, prognóstico e tratamento do câncer. A caracterização de genes mutados ou cuja expressão é alterada em tumores é de grande importância. Alguns exemplos conhecidos são os genes *p53*, que é mutado em mais de 50% dos tumores humanos (SIGAL e ROTTER 2000), c-*erbB2*, que é freqüentemente amplificado em câncer de mama (ROSS et al. 2003) e o gene supressor de tumor *Rb* cuja inativação por mutação foi mostrada em diferentes tipos de câncer como retinoblastoma, carcinoma de pulmão de células não pequenas e sarcomas osteogênicos (CLASSON e HARLOW 2002).

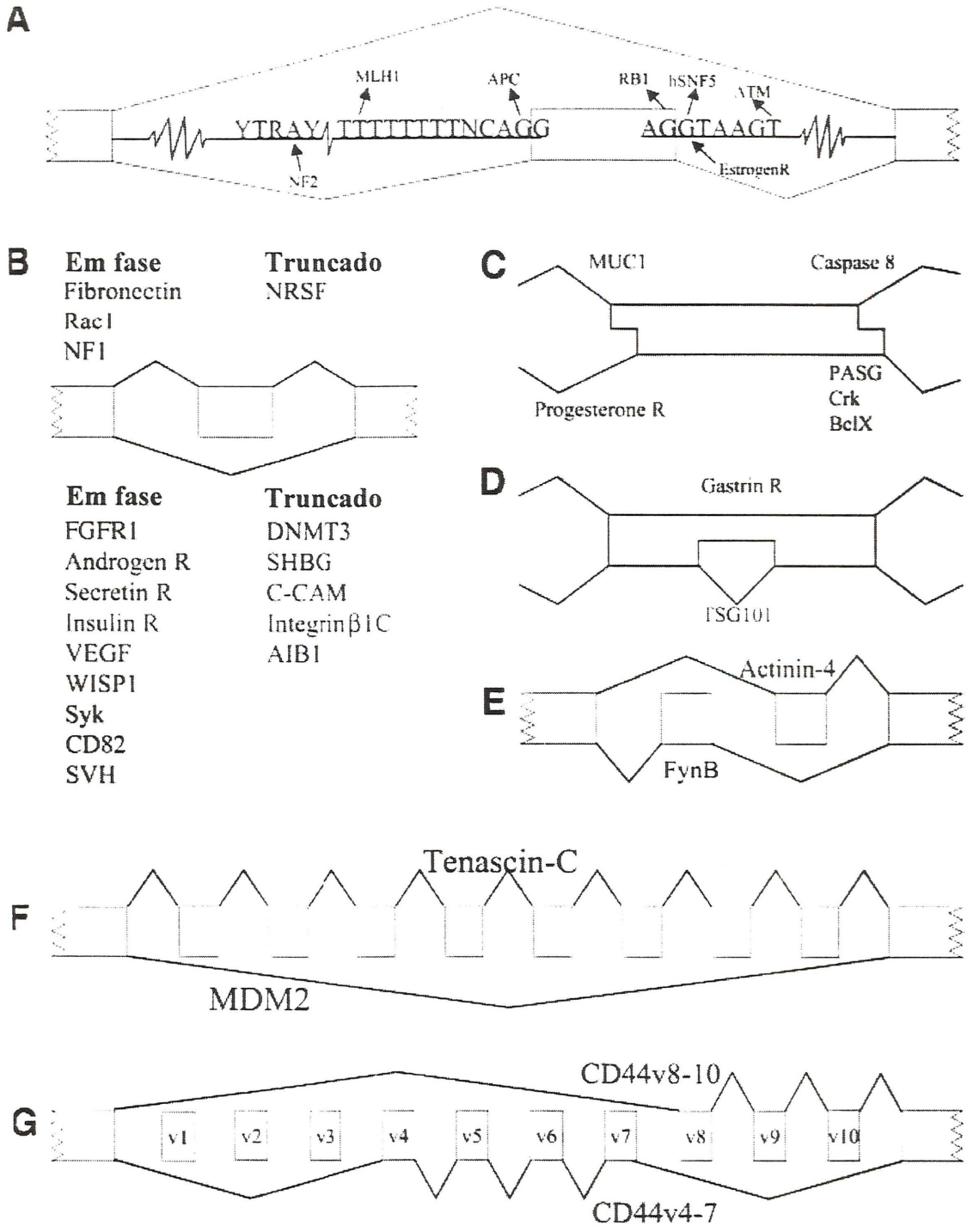
Devido à heterogeneidade dos tipos de câncer, a busca por genes alterados em tumores requer um entendimento de contextos biológicos extremamente complexos.

Um fenômeno biológico que adiciona uma variante importante a esta complexidade é o processo de *splicing* alternativo.

1.2.2 *Splicing* alternativo e câncer

Já foi demonstrado que o *splicing* alternativo está envolvido em diferentes doenças (GARCIA-BLANCO et al. 2004). Também há evidências suficientes que suportam uma correlação direta entre a expressão de determinadas variantes de *splicing* e o processo de tumorigênese (VENABLES 2004). Tais variantes são derivadas de todos os tipos de *splicing*: inclusão e exclusão de exons, sítios crípticos nas extremidades 3' e 5' do sítio de *splicing*, retenção de introns e *splicing* associado a tumor de exons mutuamente exclusivos (ver Figura 4).

Um estudo analisou a expressão de variantes de 100 diferentes eventos de *splicing* em tecido normal e em células de limfoma Hodgkin utilizando uma plataforma de *microarray*. Este estudou demonstrou que existe uma mudança no padrão de expressão das isoformas analisadas nas células transformadas (RELOGIO et al. 2005).



Fonte: Adaptada de VENABLES (2004).

Legenda: Em todos os casos, exemplos de genes afetados em câncer são indicados. R - receptor. A. Mutações genômicas, causadores de câncer, envolvidas na inclusão (em baixo) ou exclusão (em cima) aberrante de exons. Os exons são representados como retângulos. As seqüências conservadas dos sítios de *splicing* estão mostradas da esquerda para direita: A adenina do *ponto de ramificação*, o trato de polipirimidina, o sítio de *splicing* 3' e o sítio de *splicing* 5'. Os nomes de genes com setas indicam as mutações pontuais encontradas nestes genes. B. Exemplos de inclusão (em cima) e exclusão (em baixo) de exons. Genes com deleções de exons *em fase* estão mostrados no lado esquerdo. Os genes com deleções que geram proteínas truncadas estão mostrados no lado direito. C. Sítios crípticos na extremidade 5' (direita) e 3' (esquerda). D. Retenção de introns no receptor de gastrina e criação de um intron críptico no gene *tsg101 delta* 154-1054. E. *Splicing* alternativo, mutuamente exclusivo – os dois

exons nunca são encontrados simultaneamente no mesmo transcrito. F. A variante *tenascin-C* inclui oito exons variáveis enquanto a variante *mdm2* não possui oito exons comparado com a seqüência completa do gene *mdm2*. G. O gene *cd44* pode, em teoria, sofrer *splicing* de centenas de maneiras diferentes, no entanto, na realidade aproximadamente vinte variantes foram comprovadas. Duas das formas mais comuns estão mostradas aqui.

Figura 4 - Eventos de *splicing* alternativo associados ao câncer, categorizados por subtipos de *splicing*.

Em poucos trabalhos foi demonstrada a existência de uma variante de *splicing* que possui uma expressão "tumor-específica" ou seja, que tem expressão exclusiva em tecido tumoral e não em tecido normal. Portanto decidimos usar o termo "tumor-associada" para variantes que mostram uma expressão mais abundante em tecido tumoral do que em tecido normal e cujo gene independentemente das variantes não seja super expresso em tecido tumoral.

Um possível mecanismo de ação tumorigênica de uma variante de *splicing* associada a tumor poderia funcionar de duas maneiras possíveis: 1. A super-expressão de uma variante com ação tumorigênica pode funcionar semelhante ao ganho de função na transição de um proto-oncogene para um oncogene. 2. A super-expressão de uma variante nova pode também impedir a ação de uma outra variante do mesmo gene. Desta maneira, o efeito da variante associada ao tumor seria como a da perda de função de um gene supressor de tumor.

Porém é igualmente provável que a existência de uma variante no tumor seja conseqüência de diferentes processos ocorrendo no tumor e não tenha nenhuma função causadora do tumor (VENABLES 2004).

Os exemplos mais conhecidos de gene com variantes associadas a diferentes tipos de tumor são os genes *cd44*, um receptor envolvido em interações célula-célula (SNEATH e MANGHAM 1998) e *wt1*, um fator de transcrição (BAUDRY et al.

2000). O gene *cd44* contém 20 exons dos quais os exons 6 até 15 sofrem *splicing* alternativo formando variantes que são preferencialmente expressas em tumores. Foi mostrado que a inibição de algumas destas variantes pode diminuir atividades malignas nas células tumorais (YAKUSHIJIN et al. 1998). O padrão de expressão de outras variantes permite diferenciar entre células metastáticas e não-metastáticas (NAVAGLIA et al. 2003).

O gene *wt1* possui diferentes variantes de *splicing* cuja expressão aumentada em tumor afeta a regulação de proliferação, diferenciação e apoptose em células tumorais (BRINKMAN 2004).

Um estudo recente mostrou que o gene supressor de tumor p53 possui 29 diferentes sítios de *splicing* alternativo em mais de 12 tipos de câncer (HOLMILA et al. 2003). Outros exemplos de genes que apresentam variantes associadas a tumor são *cd79b* (CRAGG et al. 2002), *syk* (WANG et al. 2003a) e *bin1* (GE et al. 1999). Além de ser associada a tumor, a variante do gene *bin1* exibe uma função antagônica à função da variante nativa deste gene, mostrando a diversidade funcional que o *splicing* alternativo pode gerar. Da mesma maneira, os genes de apoptose *Bcl-x, Caspase-9, Ced-4, Caspase-2/Ich-1 e hTid-1* codificam variantes pró-e anti-apoptose (BRINKMAN 2004). Uma vez que uma mutação em uma seqüência reguladora de *splicing* pode gerar variantes de *splicing* espúrias, ela pode também induzir a geração de uma variante que leva à produção de uma proteína truncada. Neste sentido, o *splicing* alternativo pode ter conseqüências semelhantes às mutações *non-sense* (VENABLES 2004). Um exemplo bem conhecido é o gene supressor de tumor *brca1* no qual uma mutação leva à utilização de um sítio críptico que codifica uma proteína truncada em tumores de mama (HOFFMAN et al. 1998).

Estes e outros exemplos mostram o enorme potencial que a detecção e caracterização da expressão de variantes de *splicing* representam para o diagnóstico e tratamento do câncer. O primeiro passo desta caracterização destas variantes inclui: 1 - investigação do padrão de expressão da própria variante, em diferentes tecidos normais e tumorais; 2 - a avaliação do padrão de expressão das outras variantes do mesmo gene. Posteriormente deve ser investigado o efeito que o *splicing* alternativo tem na estrutura, localização e função da proteína codificada pela variante, comparado com as proteínas geradas pelas outras variantes do mesmo gene. Por exemplo, muitos genes que sofrem *splicing* alternativo associado a tumor codificam proteínas que estão presentes na superfície da célula, tornando-as bons alvos terapêuticos (VENABLES 2004).

DING *et al.* (2002) mostraram que o aumento da expressão de uma isoforma alternativa do receptor de secretina, um hormônio regulador da secreção das células ductulares do pâncreas, preveniu a ligação do receptor nativo ao seu ligante (DING et al. 2002). Neste mesmo trabalho, foi demonstrado que a redução da ativação do receptor original devido à competição com o receptor variante permitiu o crescimento tumoral. Neste caso, o bloqueio da isoforma alternativa do receptor, tanto através de anticorpos específicos, quanto eliminando seu respectivo mRNA por RNAi ou antisense, poderia ter um grande valor terapêutico.

Outra possibilidade é a ativação de variantes de *splicing* que induzem apoptose em células tumorais, eliminando assim seletivamente as células cancerosas (VENABLES 2004). Por exemplo, já foi demonstrado que através de oligonucleotídeos antisense que bloqueiam o uso do sítio constitutivo de *splicing* e permite o uso do sítio críptico do gene *bcl-x*, pode-se induzir a produção da variante

pró-apoptótica e levar à morte das células (MERCATANTE et al. 2002).

Uma forma interessante de terapia contra o câncer seria a inserção de uma seqüência que codifica uma droga com ação anti-tumoral dentro de um gene que sofre *splicing* associado a tumor. Desta maneira, o transcrito que codifica a droga será expresso na fase de leitura correta somente quando o gene "hospedeiro" sofrer o padrão de *splicing* associado a tumor (HAYES et al. 2004).

Por fim, não seria vantajoso aumentar a freqüência de *splicing* através do uso de fatores de *splicing* que atuam em *trans,* de forma que uma variante que tenha uma ação antitumor possa ser expressa em níveis elevados, uma vez que esta possibilidade implicaria em uma mudança inespecífica no padrão de *splicing* de todas as variantes de *splicing* influenciadas pelo mesmo fator de *splicing.*

Além do uso de variantes de *splicing* associadas a tumor como alvos terapêuticos, as mesmas podem ser usadas como marcador tumoral. Como mencionado anteriormente, foi demonstrado que é possível fazer a distinção entre câncer de pâncreas metastático e não metastático através da expressão da variante *cd44v10* (NAVAGLIA et al. 2003).

Uma outra possibilidade interessante seria a identificação de exons usados apenas por transcritos presentes em células tumorais que, quando situados na região codificadora, poderiam produzir um peptídeo que servisse como um marcador tumoral ou mesmo um alvo terapêutico como foi sugerido por HAYES et al. (2004). Visto que o *splicing* alternativo pode afetar de maneira significativa a estrutura das proteínas, talvez seja possível a produção de anticorpos específicos para um determinado exon de uma variante associada a tumor e utilizar esse anticorpo para o diagnóstico do câncer. Assim, fica evidente não apenas a necessidade de se estudar a

expressão de diferentes variantes de *splicing*, mas também analisá-las do ponto de vista funcional.

A análise detalhada das variantes associadas a tumor também pode ser crucial para o melhor entendimento do processo de tumorigênese. Em muitos exemplos de *splicing* alternativo associado a tumor, a função da forma alternativa de um gene é consistente com um possível papel no câncer (VENABLES 2004). Algumas das variantes podem estar associadas a processos que são importantes, mas que não necessariamente sejam exclusivos do câncer, como por exemplo, o aumento da proliferação celular.

Uma questão importante que merece ser destacada é se o padrão de *splicing* em câncer é um fator causal para células normais tornarem-se tumorais ou se o processo tumorigênico por si só altera a taxa de replicação celular e, ainda, o modo de síntese de mRNA, de maneira que a maquinaria de *splicing* começa a gerar isoformas "espúrias" de RNA mensageiro (XU et al. 2003). Se for um fator causal, a análise detalhada do processo certamente aumentaria o entendimento do processo de tumorigênese. Neste caso, o tratamento terapêutico impedindo o acontecimento do *splicing* alternativo, podia ser utilizado contra o câncer.

Existem diferentes explicações para um padrão de *splicing* diferencial em câncer: as mutações do tipo subclasse I (mutações nos próprios sítios de *splicing*) e II (nas seqüências reguladoras conservadas), alterações no padrão de expressão dos fatores de *splicing* e por fim mudanças na seqüência do mRNA em conseqüência da edição de RNA (um processo pós-transcricional que modifica nucleotídeos de adenina (A) por inosina (I)) (BRINKMAN 2004). Por exemplo, seqüenciamento de clones de cDNA do gene PTPN6 demonstrou conversões múltiplas de edição A>G,

na maioria das vezes no nucleotídeo A(7866), que representa o ponto de ramificação. Mutagenese demonstrou que esta edição causa *splicing* alternativo associado a tumor (BEGHINI et al. 2000).

Variações nas concentrações intracelulares de fatores de *splicing* podem mudar os padrões de *splicing* alternativo e existem evidências de que este é o mecanismo pelo qual o *splicing* alternativo específico de tecidos ou de diferentes fases de desenvolvimento é regulado (HANAMURA et al. 1998; BRINKMAN 2004).

Existem diferentes estudos analisando a expressão diferencial de alguns fatores de *splicing* em tecido normal e tumoral e seu efeito no padrão de *splicing* de genes específicos. Utilizando diferentes bancos de dados de expressão gênica foi encontrada super-expressão em tumor de diferentes fatores de *splicing* comparando tecido normal e tumoral de cérebro, mama e cólon (KIRSCHBAUM-SLAGER et al. 2004). Já foi demonstrado que em linhagens celulares de fibroblastos normais e transformadas, o padrão de *splicing* do gene de *β-globina* é aumentado nas células transformadas e foi também observada uma mudança no padrão de expressão das proteínas SR (CHABOT et al. 1992; MAEDA e FURUKAWA 2001). A expressão associada a tumor de variantes de *splicing* do gene *cd44* em câncer de mama também está correlacionada com um aumento na expressão de diferentes proteínas SR fosforiladas (STICKELER et al. 1999). Em tumor de pulmão, o padrão de *splicing* alterado do mesmo gene *cd44* foi observado já numa fase primária juntamente com a expressão diferencial do fator ASF/SF2 da família de proteínas SR. Isto mostra que é possível que a regulação diferencial de *splicing* pode ocorrer já no inicio da transformação (ZERBE et al. 2004).

1.3 ESTUDOS COMPUTACIONAIS COMO FERRAMENTA PARA A ANÁLISE DO *TRANSCRIPTOMA*.

Na era pré-genômica costumava-se explorar um determinado gene em todos os seus aspectos utilizando-se apenas recursos experimentais. Se fosse verificado que o mesmo possui variantes de *splicing* alternativo, os sítios de *splicing* e as variantes poderiam então ser investigados. Inúmeros estudos podem ser encontrados sobre o uso de tal abordagem. No momento, com a enorme quantidade de informações genômicas e o acesso a bancos de dados públicos, são comuns análises em massa de dados biológicos. Ferramentas computacionais vêm sendo utilizadas rotineiramente no processamento de dados biológicos, bem como na análise e interpretação dos mesmos. A utilização dessas ferramentas permite a integração de uma quantidade grande de diferentes tipos de informações originárias de fontes diversas. Devido ao grande volume de dados, a bioinformática permite a realização de análises estatísticas sólidas que não eram possíveis de serem geradas utilizando apenas ferramentas experimentais e um reduzido volume de informação.

O uso de seqüências transcritas como ferramenta para o estudo em larga escala de genes tem se mostrado bastante promissor (SOGAYAR et al. 2004). O termo *transcriptoma* vem sendo utilizado para definir o conjunto de transcritos que são expressos em um certo tecido, condição ou organismo. Existem vários métodos experimentais utilizados para decifrar o genoma e seqüenciar em larga escala os transcritos que fazem parte de um *transcriptoma*. Um exemplo importante é a geração de bibliotecas de *ESTs* (*Expressed Sequence Tags*) ou *ORESTES* (*Open Reading Frame Expressed Sequence Tags*). *ESTs* são seqüências parciais derivadas das

extremidades de clones de cDNA provenientes de uma biblioteca construída a partir de um determinado tecido. Desta maneira, em um banco de dados contendo seqüências de diferentes bibliotecas, um mesmo gene pode estar representado por várias *ESTs*. O dbEST, a divisão do GenBank que armazena as *ESTs*, contém hoje 27.646.726 seqüências, sendo 6.100.563 provenientes de tecidos humanos (dados de 24 de junho de 2005).

As *ORESTES* são um tipo específico de *ESTs* que geralmente representam a parte central de um determinado transcrito. As *ORESTES* são produzidas através de uma metodologia baseada em PCR de baixa estringência capaz de representar de forma homogênea transcritos de baixa e de alta abundância transcricional (DIAS NETO et al. 2000). Bibliotecas contendo estas seqüências vêm sendo utilizadas para a caracterização dos diferentes transcritos expressos em uma dada condição a partir de genes conhecidos ou não. Por exemplo, seqüências transcritas de bibliotecas geradas a partir de tecido tumoral podem indicar quais genes e quais variantes de *splicing* estão expressas neste tecido.

Outro tipo de metodologia capaz de gerar informações qualitativas e quantitativas sobre o perfil de expressão de um determinado tecido é o *SAGE* (*Serial Analysis of Gene Expression*) (VELCULESCU et al. 1995) e o *MPSS* (*Massively Parallel Signature Sequencing*) (BRENNER et al. 2000). A análise de diferentes bibliotecas de *SAGE* e *MPSS* pode indicar diferenças significativas entre o padrão de expressão dos genes em uma amostra normal e tumoral. Nestas tecnologias, uma seqüência (*tag*) de 10 nucleotídeos no caso de *SAGE* e 13 nucleotídeos no caso de *MPSS* é extraída e seqüenciada a partir da extremidade 3' dos transcritos presentes em

uma determinada amostra, produzindo bibliotecas contendo aproximadamente 100.000 *tags (SAGE)* / 1.300.000 (*MPSS*). Uma *tag* de *SAGE/MPSS* corresponde à seqüência adjacente ao sítio de restrição da enzima de restrição *NlaIII/DpnII*, respectivamente, mais próximo da extremidade 3' de transcritos que contêm uma cauda poli A. Embora curto, o tamanho das *tags* é, em geral, suficiente para identificar o gene que as gerou. Através da contagem do número de vezes que uma *tag* aparece na biblioteca é possível estimar a expressão de um determinado gene na amostra analisada. Da mesma forma, é possível investigar a expressão diferencial de um gene entre tecidos normais e tumorais. Embora estas tecnologias sejam mais sensíveis e quantitativas do que a geração de bibliotecas de *ESTs*, elas raramente permitem a identificação de diferentes variantes de um mesmo gene uma vez que a maioria das *tags* está localizada na extremidade 3' dos transcritos que em geral é a mesma para todas as variantes.

Existem outras limitações das técnicas de *SAGE* e *MPSS* (BOON et al. 2002; JONGENEEL et al. 2003):

- Quando o sítio de restrição da enzima utilizada está localizado perto da cauda poli A, a *tag* poderá ter muitos nucleotídeos "A", podendo representar diferentes transcritos inespecíficamente.

- Existem genes que não possuem *tags* confiáveis, seja devido à falta de sítios de restrição da enzima *NlaIII/DpnII*, ou porque a *tag* gerada do mesmo pode ser gerada a partir de mais de um gene único.

Além disso, é importante considerar que genes com sítios de poliadenilação alternativos ou com evento de *splicing* alternativo na extremidade 3' do transcrito podem gerar diferentes *tags* o que leva à geração de mais de uma *tag* por gene. Para

analisar o padrão de expressão daquele gene é necessário considerar todas as *tags* de todas as formas possíveis do gene.

Um estudo recente demonstrou que uma outra variável a ser considerada é a existência de SNPs (*Single Nucleotide Polimorphisms*) (Silva et al. 2004). Estas alterações genômicas podem alterar a seqüência de *tags* de *SAGE* por um nucleotídeo ou criar um sítio de restrição da enzima *NlaIII* em um lugar inesperado, desta maneira causando a geração de uma *tag* não esperada.

Por fim, é importante considerar eventos de *internal priming*, que ocorrem quando existe uma seqüência rica em nucleotídeos A na seqüência genômica do gene candidato (BOON et al. 2002). Esta seqüência rica em adeninas seria semelhante a uma cauda poli A e levaria a geração de *tags* que são artefatos experimentais e que não são as *tags* mais próximas da extremidade 3' do transcrito.

Também existem limitações que são comuns à metodologia de *ESTs* (JONGENEEL et al. 2003):

- Possíveis contaminações de cDNA mitocondrial e ribossomal podem gerar *tags* não verdadeiras.

- Erros de seqüenciamento impedem o reconhecimento do gene que originou a *tag* obtida.

- Repetições na seqüência genômica podem causar a geração da mesma *tag* a partir de diferentes genes.

Por fim a metodologia de *microarray* é uma ferramenta adicional para se estudar o *transcriptoma* (BRENTANI et al. 2005). Uma plataforma de *microarray* é uma coleção de fragmentos de DNA imobilizados, distribuídos de uma maneira

organizada e documentada em uma lâmina. Durante a análise de uma determinada amostra, a intensidade de cada fragmento correlaciona com a abundância da mRNA complementar correspondente. Plataformas de *microarray* têm sido utilizadas para investigar a expressão de genes em geral, sem diferenciação de variantes de *splicing*. Embora esta metodologia seja eficiente para a investigação do padrão de expressão de variantes de *splicing* ela não é adequada para a descoberta de novas variantes de *splicing* (LEE e ROY 2004).

1.3.1 A utilização de bancos de dados de *ESTs (Expressed Sequence Tags)* para o estudo de *splicing* alternativo

A falta de conhecimento sobre os mecanismos básicos de controle de *splicing* alternativo não permite o desenvolvimento de ferramentas computacionais que sejam capazes de predizer o padrão de *splicing* para um determinado gene exclusivamente a partir da seqüência genômica. Assim, a análise de seqüências transcritas vem sendo utilizada, dentre outras finalidades, na caracterização da variabilidade do *transcriptoma* humano gerada por *splicing* alternativo (MIRONOV et al. 1999; CROFT et al. 2000). As análises podem ser feitas a partir da avaliação de grupos de seqüências de cDNA que são provenientes de um mesmo gene, buscando diferenças entre as mesmas que sejam condizentes com *splicing* alternativo, tais como deleções ou inserções de blocos de seqüências correspondentes a exons.

O uso em conjunto de seqüências expressas e da seqüência genômica representa um diferencial importante que vem sendo explorado por vários grupos de pesquisa (MIRONOV et al. 1999; CROFT et al. 2000; HIDE et al. 2001). Vários estudos têm analisado a taxa de *splicing* alternativo no genoma humano e os

resultados sugerem que 30% - 74% dos genes humanos tenham pelo menos duas variantes de *splicing* (MIRONOV et al. 1999; BRETT et al. 2000; CROFT et al. 2000; MODREK et al. 2001; KAN et al. 2002; JOHNSON et al. 2003).

MIRONOV et al. (1999), por exemplo, encontraram variantes de *splicing* em 133 de 192 genes humanos após o alinhamento das respectivas *ESTs* com o genoma humano (MIRONOV et al. 1999). MODREK et al. (1999), encontraram 6.201 variantes de *splicing* para 2.272 genes através do mapeamento de 2,1 milhões de seqüências expressas (*ESTs* e mRNAs) no genoma, (MODREK et al. 2001). Cabe ressaltar que estes grupos não fizeram o mapeamento das *ESTs* propriamente ditas, mas sim de seqüências consenso, derivadas a partir de montagens de *ESTs* provenientes de um mesmo gene. Nestes estudos foram utilizados os bancos de dados do "TIGR Human Gene Index" (QUACKENBUSH et al. 2001) e do *"UniGene"* (SCHULER et al. 1996), respectivamente.

As grandes limitações dessas metodologias dizem respeito aos artefatos gerados durante o processo de montagem das seqüências. Os artefatos ocorrem principalmente devido a: i) baixa qualidade das seqüências, que dificulta a determinação precisa de sobreposições, ii) presença de formas alternativas de *splicing*, que pode dificultar o agrupamento de seqüências derivadas de um mesmo gene, iii) existência de famílias gênicas com grande similaridade em nível de nucleotídeos e cujas seqüências podem ser artificialmente agrupadas em um mesmo consenso. Outro problema detectado nos bancos de dados de seqüências expressas é a presença de uma significativa porção de seqüências derivadas de contaminação das bibliotecas de cDNA com DNA genômico

Em nosso laboratório foi desenvolvida uma abordagem de estudo do *transcriptoma* na qual as seqüências de cDNAs são diretamente mapeadas no genoma humano sem a necessidade de produção de seqüências consenso, evitando assim artefatos de montagem destas (SAKABE et al. 2003; GALANTE et al. 2004).

1.3.2 Estudos em larga escala sobre *splicing* alternativo associado a tumor

Além dos inúmeros estudos sobre genes individuais que têm uma variante que é diferencialmente expressa em tecidos tumorais, foram publicados diferentes estudos nos quais a associação entre o *splicing* alternativo e o câncer foi investigada em larga escala. Diferentes abordagens foram utilizadas para investigar o *splicing* alternativo no *transcriptoma*, e cada um dos estudos determinou filtros diferentes para buscar variantes associadas a tumor.

XIE et al. (2002), alinharam seqüências de *ESTs* e seqüências de mRNAs contra o genoma para encontrar variantes de *splicing*. Desta maneira acharam 20.301 *clusters* (agrupamentos de seqüências de cDNA) dos quais 8.254 possuíam pelo menos duas variantes. Em seguida, foram analisados os números de *ESTs* e mRNAs e suas origens teciduais para selecionar variantes específicas de um determinado tecido. Uma variante foi definida como específica a tumor quando possuía pelo menos duas *ESTs* oriundas de diferentes bibliotecas ou um mRNA de um mesmo tecido tumoral representando-a. Este estudo não apresentou o número final de candidatos específicos a tumor e não apresentou validações experimentais dos seus candidatos. (XIE et al. 2002).

Num segundo estudo, WANG et al. (2003b) alinharam apenas seqüências de *ESTs* contra seqüências de mRNAs para encontrar variantes de *splicing*. A partir de

um total de 26.258 variantes de *splicing* foram encontradas 845 variantes associadas a tumor. Setenta e seis (76) variantes foram validadas experimentalmente em amostras de tumor de pacientes, sendo que 55 apresentaram um resultado conclusivo e destas, 45 foram validadas como sendo realmente associadas a tumor.

Em um terceiro estudo, XU et al. (2003) geraram dados muito interessantes analisando variantes de *splicing* a partir de alinhamento de seqüências de cDNA com o genoma. Eles calcularam valores estatísticos para determinar quais variantes estavam realmente associadas a tumor e demonstraram através de informação sobre os genes candidatos disponível na literatura que a maioria das 89 genes selecionados com variantes associadas a tumores existe realmente. O estudo demonstrou que 78% das variantes associadas a tumor foram confirmadas por uma seqüência *full insert* (seqüências geradas a partir do seqüenciamento completo de clones de cDNA). Este fato levou a conclusão que a maioria das variantes não era espúria. Além disso, demonstraram que mais seqüências *full insert* dos genes candidatos foram gerados originalmente de tecidos tumorais quando comparado com um conjunto de genes sofrendo *splicing* alternativo não associado a tumor. Por fim, a investigação das categorias funcionais dos genes demonstrou que 52% das variantes associadas a tumor provinham de genes envolvidos na tumorigênese ao passo que apenas 20% dos genes de um conjunto controle de genes provinha de genes envolvidos na tumorigênese. Por exemplo, os genes supressores de tumor estavam representados 7 vezes mais no grupo dos genes com variantes associadas a tumor do que em um conjunto controle de outros genes. Os autores sugeriram ainda que a expressão da variante de *splicing* associada ao tumor tem um importante papel na formação deste. No entanto, apesar da utilização de diferentes abordagens computacionais e

estatísticas, nenhum candidato foi validado experimentalmente.

Por fim, HUI et al. (2004) geraram uma análise semelhante na qual alinharam primeiro todas as seqüências *full insert* com o genoma e depois todas as seqüências parciais de cDNA *(ESTs* e *ORESTES)* contra as seqüências *full insert*. Levando-se em consideração apenas bibliotecas não normalizadas, foi feita uma análise estatística que revelou a existência de 2.149 (8%) variantes associadas a tumor de um total de 26.812 variantes. Nove variantes candidatas foram validadas experimentalmente, dos quais oito demonstraram o resultado esperado. Além disso, 25 genes candidatos foram comparados com resultados de experimentos de *microarray* de um outro grupo. De 13 genes (52%) os padrões de expressão preditos se mostraram consistentes com os resultados de *microarray* (HUI et al. 2004).

Estes quatro artigos foram publicados durante o período do meu doutorado. Desta maneira, aproveitamos a oportunidade de incorporar os aspectos interessantes de cada um dos trabalhos em nosso estudo. No entanto vale ressaltar alguns aspectos exclusivos da a nossa análise e que não foram diretamente abordados nos estudos mencionados anteriormente. Por exemplo, é importante notar que apesar de não terem sido identificadas nesses trabalhos variantes de *splicing* que possuíssem realmente uma expressão "específica de tumor", o termo específico de tumor foi mantido ao passo que em nossa análise preferimos utilizar o termo associado a tumor para descrever esse tipo de expressão. Além disso, nos trabalhos discutidos não está claro se as análises foram feitas de forma compartimentalizada por tipo de tecido ou se as variantes identificadas estão associadas a todos os tipos de tumor de uma forma geral. Também vale ressaltar que nenhum dos estudos citados verificou experimentalmente

a expressão simultânea da variante tumor específica e a do gene independentemente das variantes em tecido normal. A análise simultânea permite diferenciar os genes para os quais todas as variantes estão com expressão alterada em tumores daqueles para os quais apenas uma variante específica possui expressão alterada em tumor e pode realmente ser denominada uma variante associada a tumor.

Quando um gene demonstra super-expressão em tumor espera-se que suas variantes também demonstrem uma super-expressão em tumor. No entanto, quando um gene não é super-expresso em tumor e a maquinaria de *splicing* gera variantes super-expressas, existe a possibilidade de uma regulação de *splicing* alterada nesta condição. Como a análise detalhada destas variantes resultando de *splicing* alternativo diferencialmente regulado em tumor pode levar ao melhor entendimento do processo de tumorigênese, decidimos neste trabalho buscar variantes associadas a tumor que tem o potencial de serem geradas devido à regulação de *splicing* associada à doença.

OBJETIVOS

2 OBJETIVOS

2.1 OBJETIVO PRINCIPAL

- Identificação de variantes de *splicing* associadas a tumores através da construção de um sistema computacional para a análise em larga escala de um banco de dados de genes que sofrem *splicing* alternativo e a validação experimental das mesmas.

2.2 OBJETIVOS SECUNDÁRIOS

- Desenvolvimento de um sistema de priorização de exons que provêm de variantes de splicing associadas a tumor.
- Validação experimental dos exons candidatos em linhagens celulares para o melhoramento do sistema de priorização.
- Validação experimental dos exons candidatos finais em amostras pareadas normal/tumor de pacientes com câncer.

MATERIAL E MÉTODOS

3 MATERIAL E MÉTODOS

3.1 BANCO DE DADOS DO *TRANSCRIPTOMA* HUMANO

O banco de dados do *Transcriptoma* Humano contém informações sobre as regiões transcritas do genoma humano. Tais regiões foram definidas *in silico* através do alinhamento entre seqüências de cDNA (ESTs e clones de cDNA completamente seqüenciados) e seqüências genômicas. Para a construção do nosso banco de dados, foram utilizados os seguintes bancos de dados: *UniGene* versão *153*, *dbEST* versão Julho de 2002 e seqüência genômica humano versão *29* obtida do Centro Nacional para Informação de Biotecnologia (*NCBI*). A montagem do genoma humano mascarada para seqüências repetitivas pelo programa *RepeatMasker* (http://www.repeatmasker.org/), foi utilizada como fonte da informação genômica. As seqüências de cDNA humano foram obtidas das divisões *dbEST* e *nr* do *GenBank*, e também foram filtradas como a seqüência genômica. Para o mapeamento das seqüências de cDNA na seqüência genômica foi utilizado um programa implementado de MEGABLAST (ALTSCHUL et al. 1997) chamado de pp-Blast (OSORIO et al. 2003). Em total foram utilizados 3422614 *ESTs* (incluindo *ORESTES*) e 52903 *full inserts* para a construção do banco.

Depois do alinhamento de todas as seqüências os resultados de MEGABLAST foram filtrados (ver em seguida) e armazenados em um banco de dados (MySQL). Este banco contém informação das características das seqüências alinhadas incluindo o local de alinhamento, a orientação das seqüências alinhadas e a

origem tecidual (SAKABE et al. 2003; GALANTE et al. 2004). Para excluir artefatos de alinhamento do nosso do banco de dados foram definidos critérios: os alinhamentos das seqüências de cDNA com o genoma precisavam exibir um grau de identidade de pelo menos 93% ao longo de pelo menos 45% do tamanho total de uma *EST* ou ao longo de pelo menos 55% do tamanho total de uma seqüência *full insert*. Para seqüências que alinharam em mais de um lugar no genoma apenas o melhor alinhamento foi mantido. Para escolher o melhor alinhamento foi calculado um valor resultante da multiplicação da identidade do alinhamento pelo tamanho do alinhamento.

Para agrupar as seqüências parciais e completas de cDNA correspondentes a um mesmo gene foram utilizadas as coordenadas de mapeamento dessas seqüências no genoma humano (SAKABE et al. 2003) (ver Figura 5 para um exemplo). Foram agrupadas em um mesmo *cluster* todas as seqüências que compartilhavam pelo menos um limite exon/intron ± 5 bp. Quando um *cluster* não continha nenhum limite exon/intron, duas seqüências de cDNA precisavam ter pelo menos uma sobreposição de 100 pares de bases para pertencer ao mesmo *cluster*. Este banco possui 318.272 *clusters*, dos quais 21.306 possuem pelo menos uma seqüência completa de mRNA.

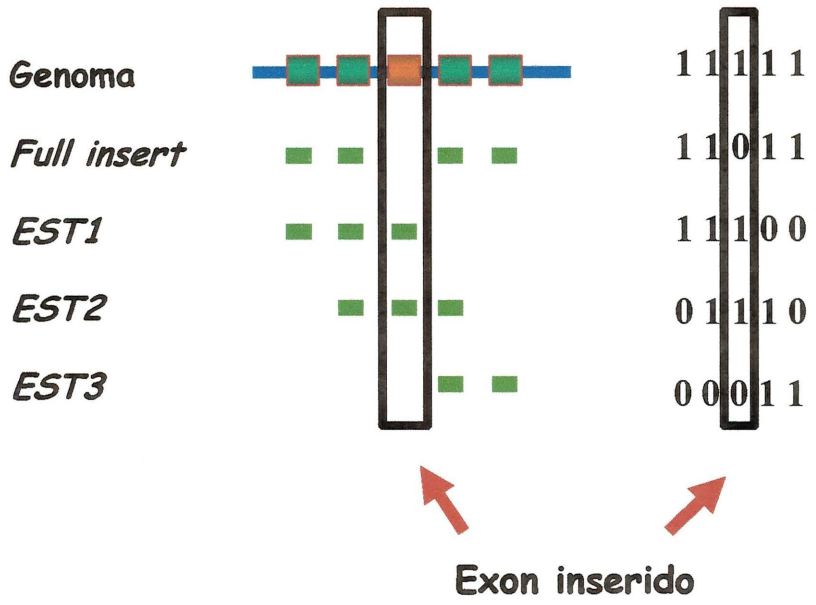
Essa estratégia de agrupamento evita parcialmente os artefatos gerados através do alinhamento direto de seqüências transcritas, geralmente associados à baixa qualidade das seqüências e à presença de extensas famílias gênicas no genoma. O alinhamento das seqüências expressas contra a seqüência genômica fornece ainda uma maneira indireta de eliminar o problema relacionado à presença de contaminação das seqüências expressas com DNA genômico. Isso porque o alinhamento entre seqüências expressas e seqüências genômicas geralmente

apresentam interrupções devido à presença de introns na seqüência genômica os quais são removidos das seqüências expressas após a ocorrência do *splicing*. Assim sendo, a presença de interrupções nos alinhamentos podem ser utilizadas para indicar se uma determinada seqüência expressa apresenta ou não *splicing*.

Após o agrupamento das seqüências e a definição dos exons através do alinhamento com a seqüência genômica, é possível identificar eventos de *splicing* alternativo comparando as coordenadas dos exons de cada uma das seqüências que fazem parte de um mesmo *cluster*. Para facilitar e automatizar essa comparação foi desenvolvido um sistema de representação dos exons de uma determinada seqüência em forma de matrizes binárias (ver seção 3.2).

3.2 CONSTRUÇÃO DE MATRIZES BINÁRIAS

Uma maneira simples de representar o uso alternativo de exons para todas as seqüências de um mesmo *cluster* de cDNA é uma matriz binária. Nesta matriz cada exon é representado por uma coluna e cada seqüência por uma linha (Figura 5). Se um determinado exon está representado em uma dada seqüência, a respectiva célula na matriz recebe 1 (um), caso contrário ela recebe 0 (zero). Como todos os exons foram representados nestas matrizes binárias, o uso alternativo dos mesmos pode ser facilmente verificado com um rastreamento das linhas da matriz buscando um padrão tipo $10_{(n)}1$. Assim, um banco de dados (MySQL) foi montado contendo toda a informação de variantes de *splicing*, mapeamento no genoma, expressão tecidual e anotação de cada *cluster* de cDNA.



Legenda: Os retângulos verdes representam exons presentes nas seqüências de cDNA que formam o *cluster*. Se um determinado exon está representado em uma dada seqüência, a respectiva célula na matriz recebe 1 (um), caso contrário ela recebe 0 (zero). O retângulo vermelho representa um exon inserido.

Figura 5 - Representação gráfica da matriz de *splicing*.

As vantagens do uso destas matrizes são: a) a visualização simplificada; b) os dados são facilmente extraídos das mesmas, o que facilita estudos em larga escala; c) elas são úteis para a análise de *clusters*, no intuito de se identificar todas as formas únicas de *splicing*; d) uma vez que as matrizes foram geradas a partir do mapeamento de *ESTs* no genoma, isso garante que artefatos decorrentes da montagem de *EST*s seguida do mapeamento dos consensos dos mesmos no genoma sejam evitados. A desvantagem deste método é que ele não permite a identificação em larga escala de outras formas de *splicing* alternativo, tais como o uso de sítios crípticos de *splicing* e o uso de introns como seqüências exônicas. Uma matriz binária foi construída para cada *cluster* gerado pelo mapeamento descrito acima.

Todas as análises computacionais foram geradas utilizando programas de computação escritos na linhagem de programação *Perl*.

3.3 VALIDAÇÃO DO BANCO DE DADOS DE *SPLICING*

O banco de dados de *splicing* alternativo foi validado manualmente através da procura de genes já descritos na literatura por possuírem variantes de *splicing*. Para tanto foi utilizado o banco de dados de artigos científicos publicados PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) com as palavras chave *alternative splicing*. A lista de genes com *splicing* selecionados com base na literatura para a validação do nosso banco de dados e as respectivas referências bibliográficas estão apresentadas na seção 4.4.2.

3.4 ANÁLISE ESTATÍSTICA: CÁLCULO DE VALOR Z

Para validar a significância estatística da associação a tumor de cada candidato, foi calculado um valor Z monocaudal para cada exon candidato considerando-se o número de *ESTs* que indicava a associação a tumor e o número total de *ESTs* existentes em todas as bibliotecas de um dado órgão (WANG et al. 2003b). Desta maneira, o valor Z fornece uma indicação da probabilidade da associação tumoral do exon em um órgão específico:

$$Z = (p_t - p_n)/\sqrt{p(1-p)(1/n_n+1/n_t)}.$$

Para um dado exon, p_t e p_n são as freqüências de expressão do exon em tecido tumoral e normal de um determinado órgão (o número de *ESTs* tumorais/normais

contendo o exon específico dividido pelo número de *ESTs* tumorais/normais de todas as bibliotecas). "p" é a média geométrica da freqüência do exon em bibliotecas normais e tumorais e n_n e n_t são os números de *ESTs* nas bibliotecas normais e tumorais que foram considerados para cada exon estudado.

Para minimizar amostragens tendenciosas, consideramos somente bibliotecas contêm um número de seqüências igual ou maior do que a biblioteca na qual foi identificada a variante associada a tumor em um órgão específico.

Valores Z tendo um $P \leq 0.05$ foram considerados significantes. Foram selecionados exons que demonstraram uma super-expressão em pelo menos um tecido de origem tumoral.

3.5 GERAÇÃO DE *SAGE TAGS* VIRTUAIS

Uma vez que queríamos excluir as variantes candidatas pertencendo a genes super expressos em tumor, investigamos a expressão dos nossos genes candidatos em tumor. Comparamos a freqüência de aparência da *tag* correspondendo ao gene candidato em bibliotecas normais com a freqüência da mesma em bibliotecas tumorais.

Uma "*SAGE tag*" virtual é uma predição da seqüência de 10 pb que poderia ser gerada a partir de um determinado gene pelo experimento de *SAGE* (BOON et al. 2002). Para cada gene candidato, foi selecionada uma seqüência completa de cDNA (*full inserts*) do nosso banco de dados. Uma vez que na geração de uma biblioteca de *SAGE* usam-se seqüências poli T para a isolação dos transcritos de mRNA, geramos *tags* apenas das seqüências completas que continham uma cauda poli A. Para cada

seqüência completa representativa de cada gene foi gerada uma *SAGE tag* virtual determinada pelo sítio de restrição da enzima *NlaII* localizado mais próximo da extremidade 3' do transcrito.

Foram excluídas as *tags* localizadas em regiões repetitivas. Estas *tags* podem não representar apenas um gene, uma vez que as seqüências repetitivas genômicas aparecem em diferentes regiões do genoma e, portanto a *tag* pode não representar apenas o gene a ser analisado. A especificidade das *tags* também foi avaliada verificando-se se uma determinada *tag* poderia representar mais do que um único gene e estas *tags* inespecíficas foram eliminadas.

A freqüência de cada *tag* foi contada nas bibliotecas tumorais e normais do mesmo tecido. Como mencionado anteriormente, foi calculado um valor Z monocaudal indicando a probabilidade de super-expressão em tecido tumoral do gene num órgão específico:

$$Z = (p_t - p_n)/\sqrt{p(1-p)(1/n_n+1/n_t)}.$$

Para um dado gene, p_t e p_n são as freqüências da *tag* observada nas bibliotecas normais e tumorais, respectivamente, do mesmo tecido (= o número de *tags* nas bibliotecas normais ou tumorais / pelo número total de *tags* normais ou tumorais do mesmo tecido). "p" é a média geométrica da freqüência da *tag* em bibliotecas normais e tumorais e n_n e n_t são os números de *tags* nas bibliotecas normais e tumorais que foram considerados para cada gene estudado.

Para minimizar amostragens tendenciosas, consideramos somente bibliotecas contêm um número de *tags* igual ou maior do que a biblioteca na qual foi identificado o gene candidato em um órgão específico. Valores Z correspondendo a um $P \leq 0.05$ foram considerados significantes e portanto, candidatos determinados

como significantes estatisticamente ainda têm uma probabilidade de serem candidatos falso-positivos de ≤ 0.05. Foram excluídos todos os genes que demonstraram uma super - expressão em tecido tumoral.

3.6 AVALIAÇÃO EXPERIMENTAL DO PADRÃO DE EXPRESSÃO DE VARIANTES DE *SPLICING*

O padrão de expressão dos genes, bem como de suas possíveis variantes de *splicing*, foi avaliado por *RT-PCR*. Inicialmente, o padrão de expressão foi avaliado em tecidos normais, utilizando RNA adquirido comercialmente (*Clontech®*). Em seguida, foi utilizado RNA de linhagens celulares tumorais. Cada exon foi testado em linhagens tumorais derivadas de tecidos na qual a associação com tumor foi observada. Ao final da validação, os produtos de PCR foram clonados em plasmídios e seqüenciados para confirmar a especificidade da amplificação obtida. Por fim, os candidatos foram validados em amostras de pacientes. Além disso, quando possível, validamos os candidatos em amostras normais pareadas com as respectivas amostras tumorais.

3.6.1 RNA de amostras de pacientes

Amostras de diferentes tipos de tumor foram solicitadas junto ao Banco de Tumores do Hospital A. C. Camargo. Foram utilizadas apenas amostras cedidas ao Banco de Tumores após o consentimento informado dos pacientes e todas as precauções pertinentes para manter o sigilo e a confidencialidade dos dados dos pacientes foram adotadas. Além destes tumores, foram utilizadas também 11 amostras de RNA de glioblastoma gentilmente cedidas pelo Dr. Gregory Riggins da Universidade John's Hopkins Baltimore – Estados Unidos da América.

A extração de RNA dos tumores foi feita com o reagente Trizol (Invitrogen®). Para tanto, os tumores ainda congelados foram imersos em 1 ml do reagente Trizol (sobre uma placa de Petri estéril) e cortados em pequenos fragmentos com o auxílio de um bisturi estéril. Posteriormente, estes fragmentos foram transferidos para tubos cônicos de poliestireno de 5 ml (*Falcon*®) aos quais foi acrescido 1 ml do reagente Trizol. Com auxílio de um *Polytron* (*Kinematica® AG*), as amostras foram completamente homogeneizadas e o protocolo foi seguido conforme as especificações do fabricante.

3.6.2 Linhagens Celulares Tumorais

Linhagens celulares tumorais, disponibilizadas pela ATCC® (*American Type Culture Collection*), foram utilizadas para avaliar o padrão de expressão das variantes de *splicing*. As linhagens tumorais utilizadas foram: A172 (glioblastoma), T98G (glioblastoma multiforme), H358 (adenocarcinoma de pulmão), H1155 (adenocarcinoma de pulmão), DU145 (carcinoma de próstata), PC3 (adenocarcinoma de próstata), MDA-436 (adenocarcinoma de mama), MCF-7 (adenocarcinoma de

mama) e SW-480 (adenocarcinoma coloretal). Além destas linhagens, foram utilizadas também duas linhagens de astrócitos cedidas pelo Dr. Gregory Riggins da Universidade John's Hopkins, Baltimore – Estados Unidos da América: uma linhagem de astrócito normal (Vnox); e THRNox, uma linhagem de astrócitos com mutações nos genes *hTERT* e *Ras*. Cada linhagem foi cultivada segundo as especificações do fornecedor. Neste laboratório também foi utilizado RNA humano de referência universal (Universal Human Reference RNA) (Stratagene no: 740000). Este RNA contém uma mistura de RNA de linhagens tumorais de adenocarcinoma de mama, hepatoblastoma de fígado, adenocarcinoma da cérvix uterina, carcinoma embrionário de testículo, glioblastoma de cérebro, melanoma, lipossarcoma, linfoma histiocítico de macrófagos, leucemia linfoblástica de linfoblastos T e plasmocitoma de linfócitos B.

As linhagens celulares foram cultivadas em meio apropriado até obtenção de confluência (aproximadamente $4x10^4$ células/cm^2) e submetidas à extração de RNA pelo método de sedimentação em Cloreto de Césio (CHIRGWIN et al. 1979). Inicialmente, o meio de cultura foi aspirado e 9 ml da solução de lise (4 M Isotiocianato de Guanidina, 25 mM Citrato de Sódio – pH 7.0, 0.1 M β-mercaptoetanol) foram adicionados à garrafa de cultura (75 cm^2). Em seguida, o lisado celular foi homogeneizado e transferido para um tubo de ultracentrífuga contendo 4 ml de solução de Cloreto de Césio (5.7 M CsCl e 25 mM NaAc) e então centrifugado a 150000 xg por 17 horas a 20° C (rotor SW40Ti, *Beckman*®). Após a centrifugação, formou-se um precipitado de RNA e o sobrenadante contendo proteínas e DNA foi descartado. Finalmente, a parede interna do tubo foi limpa e o RNA solubilizado em 50 a 200 µl de água DEPC. A dosagem de RNA foi feita em

espectrofotômetro apropriado a 260 nm de comprimento de onda. Além disso, também foi avaliada a razão entre as leituras a 260 e 280 nm, a qual indica a pureza do material obtido.

3.6.3 Avaliação da Qualidade dos RNAs Extraídos

A qualidade dos RNAs foi avaliada em relação à degradação do material e à contaminação com DNA genômico. A integridade dos RNAs foi visualizada aplicando-se 1 μg de RNA total em gel de 1% agarose. Antes de ser aplicado, o RNA foi desnaturado a 65°C por 5 minutos, sendo mantido em condição desnaturante em tampão de amostra contendo uréia (2X TAE, 30% Glicerol, 7 M Uréia, traços de Azul de Bromofenol). A coloração do material foi feita com brometo de etídio e o gel foi visualizado em luz UV. Foram considerados íntegros os *RNA*s que apresentaram as bandas correspondentes aos *RNA*s ribossômicos 28S e 18S bem evidentes e na razão 2:1.

A contaminação com DNA genômico foi verificada através do teste de *hMLH1* (*human mut-L homologue 1*). Este teste consiste em uma *PCR*, na qual utilizam-se 200 ng de *RNA* total e *primers* desenhados nos introns 12 e 13 do gene *hMLH1*, de maneira que, qualquer amplificação obtida deve-se à contaminação com DNA (tamanho esperado do fragmento: 250 pb). Os *primers* utilizados foram: **HMLH-F**: 5' TGG TGT CTC TAG TTC TGG 3' e **HMLH-R**: 5' CAT TGT TGT AGT AGC TCT GC 3'.

As seguintes condições de amplificação foram utilizadas: 1,5 mM $MgCl_2$, 0,1 mM dNTPs, 0,4 μM de cada *primer*, 1 U de *Taq DNA polimerase* (*Invitrogen®*), em tampão apropriado. A reação foi iniciada com desnaturação a 94°C por 5 minutos,

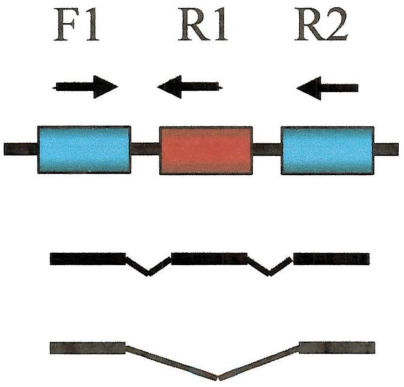
seguida de 35 ciclos de: 45 segundos a 94°C, 45 segundos a 55°C e 1 minuto a 72°C, e extensão final de 6 minutos a 72°C. Os produtos de amplificação foram visualizados em gel de poliacrilamida 8% corado em prata. As amostras que apresentaram alguma amplificação e portanto estavam contaminadas com DNA genômico, foram tratadas com *DNAse* (*Invitrogen*®), segundo instruções do fabricante. Para confirmar se a contaminação havia sido eliminada, o teste de *hMLH1* foi repetido.

3.6.4 RNAs de tecidos normais

Devido à dificuldade de obtenção de tecidos normais de cérebro, pulmão, cólon e próstata os mesmos foram comprados (*Clontech*® nº: K4005-Z). Os RNAs comprados também tiveram sua qualidade avaliada conforme descrito no item anterior.

3.6.5 Construção dos *primers*

Os *primers* foram construídos de maneira que pudéssemos avaliar o padrão de expressão das possíveis formas de *splicing* concomitantemente. Os *primers* foram construídos em exons vizinhos a exons candidatos a serem associados a tumor (ver Figura 6). Utilizando os *primers* F1 e R2 podíamos investigar o padrão de expressão das duas bandas distintas em caso de expressão das duas formas de *splicing*. No entanto, utilizando os *primers* F1 e R1 podíamos analisar apenas a expressão da variante tumor associada. O programa '*Oligotech*' foi utilizado para a construção dos *primers*. Através desse programa é possível determinar a temperatura de *annealing* assim como verificar a ocorrência de estruturas secundárias e de dímeros.



Legenda: As setas indicam os *primers* utilizados na amplificação dos fragmentos. Estão representados os fragmentos possíveis de serem obtidos. Os retângulos azuis representam exons flanqueadores. O retângulo vermelho é o exon candidato a ser associado a tumor.

Figura 6 – Representação gráfica das regiões escolhidas para a construção dos *primers* utilizados nas *RT-PCRs* dos exons candidatos

3.6.6 *RT-PCR (Reverse Trancriptase-Polimerase Chain Reaction)*

Foram utilizados 2 μg de RNA total de um determinado tecido e ainda: 1 μL de oligo dT (0,5 μg/μL), 1 μL de dNTPs (10 mM), em volume final de reação de 12 μL. Este foi incubado a 65° C durante 5 minutos e armazenado em gelo. Em seguida, foram adicionados 4 μL de 5X First strand buffer, 2 μL de DTT 0,1 M, 200 U de RNAse out e 40 U de *SuperScript II* (*Gibco*®). A reação seguiu-se com incubação a 42° C por 1 hora e depois a 70° C por 15 minutos.

Foi feito um teste para avaliar a qualidade do cDNA (DNA complementar). Para tanto, foi feita uma PCR utilizando-se *primers* específicos (*Forward*: 5' CTG CAC CAC CAA CTG CTT A 3' e *Reverse*: 5' CAT GAC GGC AGG TCA GGT C 3') para os exons 6 e 7 do gene *GAPDH* (gliceraldeído desidrogenase). Esta reação foi feita nas seguintes condições: 0,5 μl de cDNA, 1,0 mM $MgCl_2$, 0,1 mM dNTPs, 0,4 μM de cada *primer* e 1 U de Taq *DNA* polimerase (*Invitrogen*®), em tampão

apropriado e volume final de 20 µl. A amplificação foi iniciada com desnaturação a 94°C por 5 minutos, seguida de 22 ciclos de: 1 minuto a 94°C, 45 segundos a 60°C e 1 minuto a 72°C, e extensão final de 10 minutos a 72°C. Este gene é altamente expresso em todos os tecidos, de maneira que em uma boa síntese de cDNA observamos uma forte amplificação (tamanho do fragmento esperado: 300 pb).

O protocolo inicial para cada candidato foi feito utilizando-se 1 µL de cDNA, 1,4 mM $MgCl_2$, 0,1 mM dNTPs, 4,0 µM de cada *primer* e 1 U de Taq DNA polimerase (*Invitrogen*®), em tampão apropriado. A reação iniciou-se com desnaturação a 94°C por 5 minutos, seguida de 35 ciclos de: 45 segundos a 94°C, 45 segundos a 60°C e 1 minuto a 72°C, e extensão final de 10 minutos a 72° C. O produto foi visualizado em gel de poliacrilamida 8% corado com prata 0,2%.

3.6.7 Clonagem e seqüenciamento dos produtos de PCR

Após a amplificação do fragmento desejado, o mesmo foi clonado em vetor plasmidial (PCR^{21}), utilizando-se o TA *cloning kit* (*Invitrogen*®), seguindo-se as instruções do fabricante. Cinco colônias de cada candidato foram selecionadas para um *screening*, feito por PCR direto da colônia. A reação foi feita com *primers forward* (5' CGC CAG GGT TTT CCC AGT CAC GAC 3') e *reverse* (5' TCA CAC AGG AAA CAG CTA TGA C 3') de PUC, que flanqueiam a região onde o fragmento foi clonado. As condições foram as mesmas da *RT-PCR*, mas com 29 ciclos ao invés de 35, para a obtenção de um produto mais limpo. As colônias positivas foram submetidas ao seqüenciamento, feito a partir deste produto em seqüenciador automático *ABI 3100* (*Applied Biosystems*®), segundo especificações do fabricante.

RESULTADOS

4 RESULTADOS

A seguir estão apresentados os resultados que compõem o artigo já publicado e alguns resultados complementares. A composição de resultados publicados (item 4.2), refere-se aos principais dados obtidos durante o projeto de Doutorado. Este trabalho mostra o desenvolvimento de um sistema de busca de variantes de *splicing*. Posteriormente serão apresentados alguns resultados não publicados de uma ampliação da validação experimental de uma variante de *splicing* tumor associada que foi encontrada utilizando este mesmo sistema (item 4.3).

4.1 IDENTIFICAÇÃO DE VARIANTES DE *SPLICING* ASSOCIADAS A TUMORES

Como mencionado anteriormente o objetivo do presente trabalho é a identificação de variantes de *splicing* associadas a tumores através da construção de um sistema computacional e sua validação experimental. A integração das análises computacionais com as validações experimentais levou ao desenvolvimento de um sistema que foi ajustado gradualmente. Aqui serão descritos os critérios que foram implementados no sistema de seleção.

Na busca de variantes de *splicing* associadas a tumor decidimos nos focar nos eventos de *splicing* do tipo inserções de exons, já que a presença de uma seqüência

adicional (o exon inserido) em tumores permitirá o desenvolvimento de possíveis estratégias terapêuticas ou diagnósticas.

A utilização de matrizes binárias, descrita na seção 3.2 permite a representação individual do uso alternativo de todos os exons de todos os genes. Portanto, investigamos a associação a tumor para cada exon individualmente. Como conseqüência dessa abordagem existe a possibilidade de dois exons candidatos a serem associados a tumor pertencerem a uma mesma variante. Por outro lado isto também implica que um determinado exon candidato pode pertencer a mais de uma variante de *splicing* do mesmo gene.

Como mencionado anteriormente, o nosso banco do *transcriptoma* agrupa seqüências transcritas correspondentes a um mesmo gene. Cada agrupamento corresponde a um *cluster*. A partir de um levantamento do tecido de origem de cada seqüência transcrita é possível inferir o padrão de expressão de cada gene. Desta maneira, é possível procurar variantes e exons associados a um tecido ou a tumor. Entretanto, é necessário considerar que a informação tecidual disponível para cada biblioteca de cDNA nem sempre é completa e/ou correta.

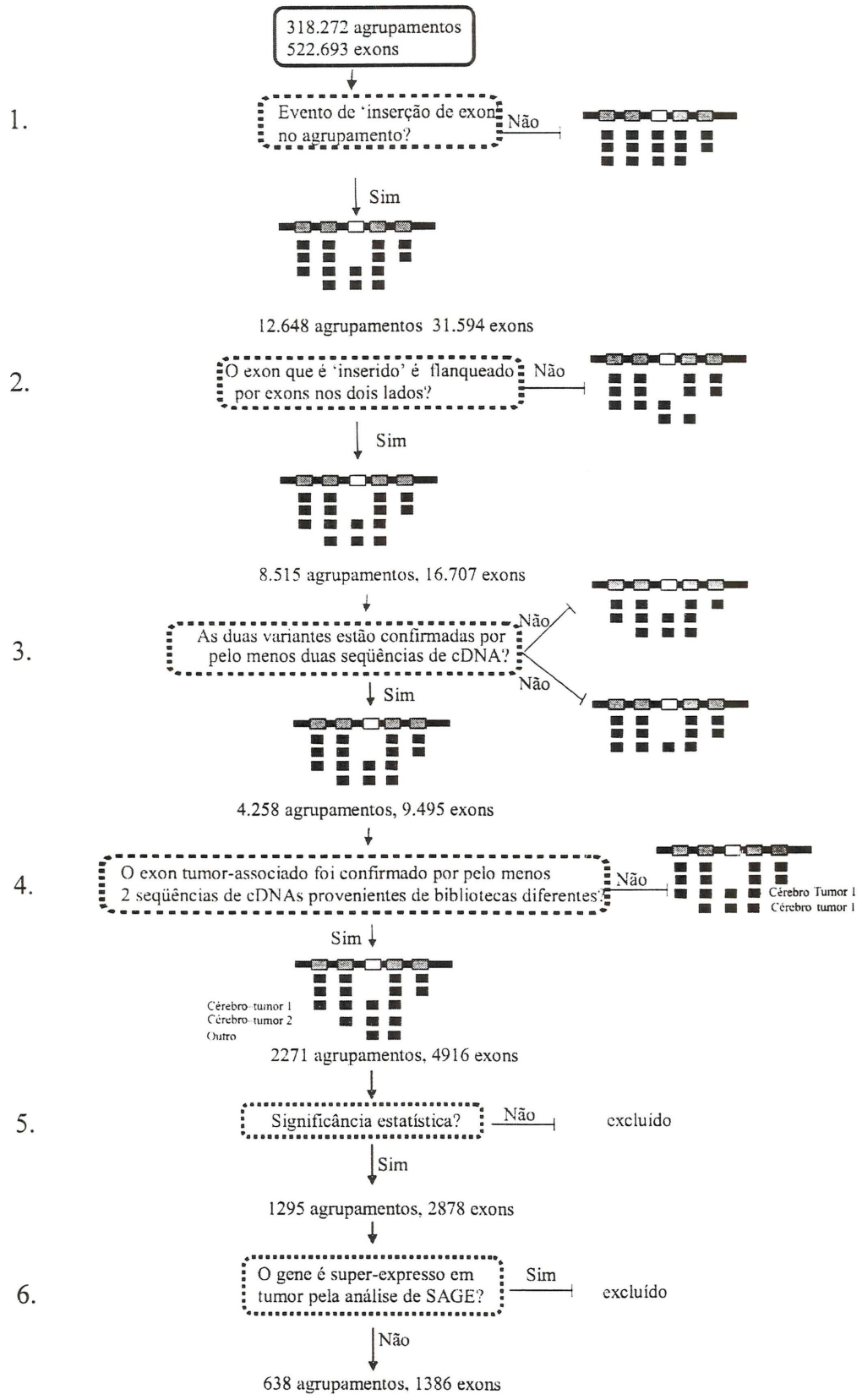
Além disso, contaminações genômicas ou processamento incompleto de mRNA podem levar à interpretação errônea da existência de uma seqüência transcrita, gerando uma informação falso-positiva sobre o número de variantes de um determinado gene. Por isso, além de alinhar seqüências de cDNA com o genoma, deve-se também filtrar os dados do *transcriptoma* para confirmar o resultado das análises. A confirmação dos dados com transcritos oriundos de mais de uma biblioteca é uma maneira indireta de demonstrar que os mesmos não são produtos de contaminação, assumindo que uma contaminação não acontecerá da mesma maneira

nas diferentes bibliotecas de cDNA.

Devido às diferentes limitações descritas acima, foram determinados critérios para selecionar os exons candidatos associados a tumor, objetivando a obtenção de um menor número possível de falso positivos.

4.2 SELEÇÃO DE CANDIDATOS ASSOCIADOS A TUMOR

Os critérios utilizados para a seleção de candidatos associados a tumor estão resumidos na Figura 7. São eles:



Legenda: A figura demonstra a abordagem utilizada para identificar exons associados a tumor. As linhas ligando retângulos representam *clusters*; os retângulos pretos em baixo dos *clusters* representam os exons presentes nas seqüências de cDNA que formam o *cluster*. O número de genes e exons, obtido depois de cada filtro de seleção, aparece na figura.

Fonte: Adaptada de KIRSCHBAUM-SLAGER et al (2004).

Figura 7 - Critérios utilizados para a seleção de variantes associadas a tumor.

1. Foram selecionados todos os *clusters* contendo pelo menos uma variante na qual existe um evento de inserção de exon. Utilizamos o termo "protótipo" para a variante onde não observamos a presença do exon inserido.
2. Para evitar a seleção de exons não verdadeiros causada por contaminações genômicas ou pela possível baixa qualidade nas extremidades das seqüências, foram selecionados apenas os exons inseridos que possuem dois exons flanqueadores.
3. Para diminuir a probabilidade de inclusão de seqüências contaminantes geradas durante o processo de produção das bibliotecas de cDNA, o sistema selecionou casos nos quais ambas as variantes (as que incluíam o exon tumor-associado, e as que não continham o exon) fossem confirmadas pela existência de pelo menos duas seqüências de cDNA oriundas de bibliotecas diferentes.
4. A expressão associada a tecido dos exons foi verificada por análise dos tecidos, que contêm a variante que apresenta o exon específico. A associação a tumor foi determinada pela existência de duas seqüências confirmando a existência do exon de cDNA de bibliotecas diferentes do mesmo tecido tumoral e a ausência de seqüências confirmando a existência do exon de bibliotecas normais do mesmo tecido.
5. Para levar em consideração o número de *EST*s que confirmavam a associação a tumor do total de *EST*s evidenciando a existência da variante candidata, um valor Z estatístico foi calculado para cada exon escolhido em todos os tecidos

nos quais o exon foi encontrado (Seção 3.4).

6. A exclusão de exons candidatos pertencendo a genes super-expressos em tumor foi realizada através de análises de padrão de expressão em tecidos normais e tumorais em nossos bancos de dados de *SAGE* (Seção 3.5).

Uma vez que podem haver problemas gerados pela detecção de *splicing* alternativo em larga escala, a nossa abordagem requer a validação experimental de todos os dados gerados pelos protocolos computacionais, os quais podem conter candidatos falso-positivos (MODREK e LEE 2002). Esta validação também pode demonstrar se a anotação de uma variante como associada a tumor ou ao tecido está correta e ela mostrará se a cobertura de *ESTs* é suficiente para permitir a seleção computacional de candidatos como associados a tumor.

4.3 ARTIGO INTITULADO: "IDENTIFICATION OF HUMAN EXONS OVER-EXPRESSED IN TUMORS THROUGH THE USE OF GENOME AND EXPRESSED SEQUENCE DATA" (KIRSCHBAUM-SLAGER et al. 2005).

Este artigo relata o desenvolvimento do sistema computacional para a busca de exons associados a tumor e a validação experimental do mesmo. Os resultados que foram publicados como informação suplementar foram inseridos no item 4.1.2.

Physiol Genomics 21: 423–432, 2005.
First published March 22, 2005; doi:10.1152/physiolgenomics.00237.2004.

Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data

Natanja Kirschbaum-Slager, Raphael Bessa Parmigiani, Anamaria Aranha Camargo, and Sandro José de Souza
Ludwig Institute for Cancer Research, São Paulo Branch, Sao Paulo, Brazil

Submitted 12 October 2004; accepted in final form 15 March 2005

Slager-Kirschbaum, Natanja, Raphael Bessa Parmigiani, Anamaria Aranha Camargo, and Sandro José de Souza. Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. *Physiol Genomics* 21: 423–432, 2005. First published March 22, 2005; doi:10.1152/physiolgenomics.00237.2004.—Alternative splicing is one of the major sources of the large transcriptional diversity found in human cells. Splicing variants have been shown to be associated with features like spreading and progression in several human tumors. Therefore, such variants may be of great importance as both diagnostic and therapeutic tools. Here, by using a set of criteria regarding the expression pattern of splicing variants and statistical analyses, we were able to screen the genome for exons overexpressed in tumors of specific tissues. However, as in other analyses attempting to identify tumor-associated variants, our list of candidates was seriously inflated with cases of genes differentially expressed in tumors. To exclude these cases and increase the probability of finding bona fide regulated splicing variants, we performed a serial analysis of gene expression (SAGE), excluding those genes that were shown to be upregulated in tumors. This allowed us to predict the overexpression of single exons in specific tumors. Our final group of candidates includes 1,386 exons belonging to 638 genes. Experimental validation of a few candidates in normal tissue, tumor cell lines, and patient samples suggests that most of these candidates are indeed tumor-associated exons. Further functional classification of our candidate genes shows that our final list is slightly inflated with cancer-related genes.

alternative splicing; tumor; transcriptome; serial analysis of gene expression

ALTERNATIVE SPLICING is one of the main sources of the variability found in the human transcriptome (3). There are four different types of alternative splicing: exon skipping/usage, alternative usage of a donor site, alternative usage of an acceptor site, and intron retention (20). Several bioinformatics analyses have indicated that at least one-half of all human genes undergo alternative splicing (7, 11, 17, 22, 23). In roughly 80% of these cases, alternative splicing invokes changes in the coding region (CDS) of genes, resulting in structural changes of the respective protein product (14, 23).

The biological impact of alternative splicing is perceptible, for example, in *Drosophila*, in which sex determination is triggered by alternative splicing of a master gene (25). Furthermore, ~15% of all human genetic diseases are believed to be caused by mutations in the splicing acceptor/donor sites, generating changes in the splicing pattern of one or more genes, which implies that alternative splicing also plays an important role in pathogenicity (19).

An apparent link between certain cancer types and alternative splicing is being investigated (for a review, see Caballero et al., Ref. 8). Several splicing variants from different genes, including *cd44*, *wt1*, *cd79b*, *bin1*, and *Syk*, have been shown to be associated with different aspects of tumorigenesis (1, 10, 13, 26, 32).

The increasing amount of cDNA libraries constructed from a diversity of both tumor and normal tissues and cell lines allows several types of computational analyses. This, together with the release of the final sequence of the human genome (16, 31), permits genome-wide analyses of alternative splicing and the search for tumor-associated splicing variants. Several groups have performed such analyses and have reported the differential expression of splicing variants in tumors (15, 33–35). None of these studies, however, systematically verified the expression pattern of the prototype variant of the same candidate gene (33, 15). Hence, it cannot be ruled out that the variants selected by their analyses as being tumor specific are variants of genes that are generally overexpressed in tumors. Furthermore, none of those studies has investigated the expression of splicing variants within tumors of one specific tissue.

Here, by using strict selection and statistical criteria, we were able to screen the genome for exons overexpressed in tumors. Tumor-associated exons are those that appear preferentially in splicing isoforms found to be overexpressed in tumors. Such exons could be of major diagnostic value, allowing the early detection of tumors based on their specific expression. New epitopes encoded by tumor-associated exons may be targeted by antibodies as well. Eventually, this should permit drug design, as the protein encoded by a spliced variant may be a therapeutic target. Here, we show by experimental, statistical, and literature validation that our set of candidates is enriched with bona fide tumor-associated splicing variants.

MATERIALS AND METHODS

cDNA mapping and clustering. All human cDNAs available in dbEST (July 2002, Ref. 4) and mRNA sequences from known human genes from UniGene release 153 (29) were aligned to the masked human genome sequence [build 29, obtained from the National Center for Biotechnology Information (NCBI)] by use of pp-Blast (27), an implementation of MEGABLAST (37) for a parallel cluster. The parameters used in MEGABLAST were: −f T −J F −F F −W 24. The MEGABLAST output was parsed, and a MySQL database was loaded with the mapping information. Spurious hits were excluded from the mapping database by use of an additional set of alignment criteria. These include a minimum degree of identity for a cDNA/genome alignment set to 93% over at least 45% of the total expressed sequence tag (EST) length or 55% of the total length of the full-insert sequence. Furthermore, for sequences mapping to more than one

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: S. J. de Souza, Ludwig Institute for Cancer Research, São Paulo Branch, Rua Prof. Antonio Prudente 109, 4 andar, São Paulo, 01509-010, SP, Brazil (e-mail: sandro@compbio.ludwig.org.br).

location on the genome, a score associated with a higher identity over a longer alignment was assigned. Clustering of cDNA sequences was based on their genomic coordinates as described by Sakabe et al. (28). Briefly, if two sequences shared at least partially the same gene structure, they were joined into the same cluster. If no exon/intron boundary was defined, a sequence had to have at least a 100-bp overlap with another sequence at the genome level to be added to the respective cluster.

Construction of the binary matrices. All sequences were represented as binary matrices, and each expressed exon was represented by 1 (one) and each skipped exon by 0 (zero). Variants were defined to skip an exon when they included two flanking exons next to an absent one (represented as 10+1, meaning that at least one exon is skipped between two flanking exons).

Z-statistics. After a screening for variants that included exons at the exact position of an exon skipping in another variant of the same cluster, a *Z*-statistic was calculated for each exon. This way, the probability of tumor association of the exon to a specific tissue, based on the numbers of ESTs confirming the variant in either tumor or normal tissue, was evaluated (33)

$$Z = (p_t - p_n)/\sqrt{p(1 - p)(1/n_n + 1/n_t)}$$

For a given exon, p_t and p_n are the expression frequencies of the exon in tumor and normal tissues, respectively, in a specific tissue (the no. of tumor or normal ESTs containing the specific exon ÷ total no. of tumor or normal ESTs from all libraries). To minimize sampling bias of small libraries, we only took into account libraries that had at least the size of the smallest library in which a transcript containing the specific exon was found in the specific tissue. The p is the geometric average frequency of the exon in tumor and normal libraries, and n_n and n_t are the numbers of ESTs in the normal and tumor libraries, respectively, taken into account for each specific exon in each tissue. In each tissue, *Z*-values having a $P \leq 0.05$ were considered significant. (It should therefore be noted that the statistically significant candidates still have a probability $P < 0.05$ of being a false-positive candidate.)

Serial analysis of gene expression tag assignment. A virtual serial analysis of gene expression (SAGE) tag is a prediction of the 10-bp sequence downstream of the 3′-most *Nla*III site of the transcript that might theoretically be produced by a SAGE experiment (5). One representative full-insert mRNA was selected from those candidate clusters that included at least one full-insert mRNA showing at least either a poly A signal and/or a poly A tail. This full insert was then assigned a virtual SAGE tag (5). The tag was assigned only to the 3′-most *NlaIII* site of the transcript. This tag was used to query all SAGE libraries of the same tissue in which we characterized the putative overexpressed exon. The frequency of each tag was counted in tumor and normal libraries of the same tissue.

Again a *Z*-statistic was calculated

$$Z = (p_t - p_n)/\sqrt{p(1 - p)(1/n_n + 1/n_t)}$$

For a given gene, p_t and p_n are the expression frequencies of the specific 3′-most SAGE tag in tumor and normal libraries, respectively, in a specific tissue (the no. of tumor or normal tags ÷ total no. of tumor or normal tags from all libraries in that tissue). The p is the geometric average frequency of the tag in tumor and normal libraries, and n_n and n_t are the numbers of tags in the normal and tumor libraries taken into account for each specific exon. *Z*-values having a $P \leq 0.05$ were considered significant (It should therefore be noted that the statistically significant candidates still have a probability $P < 0.05$ of being a false-positive candidate.)

Experimental validation. Total RNA derived from five different normal human tissues (lung, prostate, breast, brain, colon) was purchased from Clontech Laboratories and used for cDNA synthesis.

Human tumor cell lines were obtained from the American Type Culture Collection (ATCC) and maintained in appropriated medium as recommended by this organization (**http://www.atcc.org**). The following human tumor cell lines were used: A172 and T98G (glioblastoma), DU145 and PC3 (prostate), MCF-7 and MDA-MB$^-$ (breast), H1155 and H358 (lung), and SW480 (colon).

Patient samples were obtained from the Hospital A. C. Camargo tumor collection and prepared by manual dissection. All patient samples were collected after explicit informed consent, and the study was approved by the Institutional Ethics Committee.

Total RNA was extracted from tumor cell lines and tumor/normal patient samples by a conventional CsCl-guanidine thiocyanate gradient method (9), and RNA integrity was analyzed using agarose gels. Genomic DNA contamination of the total RNA was tested with PCR, using hMLH1 primers located at intronic sequences flanking exon 12 (forward, 5′-TGG TGT CTC TAG TTC TGG-3′; reverse, 5′-CAT TGT TGT AGT AGC TCT GC-3′).

Reverse transcription was carried out using the Superscript First Strand Synthesis Kit, according to the manufacturer's instructions (Invitrogen). RT-PCR reactions were carried out in a 25-μl reaction mixture containing 1 μl of cDNA, 1× *Taq* DNA polymerase buffer, 0.1 mM dNTPs, 6 pmol of each primer (for sequences of primers, see Supplemental Material; available at the *Physiological Genomics* web site),[1] 1 mM $MgCl_2$, and 1 U *Taq* DNA polymerase (Invitrogen). Standard PCR conditions were as follows: 4 min at 94°C (initial denaturation), 35 cycles of 45 s at 94°C, 45 s at 58°C, and 1 min at 72°C, with a final extension step of 10 min at 72°C. RT-PCR products were analyzed on 8% silver-stained polyacrylamide gels and on 2% ethidium bromide-agarose gels. Sequencing reactions were carried out using DYEnamic (ET Terminator Cycle Sequencing Kit, Amersham Pharmacia) and an ABI 377 prism sequencer (Perkin Elmer), according to the supplier's recommendations.

RESULTS

Transcriptome database. The database used in this work contains data obtained from alignments of all cDNA sequences to the human genome sequence (12, 28). In addition to the representation of all data concerning the alignment and clustering of the sequences, the database also contains binary matrices that were constructed for each transcript (28). In such a matrix, a transcribed exon is represented by a one (1) and an absent exon is represented by a zero (0). This approach facilitates the analysis of exon skipping/exon usage throughout the genome and the comparison of the different transcripts and exons with each other.

Our database contains 3,475,514 expressed sequences from 7,167 cDNA libraries from different tissues (see Table 1), of which 52,903 represent full-insert sequences (completely sequenced cDNA clones). Four thousand, two hundred and forty-nine (4,249) of these libraries were constructed from tumor samples and tumor cell lines, generating 1,427,390 sequences, while the remaining 2,918 libraries were constructed from normal samples, generating 2,048,124 sequences. We will refer to libraries constructed from either tumor samples or tumor cell lines as tumor libraries. Our analysis was performed on both normalized and nonnormalized libraries.

Database validation. Our clustering strategy (28) generated 318,272 cDNA clusters, 21,306 containing at least one full-insert mRNA. Of all clusters containing at least one full-insert mRNA, 52% undergo exon skipping (12), which is in agree-

[1]The Supplemental Material for this article (Supplemental Tables S1–S6, Supplemental Figs. S1–S4, Supplemental File S1) is available online at **http://physiolgenomics.physiology.org/cgi/content/full/00237.2004/DC1**.

Table 1. *No. of tumor and normal libraries, ESTs, and tissue groups in the exon-skipping database*

	Libraries	Tissue Groups	ESTs
All libraries	7,167	60	3,475,514
Tumor libraries	4,249	42	1,427,390
Normal libraries	2,918	53	2,048,124

Both types of libraries include cell lines and patient tissue. EST, expressed sequence tag.

ment with the splicing rate reported in the literature (11, 17, 23). Considering all clusters in the database, we found that 12,648 present at least one exon-skipping variant. Our analysis was performed on this latter set of clusters.

The suitability of our exon-skipping database for the current analysis was manually evaluated through the analysis of 61 genes described in the literature to have at least 2 splicing isoforms. Compared with the literature, 62% of those genes (38/61) were shown to have the same or a larger number of variants in the exon-skipping database than the number of variants published (see Supplemental Table S1). It should be noted that, although examples from the literature include all types of alternative splicing, our database considers only exon skipping/usage. This validation step confirmed that the exon-skipping database sufficiently covers the repertoire of splicing variants represented in the sequence databases.

Tumor-specific exons. We screened our database for potential tumor-associated exons, which appear in isoforms found to be exclusively expressed in tumor samples and tumor cell lines. Our clustering strategy and matrix representation allowed us to screen our database for exons that were not expressed in transcripts from any normal tissue but were expressed in transcripts from tumor libraries. For each gene, at least one transcript showing exon skipping was chosen to represent the cluster; the exon-skipping event should be confirmed by at least two cDNA sequences from different libraries. We screened for variants that would show the expression of an exon at the exact position of the skipped exon in this prototype transcript. Exons fitting into this category had to be flanked by at least two other exons; they should be represented by sequences derived from tumor libraries only. We increased the stringency of our analysis by only selecting those exon usage events that were confirmed by at least two cDNA sequences derived from different tumor libraries. Because of these stringent criteria, we were able to identify only 11 tumor-specific exons from 11 different genes (see Supplemental Material). The variants skipping these exons were expressed in several normal tissues.

Tumor-associated exons expressed in specific tissues. On the basis of the low number of candidates identified by our first approach, we decided to investigate whether a given exon could be tumor specific when its expression pattern was analyzed within one specific tissue. The search criteria for our candidates and the number of clusters filtered in each step are summarized in Fig. 1. A certain variant was defined as a candidate when it was associated with tumor samples and cell lines within one tissue only, although it could appear in normal samples of other tissues. For this purpose, all libraries were divided into tissue groups according to their annotations. Within each of the 60 selected groups, libraries were subdi-

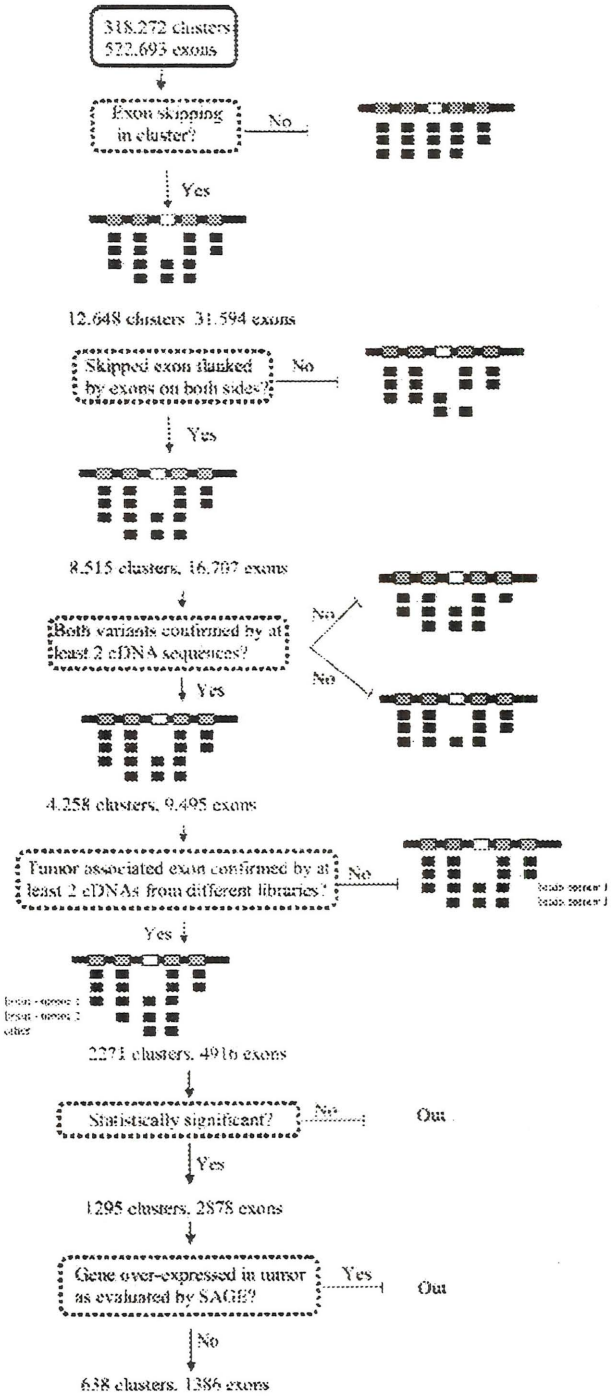


Fig. 1. Flow chart describing the approach used here to identify tumor-associated exons. Thick black lines with white boxes represent clusters; the black boxes under the clusters represent the exons present in the cDNA sequences that align to the cluster. The nos. of genes and exons obtained after each step of the screening are listed.

Table 2. *No. of candidate exons after original screening criteria, after statistical filter, and after SAGE filter*

	After Initial Criteria	After Statistical Filter	After SAGE Filter
Tumor-associated exons in all tissues	4,916	2,878	1,386
Tumor-associated exons in brain	269	233	172
Tumor-associated exons in breast	461	272	183
Tumor-associated exons in prostate	203	192	138
Tumor-associated exons in lung	239	193	156
Tumor-associated exons in colon	847	266	235

SAGE, serial analysis of gene expression.

vided into those derived from either tumor or normal tissue (see Supplemental Table S2). Only those 37 groups that included both tumor and normal libraries were used. This tissue-specific analysis increased the number of candidates to 2,271 genes, including 4,916 tumor-associated exons within different tissue types (a list of all candidate genes is available; see Supplemental Table S3). Of these genes, 2,108 contained at least one full-insert mRNA (containing 4,647 candidate exons).

Statistical filter for the tumor-associated variants. Tumor association of each of the candidate exons was tested for its statistical significance. A Z-score was calculated for each candidate exon (see MATERIALS AND METHODS) per tissue (33). This statistical approach takes into account the total number of ESTs for each tissue group in either normal or tumor libraries (see MATERIALS AND METHODS). Of the total number of candidate exons (4,916), 2,878 (59%) were shown to be significantly associated with tumors ($P < 0.05$). Of all candidates potentially associated with brain tumors (269 candidate exons), 233 (87%) exons presented a significant Z-score ($P < 0.05$). For prostate, lung, breast, and colon, 192 of 203 (95%), 193 of 239 (81%), 272 of 461 (59%), and 266 of 847 (31%) candidate exons, respectively, were shown to be significantly associated with tumors within the respective tissue (Table 2, *column 3*, and Supplemental Table S4).

Experimental validation. Seven candidates were randomly selected to be screened for expression of their putative tumor-associated exons by RT-PCR on cDNA from normal tissues and tumor cell lines. Five candidates were selected from brain, one from breast, and one from prostate, all of them passing the statistical test ($P < 0.05$) in the respective tissue. Three primers were designed for each candidate: two on the exons flanking the candidate tumor-associated exon (flanking primers), and one on the exon itself (specific primer). The products from the reaction using one flanking and one specific primer showed overexpression of the candidate variant in the respective tumor cell line (Fig. 2, data for 3 candidates). However, when using the two flanking primers, we observed that the variant skipping the exon was also overexpressed in the tumor cell lines (Fig. 2). This raised the possibility that our analysis was inflated with genes overexpressed in tumors instead of tumor-associated variants.

SAGE analysis. On the basis of the above observations, we implemented an additional filter selecting those candidate exons that did not belong to genes overexpressed in tumors. A virtual SAGE analysis was performed to verify whether the candidate genes were overexpressed in tumors from the respective tissue (5). We computationally assigned a virtual SAGE tag to one full-insert transcript of each gene (see MATERIALS AND METHODS) and statistically verified the tumor-to-normal ratio for

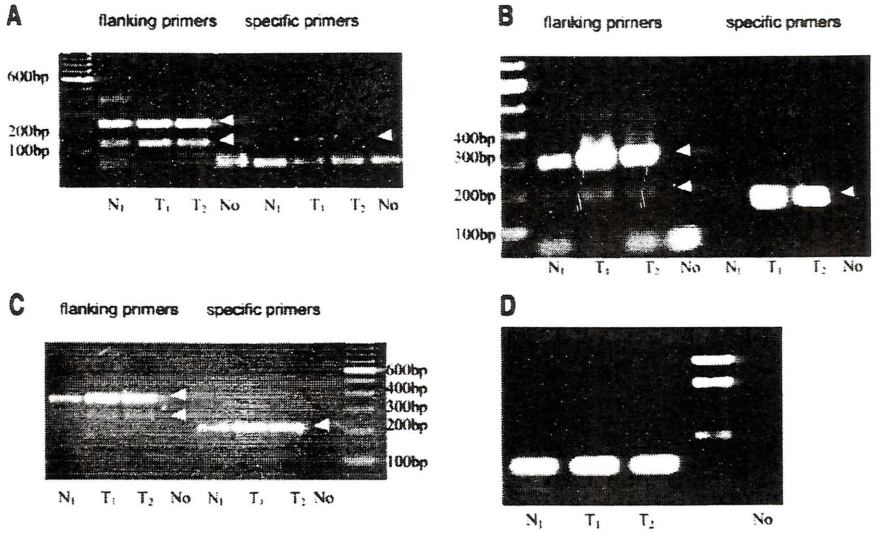


Fig. 2. Experimental validation of 3 brain tumor-associated candidates after the statistical filter. Candidates that passed the statistical filter were randomly chosen to be screened for expression of their tumor-associated exons by RT-PCR on cDNA from normal tissues and tumor cell lines. Three primers were designed for each candidate: 2 on the exons flanking the candidate tumor-associated exon and 1 annealing to the tumor-associated exon itself. We performed 2 sets of reactions, 1 using the flanking primers, which should amplify both variants, and 1 using 1 flanking and the specific primer, which should amplify only the variant expressing the candidate exon. N_1, normal whole brain tissue; T_1, T98G glioblastoma cell line; T_2, A172 glioblastoma cell line; No, "no DNA" control. Either flanking primers or specific primers were used for the amplification of variants of the following genes: *THC211630* (AJ010070; *A*), *CDK-2* (NM_052827; *B*), and *calponin 2-CNN2* (AK057960; *C*). The products from the 2nd reaction (using the primers annealing to the exon itself) showed overexpression of the candidate variant in the respective tumor cell line. However, the variant skipping the specific exon was also overexpressed in the tumor cell lines. Amplification of *GAPDH* as a positive control is shown for all samples in *D*.

that respective tag. All candidate genes showing a statistically significantly higher tag count in tumors were excluded from the list of candidates (see MATERIALS AND METHODS).

After this additional filter, 638 candidate genes, including 1,386 exons, remained in our list of candidates (Table 2, *column 4*, and Supplemental Table S5). The distribution of transcripts with more than one candidate exon is shown in Supplemental Fig. S1. The final list contained 172 candidate genes containing potential tumor-associated exons in brain, 183 in breast, 138 in prostate, 156 in lung, and 235 in colon.

We selected a few candidates for experimental validation. Using RNA extracted from tumor cell lines and normal tissues, we observed that, of the 10 candidates with conclusive results, 4 candidate genes showed that the variant containing the candidate exon was overexpressed in either brain, lung, or colon tumor cell lines, whereas the exon-skipping prototype was not (Fig. 3). The other six candidates showed a pattern similar to those in Fig. 2: both the exon-skipping variant and the variant including the selected tumor-associated exon were overexpressed in tumor tissue (results not shown).

Three of the four positively validated cases in the cell lines were also validated in patient samples (Fig. 4). Interestingly, when testing two of the six cases that were negative in cell lines, both were positively validated in some of the patient samples (Fig. 5).

Five of the six exons validated in patient samples are located inside the coding region of their respective gene. Of those five, the length of four exons is not a multiple of three and can therefore be expected to cause a change in the reading frame of its gene.

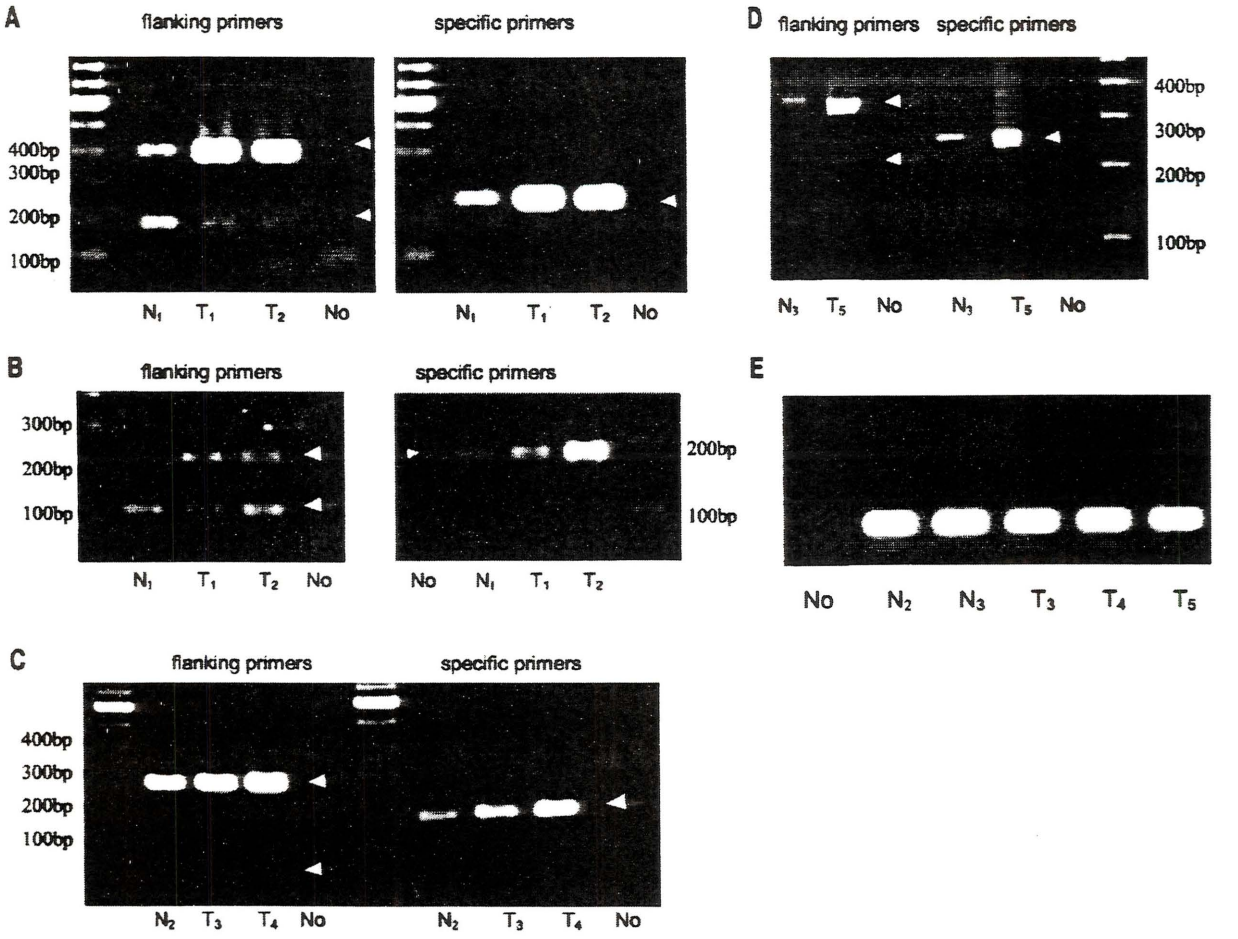


Fig. 3. Experimental validation after the serial analysis of gene expression (SAGE) filter. As in Fig. 2, candidates that passed both the statistical filter and the SAGE filter were randomly chosen to be screened for expression of their tumor-associated exons by RT-PCR on cDNA from normal tissues and tumor cell lines. Three primers were designed for each candidate: 2 on the exons flanking the candidate tumor-associated exon and 1 on the exon itself. Candidate exons corresponded to the following genes: "*Delta Tubulin*" (BC000258; *A*), *Zinc finger protein 585A* (AK074345)-T1 (*B*), *RNA terminal phosphate cyclase-like 1* (BC001025; *C*), and *karyopherin* (*importin*) *beta 1* (NM_002265; *D*). N1, normal whole brain tissue; T_1, T98G glioblastoma cell line; T_2, A172 glioblastoma cell line; N_2, normal lung tissue; T_3, H1155 lung tumor cell line; T_4, H358 lung tumor cell line; N_3, normal colon tissue; T_5, SW480 colon tumor cell line; No, no DNA control. The products from the second reaction (using the primers annealing to the exon itself) showed overexpression of the candidate variant in the respective tumor cell line. However, for the primers flanking the candidate exon, it is shown that only the variant expressing the candidate exon was overexpressed in the tumor cell lines. *A–C*: agarose gels. *D*: a silver-stained polyacrylamide gel (the variant skipping the candidate exon for this gene was only visualized this way due to its very low expression level). Amplification of *GAPDH* as a positive control is shown for all samples in *E*.

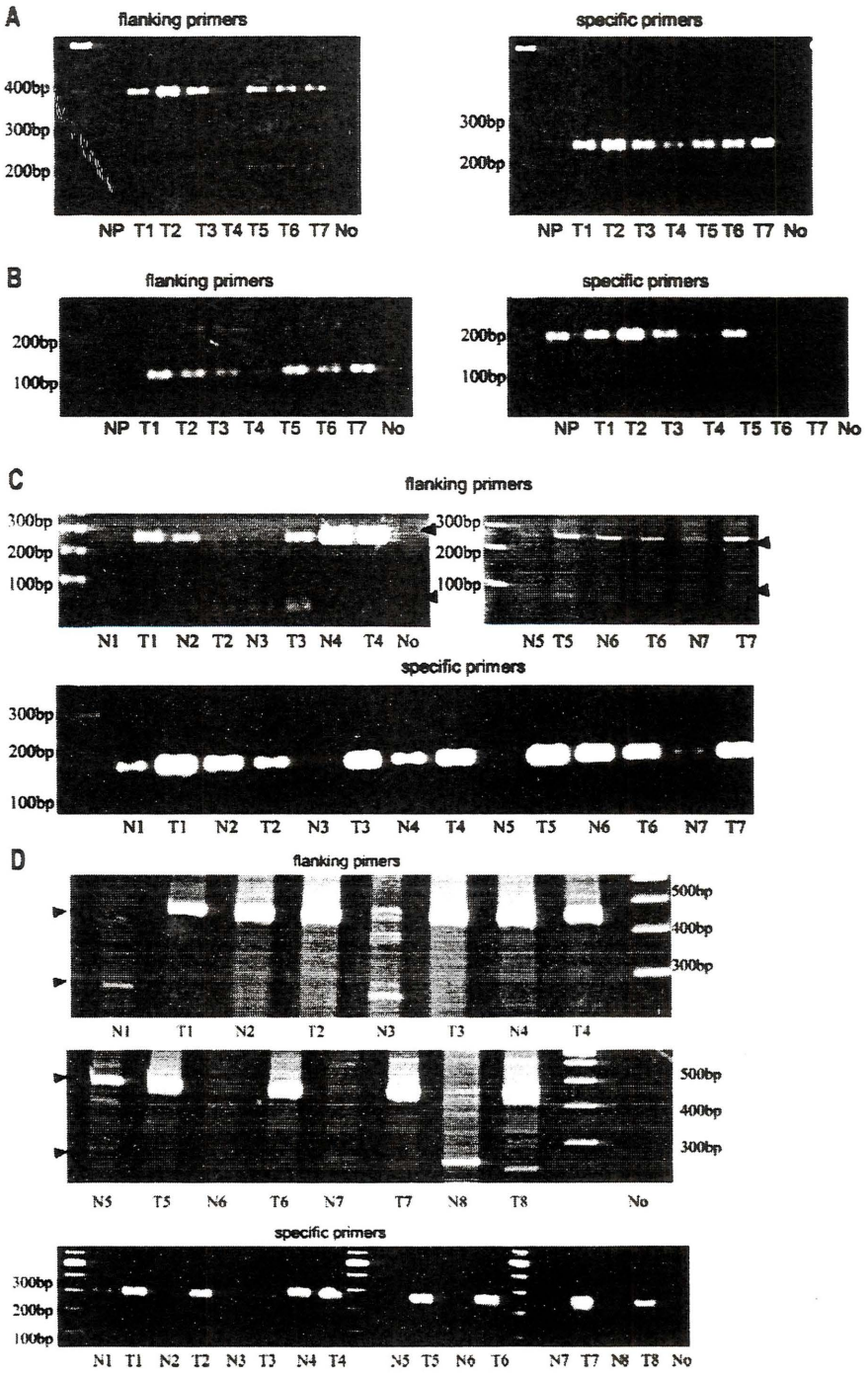


Fig. 4. Experimental validation in patient tumor samples. The 4 candidate genes that were positively validated in tumor cell lines were further validated by RT-PCR on cDNA from patient tumor samples using the same primers and conditions as before. For brain tissue, we used 7 tissue samples from glioblastoma patients and compared those with commercially obtained normal pool brain RNA (Clontech Laboratories). For lung and colon, we compared RNA from paired normal/tumor patient samples. For 3 of the 4 candidates, we obtained the same expression pattern as in cell lines in at least 2 patient samples. *A*: *Delta Tubulin* (BC000258) was validated in brain samples. The products from the reaction using the primers annealing to the exon itself showed overexpression of the candidate variant in 6 of the 7 patient samples. The primers flanking the candidate exon show that only the variant expressing the candidate exon was overexpressed in these patients. NB, pool of normal whole brain tissue; T_1–T_7, 7 different brain tumor patient samples; No, no DNA control. *B*: *Zinc finger protein 585A* (AK074345) was validated in brain samples. The products from the reaction using the primers annealing to the exon itself showed overexpression of the candidate variant in 4 of the 7 patient samples. However, the primers flanking the candidate exon showed that both the variant expressing the candidate exon and the prototype were overexpressed in these patients. NB, pool of normal whole brain tissue; T_1–T_7, 7 different brain tumor patient samples; No, no DNA control. *C*: *RNA terminal phosphate cyclase-like 1* (BC001025) was validated in paired lung samples. The products from the reaction using the primers annealing to the exon itself showed overexpression of the candidate variant in 5 of the 7 patient samples. The primers flanking the candidate exon showed that only the variant expressing the candidate exon was overexpressed in at least 2 of these patients (*patients 1* and *7*). N_1–N_7 and T_1–T_7, different paired normal/tumor lung samples, respectively; No, no DNA control. *D*: *karyopherin* (*importin*) *beta 1* (NM_002265) was validated in paired colon samples. The products from the reaction using the primers annealing to the exon itself showed overexpression of the candidate variant in 7 of 8 patient samples. The primers flanking the candidate exon showed that only the variant expressing the candidate exon was overexpressed in at least 5 of these patients (*patients 1*, *3*, *5*, *7*, and *8*). N_1–N_8 and T_1–T_8, different paired normal/tumor colon samples, respectively; No, no DNA control. Amplification of *GAPDH* as a positive control is shown in Supplemental Fig. S2.

Functional classification of the final candidate list. To analyze the functional characteristics of the final list of genes, we compared our candidates to a list of 1,127 cancer-related (CR) genes (15a). This list was a manually curated compilation based on queries of various public databases using the words "cancer" and "tumor" (for more details, see Brentani et al., Ref. 15a). The CR genes in the list constitute 5.3% of the known genes of our transcriptome database. When analyzing our 638 candidates, we found an overlap of 60 candidates (9.4%) in the CR list. Among these genes, we found *Syk* and *bin1*, which are known to have tumor-associated variants (13, 33) (for a whole list, see Supplemental Table S6). Thus there is an excess of

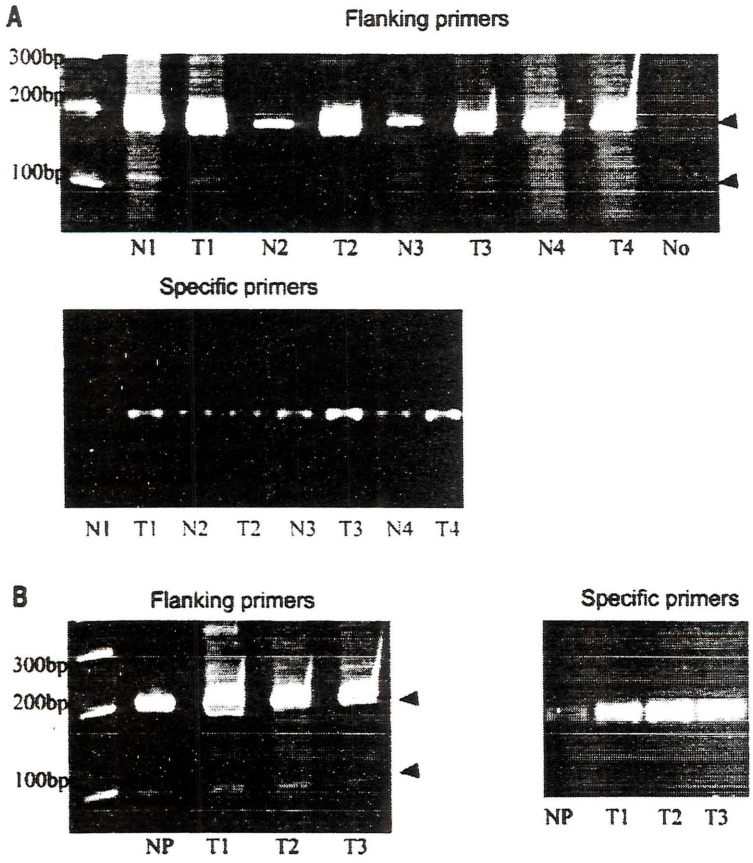


Fig. 5. We evaluated the expression pattern of 2 more candidates that were originally negative in the validation using tumor cell lines. *A*: the gene *NADH dehydrogenase (ubiquinone) Fe-S protein 2, 49-kDa (NADH-coenzyme Q reductase)* (BC001456), was validated in paired lung samples. N_1–N_4 and T_1-T_4, same paired lung normal/tumor patient samples, respectively, as in Fig. 4*C*; No, no DNA control. The products from the reaction using the primers annealing to the exon itself showed tumor overexpression of the exon in 3 of 4 paired samples. The primers flanking the candidate exon show that the variant expressing the candidate exon was overexpressed in 3 tumor samples, whereas the prototype was not overexpressed in tumors. *B*: validation of the gene "*proteasome (prosome, macropain) 26S subunit, non-ATPase, 10* (NM_002814)" in prostate normal pool and 3 patient tumor samples. The products from the reaction using the primers annealing to the exon itself showed tumor overexpression of the exon in 3 patient tumor samples. The primers flanking the candidate exon show that only the variant including the candidate exon is overexpressed in patient tumor samples. NP, normal pooled prostate cDNA; T_1–T_3, different prostate tumor patient samples; No, no DNA control.

cancer-related genes in our final list of candidates (chi-square = 7.97, 1 degree of freedom, $P = 0.005$). Comparing these results to a simulation of 200 randomly chosen sets of 638 clusters out of all UniGene clusters, we found that none of these sets presented >60 CR genes ($P < 0.005$).

The Gene Ontology (GO) terms of the final list of 638 candidates were obtained with the GOTM program (36). For 437 of the 638 genes, a GO term could be assigned (see Supplemental Fig. S3 for an overview of the distribution of the GO terms in the different GO categories and Supplemental Fig. S4 for all levels of the GO tree). The program GOstat (2) was used to analyze whether any GO category was overrepresented in our final list of candidates relative to the representation of all ontology terms in the ontology database (see Supplemental File S1). In each category, the lowest P value resulting in biologically meaningful GO terms was used. In the category "biological process," using a stringent P value cutoff of 10^{-5}, we found the GOs "intracellular protein transport" and "cell growth and maintenance" to be significantly overrepresented. In the category "molecular process," the GOs "actin binding," "receptor activity," "cytoskeletal protein binding," and "ATP binding" were significantly overrepresented (cutoff P value of 0.001). Finally, in the category "cellular components," the GO "peroxisome" was significantly overrepresented (P value cutoff = 0.001).

Literature validation. In one reported study (33) of tumor-associated splicing variants, experimental validation was performed in 76 genes chosen either by statistical criteria or by knowledge of their tumor association. All 76 candidates were experimentally validated (M. P. Lee, personal communication). To validate our candidates once more, we verified whether we could find any of our candidates in the published list of experimentally validated genes (Table 3). Thirteen of the validated candidates of this reported work were found in our initial set of candidates before the SAGE analysis. In our final list of candidates, we could only find three of their candidates.

Table 3. *Overlap of our candidates with experimentally validated candidates from Wang et al. (33)*

Gene	Overlap of Candidates Before SAGE Filter	Overlap of Candidates After SAGE Filter
NME1	+	−
CDC25C	+	−
DVL1	+	−
ERCC1	+	−
GSS	+	−
GTF3C1	+	+
IRAK1	+	−
NKTR	+	−
POLB	+	+
RAD51	+	−
SHC1	+	−
ST5	+	+
TNFRSF1A	+	−

The same comparison with a different study (35) indicated an overlapping of 21 candidates of 89 published genes (Table 4). Interestingly, we observed that 11 of the candidate genes in this last report (35) presented an overexpression in tumors as evaluated by SAGE. Taken together, these comparisons highlight the importance of filtering off genes generally overexpressed in tumors to increase the likelihood of finding bona fide tumor-associated splicing variants.

DISCUSSION

The characterization of splicing variants associated with tumors is critical for the development of new diagnostic and therapeutic strategies for the treatment of cancer. Few attempts have been made to search the human genome for tumor-associated splicing isoforms (35, 33, 15). An interesting aspect of these studies is the fact that none of them take into consideration whether the differential expression was specific to the respective splicing variant or common to all transcripts from that gene. Although some of these reports provided statistical arguments corroborating the association between the splicing isoform and tumors, most of them lack experimental validation. Wang et al. (33) showed experimental validation for the gene *RAB1A* (Fig. 2 in Wang et al., Ref. 33). There, it is possible to see that, in some samples, the prototype variant is also overexpressed in tumors. In our attempt to define exons overexpressed in tumors, we faced the same problem.

The clustering of all human cDNAs onto the human genome sequence allowed us to focus our strategy on determining the expression pattern of all exons in the human genome. We were able to seek for exons that were exclusively represented by transcript sequences derived from tumor cDNA libraries. A broad computational analysis revealed that only 11 exons were found to be expressed in tumors with no expression at all in normal tissues. This motivated us to search for exons expressed only in tumor tissues or tumor cell lines within a specific tissue. Using this strategy, we found 4,916 tumor-associated exons.

Table 4. *Overlap of our candidates with those of Xu and Lee (35), having log score >3*

Gene	Overlap of Candidates Before SAGE Filter	Overlap of Candidates After SAGE Filter
BZW2	+	+
CCT3	+	−
CGI-31	+	−
F11R	+	−
HGD	+	+
HLA-DMB	+	+
HNRPF	+	−
HNRPK	+	−
KIF2C	+	+
MBP	+	+
MED8	+	+
MGC11257	+	+
MORF4L2	+	−
NGLY1	+	+
NUDE1	+	+
SAD1	+	−
SMUG1	+	−
SPC18	+	−
TPM1	+	−
WDR4	+	+
WRB	+	−

The expression pattern of all variants was defined based on the annotation provided with the cDNA libraries publicly available. This information, however, is not always precise and clear. Because further tissue characterization of a transcript is dependent on this information, efforts are currently being made to verify the real character of all publicly available libraries (18).

To verify whether the presence of data from normalized libraries would compromise any analyses based on the relative expression levels of transcripts, we tested whether our set of candidates was inflated with sequences derived from normalized libraries. We found a number proportional to the total number of normalized data in our initial set (30% of candidates, 30% of normalized data in the initial set). This excludes the possibility that our set of candidates is enriched with artifacts due to data from normalized libraries.

To refine our analysis of tumor-associated exons, the candidates were further screened by a statistical analysis of each exon and, for a few cases, by RT-PCR validation in tumor cell lines. Roughly 41% of our candidate exons were excluded by the statistical filter.

Nowadays, experimental analysis by RT-PCR is one of the most specific ways of verifying the expression pattern of mRNA transcripts. However, while amplifying two different variants by the use of two flanking primers, competition in transcript amplification may not reflect correctly the intrinsic difference in the expression level of the two transcripts. When using a specific primer for one exon only, however, the primers may be of such specificity that they may amplify a maximum of the transcript regardless of its expression level within the cell. More sensitive methods like real-time PCR or single molecule profiling may be used to better quantify splicing variants (30, 38).

Experimental validation also showed that the whole gene, not only the candidate exon, was overexpressed in tumor cell lines. We found support for this when we performed a SAGE analysis for all of our candidates. For this approach, we assumed that the 3′-most SAGE tag is representative of the most abundant transcript of the candidate genes. We also assumed that the prototype is more abundantly expressed than the candidate variant and that the SAGE tag count is therefore an indication of the prototype expression pattern. We found that ~52% of our candidates obtained after the statistical filter represented genes overexpressed in tumors. All those cases were excluded, and a new list of candidates was produced containing 1,386 exons. Validation with those candidates showed a success rate of 40% (4/10) when tumor cell lines were used. When we used a panel of patient samples, the success rate was much higher, ~85% (5/6). This probably occurs because of both the limitations of RT-PCR and the still-limited number of SAGE libraries available today. Furthermore, the heterogeneity of tumor samples and cell line cultures provides another variable to the whole system. Only a large-scale validation scheme will allow the definitive test of our bioinformatics pipeline. However, here we show that the combination of our computational analysis with experimental validation is successful in screening for real cancer-associated exons at a success rate that would not be achieved using either a computational analysis or experimental validation alone.

Our final list of candidate genes is enriched with cancer-related genes ($P = 0.005$). As stated before, it is likely that

variants associated with cancer are found in genes that are related to cancer. On the other hand, this does not mean that variants from genes not involved in cancer would not have a functional impact on tumorigenesis (35). An ontology analysis also suggested that genes involved in intracellular protein transport and cell growth and maintenance are overrepresented in our final list of candidates. One could expect that any change in the expression level of splicing variants of genes that are connected to cell cycle and maintenance might influence cell transformation. In the category cellular components, the GO peroxisome was significantly overrepresented. Interestingly, each of the peroxisome proliferator-activated receptor (PPAR) isotypes, for example, has been shown to be involved in the pathogenesis of several tumors. PPARα induces hepatocarcinomas; PPARγ has an anti-proliferation, pro-apoptotic effect and is therefore thought to have an anti-carcinogenic effect; and PPARβ/δ is involved in the control of cell proliferation and apoptosis (21). Further investigation on the impact of the overexpression of the variants of any genes involved in the above pathways might give some insight on the possible function of specific splicing variants.

To our knowledge, this is the first report that attempts to search specifically for exons overexpressed in tumors while excluding genes that are generally overexpressed in the same tumors. Such exons may provide valuable information for future investigations on the regulation of tumor-associated alternative splicing. Finally, further experimental analysis will validate the extent to which our candidate exons are of potential diagnostic and/or therapeutic value.

ACKNOWLEDGMENTS

We thank Dr. Ricardo R. Brentani, Dr. Helena P. Brentani, Maria D. Vibranovski, Noboru J. Sakabe, and Pedro A. F. Galante for helpful discussions and/or careful reading of the manuscript.

GRANTS

N. Kirschbaum-Slager is supported by PhD fellowships from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (to 12/03) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (from 01/04). R. B. Parmigiani is supported by a PhD fellowship from CAPES. This project was supported by a Centros de Pesquisa, Inovação e Difusão Grant from FAPESP.

REFERENCES

1. **Baudry D, Hamelin M, Cabanis MO, Fournet JC, Tournade MF, Sarnacki S, Junien C, and Jeanpierre C.** WT1 splicing alterations in Wilms' tumors. *Clin Cancer Res* 6: 3957–3965, 2000.
2. **Beissbarth T and Speed TP.** GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 20: 1464–1465, 2004.
3. **Black DL.** Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72: 291–336, 2003.
4. **Boguski MS, Lowe TMJ, and Tolstoshev CM.** dbEST—database for "expressed sequence tags." *Nat Genet* 4: 332–333, 1993.
5. **Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, de Souza SJ, and Riggins GJ.** An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci USA* 99: 11287–11292, 2002.
7. **Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, and Bork P.** EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* 474: 83–86, 2000.
8. **Caballero OL, de Souza SJ, Brentani RR, and Simpson AJG.** Alternative spliced transcripts as cancer markers. *Dis Markers* 17: 67–75, 2001.
9. **Chirgwin JM, Przybyla AE, MacDonald RJ, and Rutter WJ.** Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* 18: 5294–5299, 1979.
10. **Cragg MS, Chan HTC, Fox MD, Tutt A, Smith A, Oscier DG, Hamblin TJ, and Glennie MJ.** The alternative transcript of CD79b is overexpressed in B-CLL and inhibits signaling for apoptosis. *Blood* 100: 3068–3076, 2002.
11. **Croft L, Schandorff S, Clark F, Burrage K, Arctander P, and Mattick JS.** ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet* 24: 340–341, 2000.
12. **Galante PA, Sakabe NJ, Kirschbaum-Slager N, and de Souza SJ.** Detection and evaluation of intron retention events in the human transcriptome. *RNA* 10: 757–765, 2004.
13. **Ge K, DuHadaway J, Du W, Herlyn M, Rodeck U, and Prendergast GC.** Mechanism for elimination of a tumor suppressor: aberrant splicing of a brain-specific exon causes loss of function of Bin1 in melanoma. *Proc Natl Acad Sci USA* 96: 9689–9694, 1999.
14. **Hide WA, Babenko VN, van Heusden PA, Seoighe C, and Kelso JF.** The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res* 11: 1848–1853, 2001.
15. **Hui L, Zhang X, Wu X, Lin Z, Wang Q, Li Y, and Hu G.** Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene* 23: 3013–3023, 2004.

15a. **Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium: Human Cancer Genome Project Sequencing Consortium.** The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc Natl Acad Sci USA* 100: 13418–13423, 2003.

16. **International Human Genome Sequencing Consortium.** Initial sequencing and analysis of the human genome. *Nature* 409: 860–921, 2001.
17. **Kan Z, States D, and Gish W.** Selecting for functional alternative splices in ESTs. *Genome Res* 12: 1837–1845, 2002.
18. **Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, Hide T, and Hide W.** eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* 13: 1222–1230, 2003.
19. **Krawczak M, Reiss J, and Cooper DN.** The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet* 90: 41–54, 1992.
20. **McKeown M.** Alternative mRNA splicing. *Annu Rev Cell Biol* 8: 133–155, 1992.
21. **Michalik L, Desvergne B, and Wahli W.** Peroxisome-proliferator-activated receptors and cancers: complex stories. *Nat Rev Cancer* 4: 61–70, 2004.
22. **Mironov AA, Fickett JW, and Gelfand MS.** Frequent alternative splicing of human genes. *Genome Res* 9: 1288–1293, 1999.
23. **Modrek B, Resch A, Grasso C, and Lee C.** Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* 29: 2850–2859, 2001.
24. **Modrek B and Lee C.** A genomic view of alternative splicing. *Nat Genet* 30: 13–19, 2002.
25. **Nagoshi RN, McKeown M, Burtis KC, Belote JM, and Baker BS.** The control of alternative splicing at genes regulating sexual differentiation in *D. melanogaster*. *Cell* 53: 229–236, 1988.
26. **Naor D, Sionov RV, and Ish-Shalom D.** CD44: structure, function, and association with the malignant process. *Adv Cancer Res* 71: 241–319, 1997.
27. **Osorio EC, de Souza JE, Zaiats AC, de Oliveira PS, and de Souza SJ.** pp-Blast: a "pseudo-parallel" Blast. *Braz J Med Biol Res* 36: 463–464, 2003.
28. **Sakabe NJ, de Souza JES, Galante PAF, de Oliveira PSL, Passetti F, Brentani H, Osorio EC, Zaiats AC, Leerkes MR, Kitajima JP, Brentani RR, Strausberg RL, Simpson AJG, and de Souza SJ.** ORESTES are enriched in rare exon usage variants affecting the encoded proteins. *CR Biol* 326: 979–985, 2003.
29. **Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Edwards C, Fan JB, Fang N, Fizames C, Garrett C, Green L, Hadley D, Harris M, Harrison P, Brady S, Hicks A, Holloway E, Hui L, Hussain S, Louis-Dit-Sully C, Ma J, MacGilvery A, Mader C, Maratukulam A, Matise TC, McKusick KB, Morissette J, Mungall A, Muselet D, Nusbaum HC, Page DC, Peck A, Perkins S, Piercy M, Qin F, Quackenbush J, Ranby S, Reif T, Rozen S, Sanders C, She X, Silva J, Slonim DK, Soderlund C, Sun WL, Tabar P, Thangarajah T, Vega-Czarny N, Vollrath D, Voyticky S, Wilmer T, Wu X, Adams**

MD, Auffray C, Walter NA, Brandon R, Dehejia A, Goodfellow PN, Houlgatte R, Hudson JR Jr, Ide SE, Iorio KR, Lee WY, Seki N, Nagase T, Ishikawa K, Nomura N, Phillips C, Polymeropoulos MH, Sandusky M, Schmitt K, Berry R, Swanson K, Torres R, Venter JC, Sikela JM, Beckmann JS, Weissenbach J, Myers RM, Cox DR, James MR, Bentley D, Deloukas P, Lander ES, and Hudson TJ. A gene map of the human genome. *Science* 274: 540–546, 1996.

30. **Vandenbroucke II, Vandesompele J, Paepe AD, and Messiaen L.** Quantification of splice variants using real-time PCR. *Nucleic Acids Res* 29: E68, 2001.
31. **Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, and Zhu X.** The sequence of the human genome. *Science* 291: 1304–1351, 2001.
32. **Wang L, Duke L, Zhang PS, Arlinghaus RB, Symmans WF, Sahin A, Mendez R, and Dai JL.** Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. *Cancer Res* 63: 4724–4730, 2003.
33. **Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH, and Lee MP.** Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res* 63: 655–657, 2003.
34. **Xie H, Zhu WY, Wasserman A, Grebinskiy V, Olson A, and Mintz L.** Computational analysis of alternative splicing using EST tissue information. *Genomics* 80: 326–330, 2002.
35. **Xu Q and Lee C.** Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res* 31: 5635–5643, 2003.
36. **Zhang B, Schmoyer D, Kirov S, and Snoddy J.** GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* 5: 16, 2004.
37. **Zhang Z, Schwartz S, Wagner L, and Miller W.** A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7: 203–214, 2000.
38. **Zhu J, Shendure J, Mitra RD, and Church GM.** Single molecule profiling of alternative pre-mRNA splicing. *Science* 301: 836–838, 2003.

4.4 MATERIAL SUPLEMENTAR

A seguir serão apresentados os resultados que foram publicados como material suplementar do artigo ***"Identification of human exons over-expressed in tumors through the use of genome and expressed sequence data".***

Os arquivos com conteúdo maior que duas páginas não foram incluídos na versão escrita da tese, no entanto os endereços para acessá-los estão indicados nas seções relevantes. Toda a informação suplementar aqui apresentada também está disponível em http://www.compbio.ludwig.org.br/~natanja/TAE/.

4.4.1 Seqüências dos *primers* utilizados para a validação experimental dos candidatos

Para validar o padrão de expressão de cada exon dois *primers* foram desenhados nos exons flanqueadores do exon candidato e um *primer* foi desenhado no exon candidato propriamente dito (ver Material e Métodos). As seqüências dos *primers* utilizados para a validação do padrão de expressão dos exons candidatos sequem abaixo, onde F1 e F2 correspondem aos *primers forward* e R1, R2 aos *primers reverse*.

THC211630 (AJ010070)

SPLC24F1: 5' GAG ACC TTA AAC CTC AGA A 3'

SPLC24F2: 5' CAA ATA TTA TTC CAC AGC TG 3'

SPLC24R1: 5' TAC TAA AAT CTT GCC GGG 3'

CDK-2 (NM_052827)

SPLC29F1: 5' CAT GGT CAG CTA CGG CAT 3'

SPLC29F2: 5' AAG CAG GAG CGG AAT TTC 3'

SPLC29R1: 5' CTT GGG GTC ATA GAG ATG 3'

Calponin 2 (AK057960)

SPLC40F1: 5' GGC GAG CCG AGT GAA G 3'

SPLC40F2: 5' CAG ACT CCG TGA AGA AAG 3'

SPLC40R: 5' CAG CAA TTT CTT CCT CTT CC 3'

Delta Tubulin (BC000258)

SPLC43F: 5' GCA CAG ATC TTA TGG ATG G 3'

SPLC43R1: 5' GACTCTTTCAGTGGTTTTTTC 3'

SPLC43R2: 5' GTCTGGCCAATGGAAATATG 3'

Zinc finger protein 585A (AK074345)

SPLC46F1: 5' GCA TCA CAT CCC GGT AC 3'

SPLC46F2: 5' GAC TTT GGC TTT GGA GTG 3'

PLC46R : 5' CTA TGG CAG CTC CTA TG 3'

RNA terminal phosphate cyclase-like 1(BC001025)

SPLC61F1: 5' GCTTCATAAGGCTATTGGAC 3'

SPLC61F2: 5' GTATGGTGGATCTGTGGAACATG 3'

SPLC61R: 5'GGAGTGCTGTTGCCTTAAGGAAC 3'

Karyopherin (importin) beta 1 (NM_002265)

SPLC55F1: 5' CTT TAC AGA ATC TGG TGA AG 3'

SPLC55F2: 5' GAA AAG TGA CAT TGA TGA GG 3'

SPLC55R: 5' CAT CAC TGC TGC ATC CC 3'

NADH dehydrogenase (ubiquinone) Fe-S protein 2, 49kDa (NADH-coenzyme Q reductase) (BC001456)

SPLC60F1: 5' GTA CAG CCA TTT TGA TTG ACA GC 3'

SPLC60R1: 5' CAG CTG GAC ATA CAG AAA TTG TTG 3'

SPLC60R2: 5' GAC AGC GAA ATG GAG GGG 3'

Proteasome (prosome, macropain) 26S subunit, non-ATPase, 10 (NM_002814)

SPLC66F1: 5' CCA TAC TTT GAC CGG CTA GAC 3'

SPLC66F2: 5' GGC TGT TTG AAG AAA GGG AGA AG 3'

SPLC66R1: 5' GGC CGG ATA TAA GCA GCA TG 3'

4.4.2 Tabela Suplementar 1

O banco de dados de *splicing* alternativo foi validado manualmente através da análise de 61 genes já descritos na literatura, os quais possuem pelo menos duas variantes de *splicing*. Todos os genes foram encontrados em nosso banco de dados, sendo que 38 dos 61 genes (62%) possuem um número maior de variantes do que o número já descrito na literatura. Esta tabela mostra os genes encontrados na literatura, o número de variantes encontradas para cada gene e o número de variantes encontradas no nosso banco de dados para o mesmo gene.

Tabela 1 – referente à Tabela suplementar 1 do artigo: Comparação do número de variantes de genes encontrados na literatura com o número de variantes dos mesmos proveniente do nosso banco de dados.

Gene	Número de variantes - literatura	Número de variantes - banco de dados	UniGene *Cluster*	Referência
(HAS)SEMA6B	2	3	Hs.148932	Genomics 2001 May 1;73(3):343-8
5-HT2CR	2	2	Hs.46362	J Neurochem. 2003 Dec;87(6):1402-12.
5-HT4R	6	4	Hs.113262	Neuropharmacology 2002 Jan;42(1):60-73
ABCA7	2	3	Hs.134514	Biochem Biophys Res Commun. 2003 Nov 14; 311(2): 313-8
ACF	9	2	Hs.8349	Biochim Biophys Acta. 2001 Nov 11;1522(1):22-30.
Act1	2	7	Hs.437508	Biochem.Biophys.Res 2002.296:406-412.
AIG1	2	13	Hs.390662	Mol Cells. 2001 Feb 28;11(1):35-40
Alpha II spectrin	8	1	Hs.387905	Biochemistry. 1999 Nov 30;38(48):15721-30
Ang-1	4	4	Hs.2463	Blood. 2000 Mar 15;95(6):1993-9
APC tumour suppressor gene	16	4	Hs.2463	Cancer Res. 1997 Feb 1;57(3):488-94.
ATP2A3/SERCA3	3	4	Hs.5541	Am J Physiol Cell Physiol 275: C1449-C1458, 1998
BACE	3	3	Hs.49349	Biochem Biophys Res Commun 2002 Apr 26;293(1):30-7
BDNF	7	2	Hs.439027	J Neurochem 2002 Sep;82(5):1058-64
BIN1	3	34	Hs.193163	Poc.Natl.Acad.Sci USA 1999. 96;9689-9694,
CaMKII	4	6	Hs.111460	Mol Biol Rep 2001 Mar;28(1):35-41
CC3	2	2	Hs.90753	Mol. Cel. Biol. 2000. 583-593
CD79b	2	2	Hs.89575	Blood, 2002.(199;9):3068-3076
CHODL	4	1	Hs.283725	J Biol Chem. 2003 May 23;278(21):19164-70
Cortactin	3	4	Hs.301348	J. Biol. Chem., Vol. 278, Issue 46, 45672-45679, November 14, 2003
CPM	4	2	Hs.334873	Biol Chem. 2002 Feb;383(2):263-9.
DGKH	2	3	Hs.378969	J Biol Chem. 2003 Sep 5;278(36):34364-72
Dmd	3	21	Hs.169470	Acta Histochem 2002;104(3):245-54
Dnmt1	2	7	Hs.202672	J Biol Chem. 2000 Apr 14;275(15):10754-60
ELKS	5	6	Hs.306711	Genes Chromosomes Cancer 2002 Sep;35(1):30-7
EP3	8	2	Hs.527970	Prostaglandins Other Lipid Mediat 2001 Mar;63(4):165-74

Gene	Número de variantes - literatura	Número de variantes - banco de dados	UniGene Cluster	Referência
FOXP2	3	1	Hs.282787	Hum Genet 2002 Aug;111(2):136-44
GDPD1	2	1	Hs.249795	Int J Mol Med. 2003 Dec;12(6):1003-7.
Glut9	2	5	Hs.95497	J Biol Chem. 2004 Jan 22
GRIK2 kainate GluR6 receptor subunit	5	2	Hs.307494	Gene 2001 Aug 22;274(1-2):187-97
hDII	3	3	Hs.436020	Mol Cell Endocrinol 2001 Feb 14;172(1-2):169-75
hDMP1	3	4	Hs.129506	J Biol Chem. 2003 Oct 31;278(44):42750-60. Epub 2003 Aug 12.
hGAC, glutaminase	3	8	Hs.128410	Physiol Genomics. 1999 Aug 31;1(2):51-62
Hippostasin: kallikrein-like protease (PRSS20/KLK11),	3	3	Hs.57771	Prostate. 2003 Mar 1;54(4):299-305
HOX11	3	1	Hs.89583	Gene. 2003 Dec 24;323:89-99.
hSK1	4	1	Hs.158173	Biochemistry 2001 Mar 13;40(10):3189-95
hTERT	4	1	Hs.439911	Neoplasia. 2003 May-Jun;5(3):193-7.
IP4P	3	5	Hs.334575	Biochem Biophys Res Commun 2001 Aug 10;286(1):119-25
KAI1/CD82	2	5	Hs.323949	Cancer Research 63, 7247-7255, 2003
LDHC	6	18	Hs.99881	Cancer Research 62, 6750-6755
Mdm2	11	18	Hs.212217	Oncol Res. 2000;12(11-12):451-7
MG61/PORC	4	1	Hs.386453	Gene 2002 Apr 17;288(1-2):147-57
MMP-28, epilysin	2	3	Hs.380710	J Biol Chem 2001 Mar 30;276(13):10134-44
NABC1/BCAS1	2	6	Hs.400556	Genomics 2000.65;299-302
NER	2	2	Hs.432976	Oncogene 1997, 14, 617-621
Ovarian carcinoma immunoreactive antigen (OCIA)	2	7	Hs.170291	Biochem Biophys Res Commun 2001 Jan 12;280(1):401-6
p63	6	4	Hs.137569	Anticancer Res. 2003 Sep-Oct;23(5A):3945-8.
PLGF	4	1	Hs.252820	J Reprod Immunol. 2003 Oct;60(1):53-60.
PPIL3	2	4	Hs.121076	Cytogenet Cell Genet 2001;92(3-4):231-6
PSMA	4	5	Hs.1915	Int J Cancer. 2003 Nov 1;107(2):323-9.
PTHrP	3	4	Hs.89626	Clinical Chemistry 49: 1398-1402, 2003

Gene	Número de variantes - literatura	Número de variantes - banco de dados	UniGene Cluster	Referência
RAGE	5	2	Hs.184	Biochim Biophys Acta. 2003 Oct 20;1630(1):1-6
Smac/DIABLO	3	2	Hs.169611	J. Biol. Chem., Vol. 278, Issue 52, 52660-52672, December 26, 2003
Survivin	3	2	Hs.1578	Cell Death Differ. 2002 Dec;9(12):1334-42.
Synaptoporin	2	6	Hs.26411	Mol Biol Rep. 2003 Sep;30(3):185-91.
TREK-2	3	1	Hs.365690	J Physiol 2002 Mar 15;539(Pt 3):657-68
Trisk 51	2	2	Hs.159090	Biochem Biophys Res Commun. 2003 Apr 4;303(2):669-75.
TrkB	3	6	Hs.439109	Biochem Biophys Res Commun 2002 Jan 25;290(3):1054-65
tTG	2	2	Hs.512708	J Biol Chem 2001 Feb 2;276(5):3295-301
Tyrosine kinase: p59fyn	3	18	Hs.390567	Biochem Biophys Res Commun 2002 Nov 8;298(4):501-4
WT1	4	2	Hs.1145	Clin Cancer Res. 2000 Oct;6(10):3957-65.
ZNF74	6	1	Hs.127476	DNA Cell Biol 2001 Mar;20(3):159-73

4.4.3 Genes contendo exons tumor específicos

Selecionamos do nosso banco de dados exons potenciais a serem exons específicos de tumor. Estes exons foram selecionados por serem confirmados através da existência de seqüências expressas unicamente em amostras de tumor e linhagens celulares tumorais independentemente do tipo histológico do tecido. Identificamos apenas 11 exons tumor específicos. Os genes que incluem estes 11 exons são:

1 Splicing factor 3a, subunit 3, 60kD	SF3A3
2 Peptidylprolyl isomerase E (cyclophilin E)	PPIE
3 RAP1, GTPase activating protein 1	RAP1GA1
4 Latrophilin	LPHH1
5 COBW-like protein	LOC55871
6 Immunoglobulin kappa constant	IGKC

7 Moderately similar to A43932 mucin 2 precursor		
8 Thyroglobulin	TG	
9 BRAF35/HDAC2 complex (80 kDa)	BHC80	
10 Acidic (leucine-rich) nuclear phosphoprotein 32 family, member D		ANP32D
11 Immunoglobulin lambda locus	IGL	

4.4.4. Tabela suplementar 2

Esta tabela mostra os grupos de órgãos nos quais as bibliotecas foram divididas baseado em suas anotações. Cada grupo foi subdividido em tecido tumoral e normal. Apenas os grupos contendo bibliotecas tumorais e normais foram utilizados em nossa análise.

Tabela 2 – referente à Tabela suplementar 2 do artigo: Os órgãos nos quais as bibliotecas foram divididas e o número de seqüências observadas em cada órgão.

Órgão	Bibliotecas de tumor	Seqüências tumorais	Bibliotecas normais	Seqüências normais
Baco	0	0	7	19334
Bexiga	42	26281	14	1767
Boca	8	20049	9	1607
Cabeça pescoço	912	137167	103	17110
Cartilagem	13	33300	8	18221
Células Tronco	0	0	2	2946
Cérebro	76	170498	111	276442
Cólon	754	180068	149	60042
Coração	1	116	15	45056
Cordão Umbilical	0	0	7	10588
Denis_drash	63	9243	0	0
Duodeno	1	14545	0	0
Embrião	2	38516	8	12425
Epidídimo	37	5775	3	1252
Esôfago	4	32	4	200
Estômago	213	54652	88	51002
Fígado	11	67879	47	180226
Fluido / líquido amniótico normal	0	0	62	10046
Gengiva	0	0	3	919
Glândula adrenal	6	16382	11	10959
Glândula Parétida	0	0	1	37
Glândula Pineal	0	0	3	8810
Glândula Pituitária	1	1598	7	8887
Intestino	1	799	0	0
Intestino Delgado	0	0	4	876
Laringe	2	19	3	1124
Linfonodo	8	59902	15	26102
Mama	733	103261	337	74396
Músculo	67	86377	20	65216
Nariz	4	1608	4	2666
Nasofaringe	0	0	2	1640
Olho	2	36810	40	111980
Omento maior	0	0	7	398
Osso	262	60350	19	24216
Ouvido	2	3851	3	18587
Ovário	146	87642	15	16481
Pâncreas	9	82400	23	68404
Paratireóide	1	22412	4	105
Pele	10	115790	23	35862
Placenta	4	24133	343	173562
Próstata	137	46721	154	75655
Pulmão	190	117246	121	168117

Órgão	Bibliotecas de tumor	Seqüências tumorais	Bibliotecas normais	Seqüências normais
Tonsila	0	0	4	18436
Traquea	0	0	2	29
Trato Gênito-urinário	1	5545	0	0
Útero	206	146712	13	100702
Vagina	1	5	0	0
Veia	0	0	10	9478
Vesícula Biliar	0	0	4	3068
Rim	70	81999	17	74318
Sangue	2	4320	26	134743
Sistema Nervoso	218	40822	331	70021
Tecido adiposo	0	0	7	3578
Tecido conectivo	1	1469	0	0
Testículo	12	2733	152	51473
Timo	0	0	16	5550
Tireóide	6	1743	15	7585

4.4.5 Tabela suplementar 3

A análise de candidatos associados a tumor, compartimentalizados por tipo de tecido, resultou em um número de candidatos maior: 2271 genes incluindo 4916 exons associados a tumor em diferentes tipos de tecidos. Esta tabela mostra todos os candidatos e está disponível no endereço de Internet: http://www.compbio.ludwig.org.br/~natanja/TAE/names_candidates.

4.4.6 Tabela suplementar 4

A significância estatística de associação tumoral de cada exon candidato foi avaliada. Um valor Z foi calculado para cada exon candidato (ver Material e Métodos) para cada tecido. Os candidatos com valores estatisticamente significantes estão apresentados nesta tabela, que está disponível no endereço de Internet: http://www.compbio.ludwig.org.br/~natanja/TAE/candidates_afterstat.htm.

4.4.7 Tabela suplementar 5

Implementamos um filtro adicional excluindo os exons candidatos que pertenciam a genes super-expressos em tumores. Uma análise virtual de *SAGE* foi feita verificando se os genes da nossa lista de candidatos eram super-expressos em tecidos de tumor e em linhagens celulares dos tecidos testados. A tabela suplementar 5 contém os genes candidatos que não apresentam super-expressão em tecido tumoral e está disponível no endereço de Internet: http://www.compbio. ludwig. org.br/~natanja/TAE/candidates_afterSAGE.htm.

4.4.8 Figura suplementar 1

Esta figura demonstra a distribuição dos exons candidatos por *cluster*.

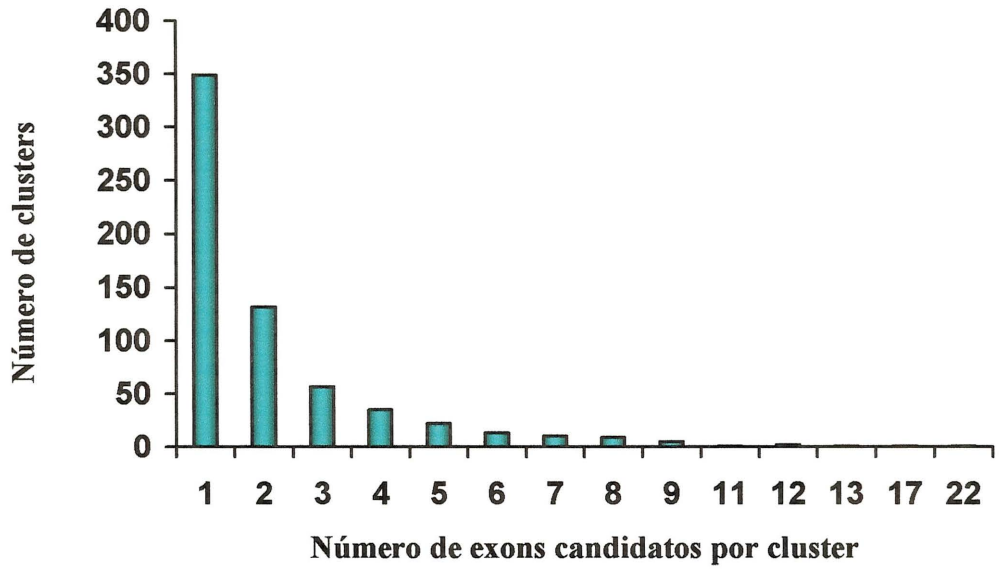
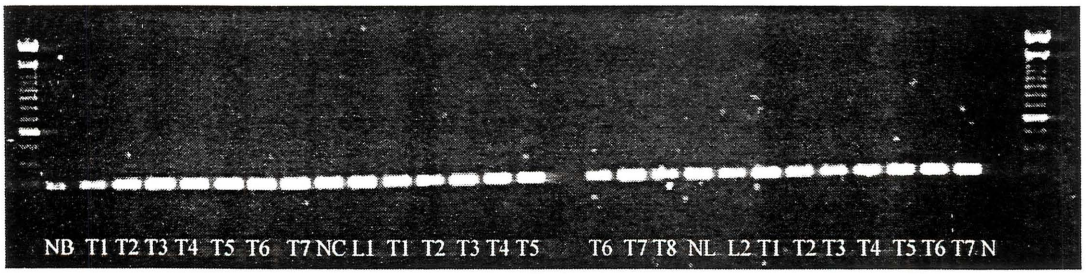


Figura 8 - Figura Suplementar 1 do artigo – Distribuição de exons candidatos por *cluster*

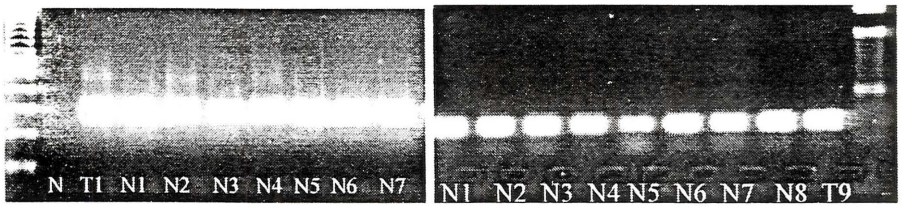
4.4.9 Figura suplementar 2

A.



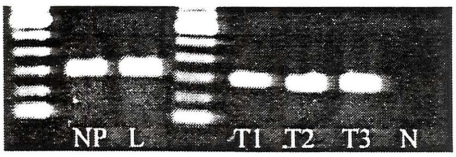
Cérebro Cólon Pulmão

B.



Pulmão Cólon

C.



Próstata

Legenda A -NB - "pool" de tecido cerebral normal, NC - "pool" de tecido de cólon normal, NL - "pool" de tecido pulmonar normal. L_1 corresponde à linhagem tumoral celular, SW480 – linhagem de cólon, L_2 - linhagem tumoral celular H1155 de pulmão. "N" corresponde ao controle negativo da reação, isto é, sem adição de cDNA. $T_1 - T_8$ correspondem às amostras tumorais de pacientes de cada tecido. B - "N" representa o controle negativo da reação, N1 – N8 - amostras normais de pulmão e cólon, T_1 - amostra tumoral de pulmão, T_9 - amostra tumoral de cólon. C - Amplificação do gene *GAPDH* como controle positivo nas amostras de próstata. NP – Tecido normal de próstata, T1 – T3 são amostras diferentes de tumor de próstata. L corresponde à linhagem celular de próstata PC3. "N" - controle negativo da reação.

Figura 9 - Amplificação do gene *GAPDH* como controle positivo em todas as amostras utilizadas.

4.4.10 Tabela suplementar 6

Comparamos nossa lista de 638 genes candidatos com uma lista de genes publicada contendo 1127 genes relacionados ao câncer (BRENTANI et al. 2003). Nesta tabela estão anotados os nossos candidatos relacionados ao câncer.

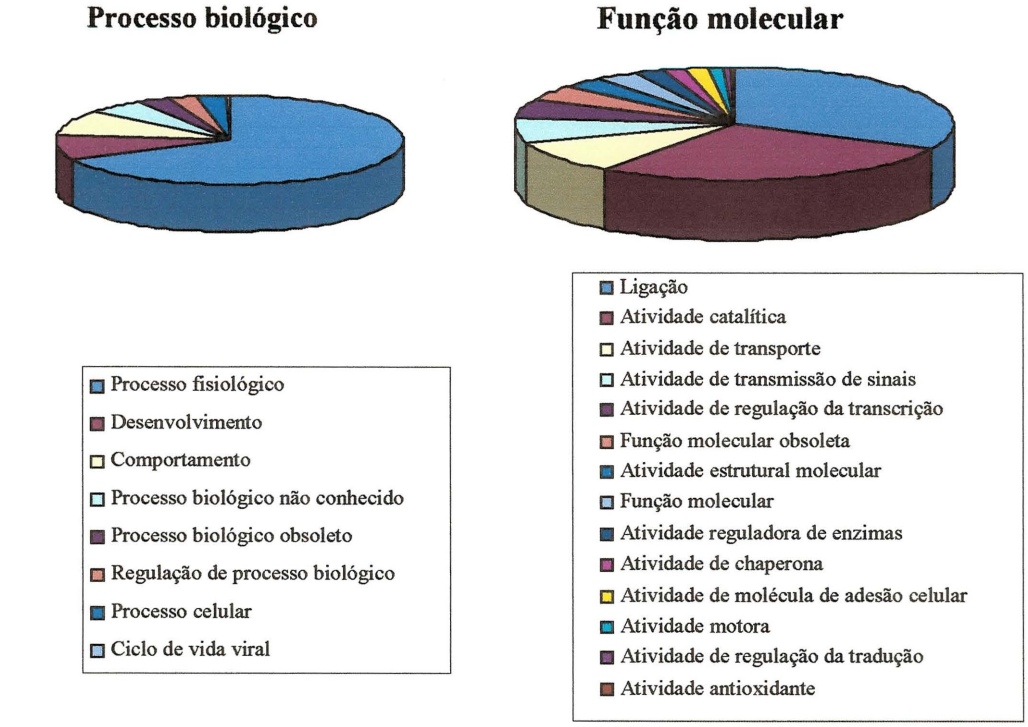
Tabela 3 - referente à Tabela suplementar 6 do artigo: Os candidatos que foram indicados a serem relacionados ao câncer por um estudo baseado em pesquisas individuais de diferentes bancos de dados públicos utilizando as palavras "Câncer" e "Tumor" (BRENTANI et al. 2003). As anotações de cada gene foram extraídas utilizando o programa SOURCE: (http://source.stanford.edu) (DIEHN et al. 2003).

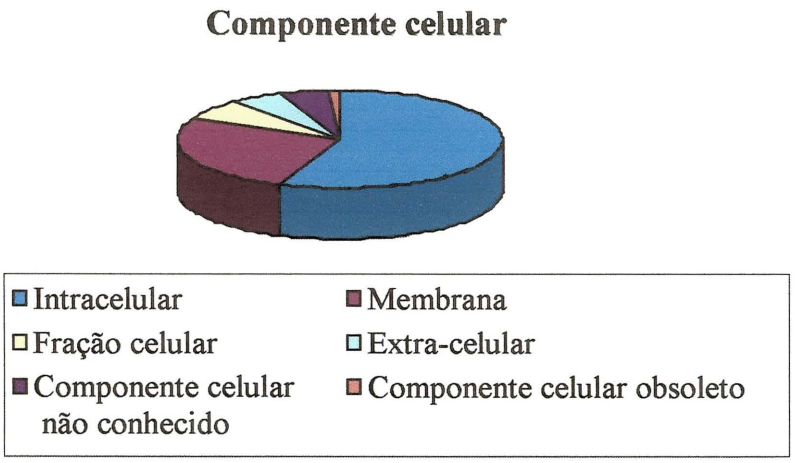
Cluster do UniGene	Nome	Símbolo
Hs.103755	receptor-interacting serine-threonine kinase 2	RIPK2
Hs.112255	nucleoporin 98kDa	NUP98
Hs.115325	RAB7, member RAS oncogene family-like 1	RAB7L1
Hs.117715	suppression of tumorigenicity 5	ST5
Hs.1211	acid phosphatase 5, tartrate resistant	ACP5
Hs.12124	elaC homolog 2 (E. coli)	ELAC2
Hs.127686	interferon regulatory factor 4	IRF4
Hs.132911	pleiomorphic adenoma gene-like 1	PLAGL1
Hs.148221	RAD52 homolog (S. cerevisiae)	RAD52
Hs.149846	integrin, beta 5	ITGB5
Hs.150136	mitogen-activated protein kinase 7	MAPK7
Hs.153934	core-binding factor, runt domain, alpha subunit 2; translocated to, 2	CBFA2T2
Hs.153961	ARP1 actin-related protein 1 homolog A, centractin alpha (yeast)	ACTR1A
Hs.1592	CDC16 cell division cycle 16 homolog (S. cerevisiae)	CDC16
Hs.160958	CDC37 cell division cycle 37 homolog (S. cerevisiae)	CDC37
Hs.174051	small nuclear ribonucleoprotein 70kDa polypeptide (RNP antigen)	SNRP70
Hs.1770	ligase I, DNA, ATP-dependent	LIG1
Hs.180107	polymerase (DNA directed), beta	POLB
Hs.180566	mucosa associated lymphoid tissue lymphoma translocation gene 1	MALT1
Hs.191842	cadherin 3, type 1, P-cadherin (placental)	CDH3
Hs.19192	cyclin-dependent kinase 2	CDK2
Hs.192023	eukaryotic translation initiation factor 3, subunit 2 beta, 36kDa	EIF3S2
Hs.192182	spleen tyrosine kinase	SYK
Hs.193163	bridging integrator 1	BIN1
Hs.194143	breast cancer 1, early onset	BRCA1
Hs.20084	retinoid X receptor, alpha	RXRA
Hs.231411	transforming growth factor beta regulator 4	TBRG4
Hs.233765	TCF3 (E2A) fusion partner (in childhood Leukemia)	TFPT
Hs.243491	caspase 8, apoptosis-related cysteine protease	CASP8
Hs.262886	inositol polyphosphate-5-phosphatase, 145kDa	INPP5D
Hs.265829	integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)	ITGA3
Hs.313	secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T-lymphocyte activation 1)	SPP1
Hs.326445	v-akt murine thymoma viral oncogene homolog 2	AKT2

Cluster do UniGene	Nome	Símbolo
Hs.350631	A kinase (PRKA) anchor protein 13	AKAP13
Hs.35140	serine/threonine kinase 4	STK4
Hs.355724	CASP8 and FADD-like apoptosis regulator	CFLAR
Hs.374378	CDC28 protein kinase regulatory subunit 1B	CKS1B
Hs.426324	tumor suppressor candidate 3	TUSC3
Hs.43080	platelet derived growth factor C	PDGFC
Hs.433618	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	MLH1
Hs.433892	adaptor-related protein complex 2, mu 1 subunit	AP2M1
Hs.436852	fibroblast activation protein, alpha	FAP
Hs.439683	karyopherin (importin) beta 1	KPNB1
Hs.443057	CD53 antigen	CD53
Hs.443960	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 11 (CHL1-like helicase homolog, S. cerevisiae)	DDX11
Hs.446352	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	ERBB2
Hs.446414	CD47 antigen (Rh-related antigen, integrin-associated signal transducer)	CD47
Hs.484782	DNA fragmentation factor, 45kDa, alpha polypeptide	DFFA
Hs.512682	carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)	CEACAM1
Hs.53250	bone morphogenetic protein receptor, type II (serine/threonine kinase)	BMPR2
Hs.63489	protein tyrosine phosphatase, non-receptor type 6	PTPN6
Hs.6764	histone deacetylase 6	HDAC6
Hs.73090	nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)	NFKB2
Hs.77500	ubiquitin specific protease 4 (proto-oncogene)	USP4
Hs.77917	ubiquitin carboxyl-terminal esterase L3 (ubiquitin thiolesterase)	UCHL3
Hs.79026	myeloid leukemia factor 2	MLF2
Hs.79058	suppressor of Ty 4 homolog 1 (S. cerevisiae)	SUPT4H1
Hs.83347	angio-associated, migratory cell protein	AAMP
Hs.95577	cyclin-dependent kinase 4	CDK4
Hs.97616	SH3-domain GRB2-like 1	SH3GL1

4.4.11 Figura suplementar 3

Esta figura demonstra a distribuição dos termos de ontologia dos nossos genes candidatos por categoria de GO obtida utilizando o programa GOTM (**http://genereg.ornl. gov/gotm/** (ZHANG et al. 2004).





Fonte: Figura obtida utilizando o programa http://genereg.ornl.gov/gotm/ (ZHANG et al. 2004).

Figura 10 - Divisão de genes candidatos por categoria de ontologia gênica.

4.4.12 Figura suplementar 4

O programa GOTM (http://genereg.ornl.gov/gotm, (ZHANG et al. 2004) foi utilizado para analisar os termos de ontologia gênica de cada um dos genes candidatos (www.ontology.com). O arquivo resultando do GOTree está disponível na página http://www.compbio.ludwig.org.br/~natanja/TAE/tree.

4.4.13 Arquivo suplementar 1

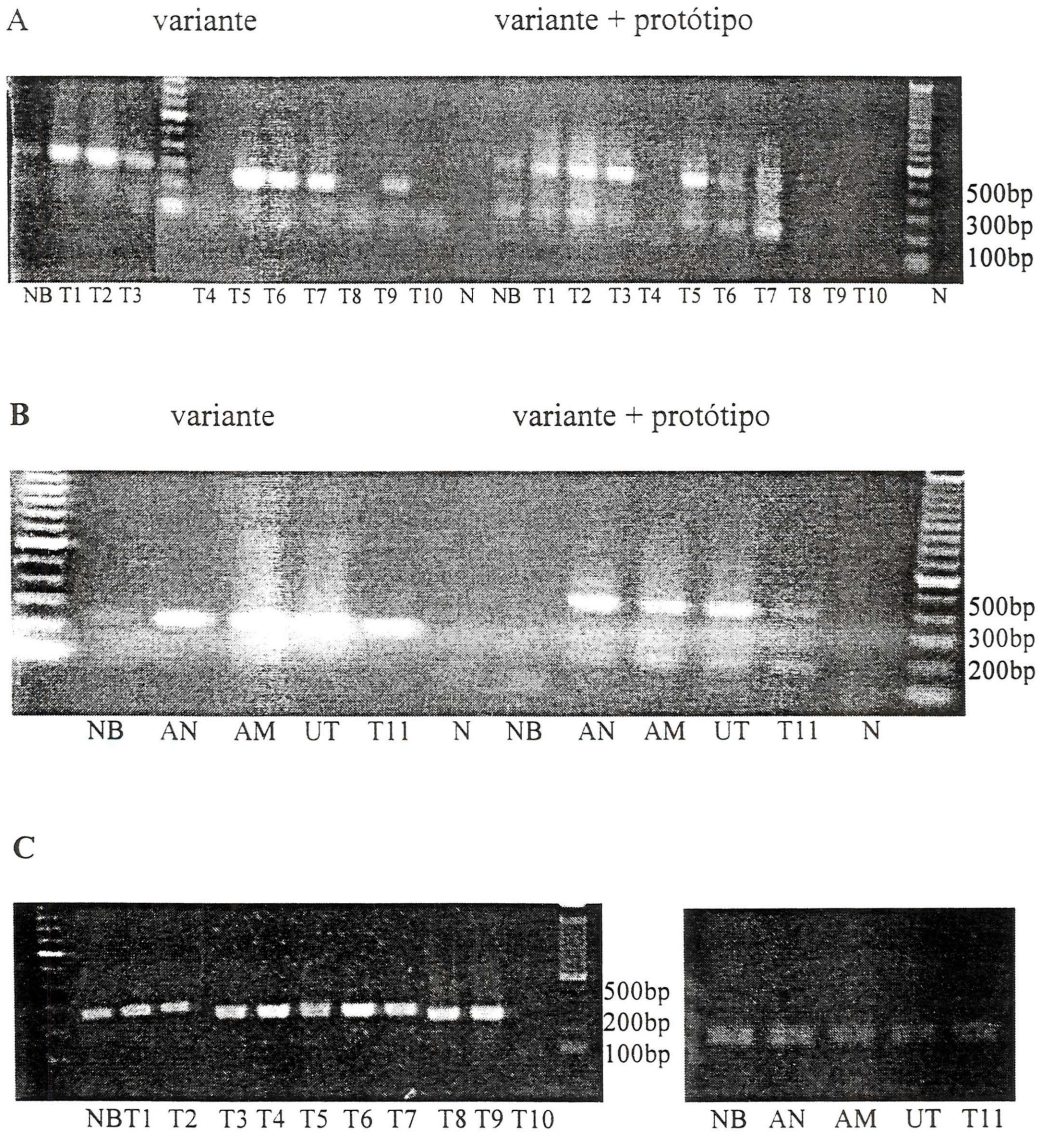
O programa GOstat (BEISSBARTH et al. 2004) foi utilizado para analisar a possível super-representação de categorias de GO na lista final de genes candidatos em relação a todas as proteínas humanas anotadas no banco de dados de ontologia. Os arquivos gerados pelo programa estão disponíveis na página http://www.compbio.ludwig.org.br/~natanja/TAE/GOstat.

4.5 VALIDAÇÃO EXPERIMENTAL DO PADRÃO DE EXPRESSÃO DO EXON ASSOCIADO A TUMOR DO GENE *TUBD1* EM AMOSTRAS DE GLIOBLASTOMA

O gene *tubd*1 está envolvido com a manutenção do citoesqueleto e com a mobilidade celular (DUTCHER 2003). Drogas que interferem com a dinâmica das microtubulinas e causando a agregação das microtubulinas ou a dissociação das mesmas já estão sendo utilizadas no tratamento contra o câncer. Acredita-se que eventos associados ao ligamento das drogas às tubulinas sejam críticos para a o desencadeamento da apoptose na célula maligna (PELLEGRINI e BUDMAN 2005). Portanto uma variante associada a tumor de gene *tubd*1 teria um impacto terapêutico interessante.

O laboratório do Dr. Greg Riggins da escola de Medicina da Universidade Johns Hopkins em Baltimore, EUA disponibilizou amostras do banco de cDNAs existente neste laboratório. Além de cDNA de amostras de pacientes de tumor de cérebro (glioblastoma) de diferentes graus, este banco contém cDNA de linhagens celulares de astrócitos com mutações nos genes *Ras* e *hTERT* que induzem a transformação destas células. Uma vez que publicamos validações experimentais da associação do gene *tubd*1 a câncer de cérebro, utilizamos estes amostras para a ampliação da validação deste gene.

A figura 11A mostra que o exon associado a tumor está super-expresso em várias amostras de glioblastoma de todos os graus. Entretanto, apesar do exon não ser super-expresso em amostras normais de cérebro, o exon é super-expresso em linhagens normais de astrócitos (ver Figura 11B).



Legenda - A. Expressão das variantes do *TUBD1* em diferentes amostras de pacientes. NB – Tecido normal de cérebro, T1–T10: amostras de glioblastoma sendo T1-T3 grau 2 (Astrocytoma), T4-T6 grau 3 (Anaplastic Astrocytoma) e T7-T10 grau 4 (Glioblastoma multiforme). B. Expressão das variantes do *TUBD1* em linhagens de astrócitos. NB – Tecido normal de cérebro, AN se refere à linhagem celular de astrócitos normal e AM à linhagem de astrócitos mutada e transformada. UT e RNA universal comprado (Stratagene) de vários tumores, como controle positivo. T11 - amostra de glioblastoma grau 4 (Glioblastoma multiforme). C. Amplificação do gene *GAPDH* como controle positivo em todas as amostras. "N" refere ao controle sem DNA.

Figura 11 - Validação por *RT-PCR* da expressão do exon associado a tumor do gene *tubd1* em amostras de pacientes de glioblastoma e em linhagens celulares de astrócitos normais e transformados por mutação.

DISCUSSÃO

5 DISCUSSÃO

Neste trabalho combinamos uma análise computacional e validações experimentais com o objetivo de buscar exons associados a tumor em tecidos específicos oriundos de genes não super-expressos em tumor.

Uma vez que mutações e eventos epigenéticos podem ser específicos de um tipo de tumor, baseamos nossos critérios de seleção na idéia de que variantes de *splicing* podem ser associadas a tumor de um tecido específico e não necessariamente em todos os tipos de tumores. Estes critérios e a análise estatística resultaram em 1295 genes candidatos, contendo 2878 exons candidatos associados a tumor em pelo menos um tecido específico.

A validação experimental de alguns exons candidatos demonstrou que nossa lista inicial continha exons pertencendo a genes que independentemente da forma de *splicing* são super-expressos em tumor. Uma vez que nosso interesse era a busca de candidatos que representam casos de *splicing* alternativo regulado diferencialmente em tumores, implementamos um filtro adicional que permitiu a exclusão de genes super-expressos em tumores. Este filtro baseou-se na exclusão de genes candidatos que apresentavam um número significativamente maior de *SAGE tags* em bibliotecas de tumor em comparação às *tags* de bibliotecas de tecido normal. Após esta análise, o número final de candidatos foi reduzido para 638 genes (contendo 1386 exons).

A utilização de bibliotecas normalizadas

Como mencionado na seção "Material e Métodos", utilizamos seqüências de cDNA tanto de bibliotecas normalizadas quanto de bibliotecas não-normalizadas. Uma vez que a intenção da geração de bibliotecas normalizadas é a representação não redundante de todos os transcritos (de alta e de baixa expressão) em uma determinada célula, elas não permitem a análise do padrão de expressão relativo de cada um dos transcritos (BONALDO et al. 1996). Testamos se a presença de seqüências de bibliotecas normalizadas na nossa análise causou um viés em nosso resultado. Refazendo a nossa busca utilizando apenas bibliotecas não-normalizadas conseguimos um número proporcional ao número de candidatos no nosso resultado inicial utilizando todas bibliotecas (encontramos 889 candidatos, que representam 30% dos candidatos obtidos pela análise estatística inicial, utilizando dados de *ESTs* que constituem 30% dos dados iniciais). Assim, o resultado da nossa análise não foi influenciado por artefatos causados pela inclusão das bibliotecas normalizadas.

A validação experimental

Nos testes estatísticos utilizados foi definida uma margem de 5% de inclusão de candidatos falso-positivos. Isto implica que existe uma possibilidade de que 69 dos 1386 exons identificados não estejam realmente associados a tumor. Uma vez que a taxa real de candidatos falso-positivos pode ser avaliada apenas experimentalmente, analisamos o padrão de expressão de alguns dos nossos candidatos utilizando a técnica de RT-PCR. Utilizando linhagens celulares para a validação de dez exons candidatos da nossa lista final, 4 dos 10 genes demonstraram o padrão de expressão esperado. Quando utilizamos um painel de amostras de

pacientes, conseguimos validar positivamente 5 de 6 candidatos testados. Estes resultados demonstram que a utilização da validação experimental é essencial e indicam que existem variáveis adicionais que podem aumentar o número de candidatos falso positivos além das margens da análise estatística. Ambas as limitações da técnica de *RT-PCR* (a serem discutidas em seguida) e o número relativamente baixo de bibliotecas de *SAGE* e *ESTs* disponíveis são algumas destas variáveis. Por exemplo, o seqüenciamento de uma quantidade maior de *ESTs* poderia revelar um padrão de expressão diferente para algumas variantes de *splicing* ou a disponibilidade de um número maior de bibliotecas de *SAGE* poderia demonstrar que um gene candidato é super-expresso em tumor.

Além disso, a heterogeneidade das amostras de tumor e das linhagens celulares adiciona uma outra variável ao sistema. Embora a utilização de linhagens celulares permita a replicação de experimentos e a análise de condições experimentais diferentes, elas não representam o contexto biológico tecidual exato de amostras tumorais de pacientes. Linhagens celulares também podem ter incorporado alterações no genoma comparado a amostras de pacientes. Por outro lado, amostras de pacientes podem estar contaminadas por diferentes tipos de células causando diferenças individuais entre amostras e este fato torna a reprodução dos experimentos mais complicada (BRENTANI et al. 2005).

A validação experimental por *RT-PCR* é uma das maneiras mais específicas para a investigação do padrão de expressão de mRNA. No entanto, a utilização de *primers* que permitam a amplificação das duas variantes de *splicing* simultaneamente não nos permite avaliar o nível de expressão das mesmas devido à ocorrência de competição entre as variantes durante a amplificação. A variante menor (na qual o

exon candidato seria excluído) teria uma vantagem durante a reação de amplificação devido ao seu tamanho menor. A utilização de *primers* específicos para cada uma das variantes poderia resolver este problema. Certamente é necessária uma validação experimental mais ampla dos candidatos com a técnica de PCR em tempo real, que permite uma determinação mais precisa da expressão diferencial (VANDENBROUCKE et al. 2001). Também existem outras técnicas sensíveis para validar o nível de expressão de variantes de *splicing* alternativo, como *single molecule profiling* (ZHU et al. 2003).

Análise de *SAGE* para excluir genes super-expressos em tumor

Como mencionado anteriormente, quatro outros estudos buscando variantes de *splicing* associadas a tumor foram publicados durante a realização do presente trabalho. Nenhum destes considera a validação experimental da expressão dos genes candidatos independentemente das variantes para a verificação da associação a tumor dos mesmos. Em um dos dois trabalhos nos quais foi implementada uma validação experimental dos candidatos foram utilizados *primers* que não permitiam a verificação da expressão do protótipo (WANG et al. 2003b). No segundo estudo, no qual foram utilizados tal *primers*, um dos genes candidatos demonstrou super-expressão no tecido tumoral (HUI et al. 2004). A nossa implementação da análise de *SAGE* permitiu a diminuição destes genes super expressos em tumores na nossa lista de candidatos.

É importante enfatizar que para a análise de *SAGE* era necessário assumir que a *tag* de *SAGE* mais próxima à extremidade 3' do gene candidato era representativa do transcrito mais abundante deste gene e que, portanto, a contagem desta *tag* refletia

corretamente o padrão de expressão do gene inteiro. Existe a possibilidade que esta *tag* represente, na realidade, uma variante de *splicing* que não é a mais abundante e que é expressa diferentemente das outras variantes do gene. Desta maneira a contagem desta *tag* não representaria o padrão de expressão do gene inteiro corretamente. Além disso, quando um exon candidato a ser associado a tumor pertence à variante mais abundante de um gene, a análise de *SAGE* excluiria o gene da lista de candidatos por ser super-expresso em tumor.

Validação em larga escala por *microarray*

Neste estudo demonstramos que a combinação da análise computacional juntamente com a validação experimental permitiu uma taxa de sucesso mais alta na busca de exons associados a tumor do que conseguiríamos com apenas uma das abordagens. Porém, apenas validações de larga escala como a de *microarray,* podem determinar a taxa de sucesso verdadeira da nossa análise. Do ponto de vista funcional, com a técnica de *micorarray* é possível medir as quantidades relativas de formas distintas de *splicing* em uma grande quantidade de diferentes tecidos e desta forma se determinar padrão de *splicing* e sua regulação nestes diferentes tecidos (LEE e ROY 2004). Para medir em um *array* o padrão de expressão de todas as variantes de *splicing* dos genes candidatos é necessária a utilização de fragmentos de todas as regiões destes genes analisados.

Estudos que comparam análises de *microarray* com predições computacionais de *splicing* alternativo específico a tecidos demonstraram a necessidade (PAN et al. 2004) e a compatibilidade (YEO et al. 2004) da combinação dos dois. LE et al. (2004) utilizaram uma abordagem para analisar *microarrays* que

permite diferenciar entre sinais considerados "ruído" e sinais que indicam a existência de diferentes isoformas diferencialmente expressas do mesmo gene (LE et al. 2004). Existem muitos estudos utilizando a técnica de *microarray* para estudar o padrão de *splicing* geral em diferentes tecidos e também *arrays* que estudam muitas isoformas de um único gene (LEE e ROY 2004).

Os exons candidatos descritos neste trabalho estão sendo utilizados para a construção de um *microarray* no Laboratório de Expressão Gênica do Instituto Ludwig (BRENTANI et al. 2005). Este *microarray* contém um subgrupo dos exons candidatos a serem associados a tumor obtidos no presente trabalho. Foram desenvolvidos novos critérios para a seleção de exons a serem imobilizados no *microarray*. Estes critérios incluíram, por exemplo, a utilização de exons com tamanho de pelo menos 100 bp, exons com conteúdo de GC entre 50 e 60%, e a exclusão de exons que contêm seqüências repetitivas para otimizar as condições de hibridização e para favorecer a hibridização de seqüências homólogas do genoma a um único fragmento presente no *microarray* (BRENTANI et al. 2003). Os exons candidatos foram alinhados contra o banco de dados *dbEST* para excluir os exons que alinharam com mais de um *cluster* de *ESTs* (BRENTANI et al. 2005). Também foi aplicado o programa *Primer3* para desenhar automaticamente *primers* de alta qualidade para a amplificação de fragmentos correspondentes aos 270 exons selecionados. Até o presente momento todas os fragmentos selecionados deste projeto foram amplificados e seqüenciados e estão sendo realizadas hibridizações com RNA de diferentes tipos de tumores e de tecidos normais.

O *microarray* dos exons obtidos neste trabalho permite a descoberta de exons como novos marcadores de tumor em uma quantidade grande de diferentes tumores.

Posteriormente, a validação experimental individual dos marcadores tumorais será feita para a análise de *splicing* associado a tumor dos mesmos. Desta maneira são necessárias menos fragmentos para imobilização, uma vez que para a validação de exons de *splicing* associados a tumor pertencendo a genes que não são super-expressos em tumor em uma lamina, seria necessário imobilizar simultaneamente fragmentos dos exons candidatos e fragmentos dos exon flanqueadores ou as junções exon/exon de todas as variantes dos genes candidatos. Quando a validação experimental individual demonstrar que um exon pertence a um gene que é associado a tumor em geral, independente das variantes de *splicing*, este ainda poderá ser investigado mais amplamente para uso como marcador tumoral.

O uso de um padrão de expressão diferencial como método diagnóstico requer um resultado muito específico que dificilmente será alcançado com a análise de um gene isoladamente. Acreditamos que a utilização dos métodos experimentais mencionados anteriormente e um painel composto por vários exons marcadores de tumor será mais adequada para diagnosticar corretamente a existência de certos tumores.

A possível função de variantes de *splicing* associadas a tumor no processo tumorigênico

A análise detalhada das variantes associadas a tumor também pode ser crucial para o melhor entendimento do processo de tumorigênese. Em muitos exemplos de *splicing* associado a tumor a função da forma alternativa de um gene é consistente com um papel possível no câncer (VENABLES 2004). Nossa análise da ontologia sugere que genes envolvidos no transporte intracelular de proteínas, crescimento e

manutenção celular, estejam super-representados na lista final de candidatos. Uma mudança no padrão de expressão de variantes de genes relacionados ao ciclo e manutenção celular certamente poderiam ter um efeito na transformação celular.

Algumas das variantes podem ser relacionadas funcionalmente a processos que são conseqüências do processo tumorigênico e não necessariamente a causa do mesmo. Por exemplo, neste trabalho demonstramos a associação da expressão de uma variante do gene *tubd1* com glioblastoma. Apesar desta associação, a variante estudada também demonstrou uma super-expressão em linhagens celulares normais de astrócitos. Portanto, a associação observada existe em condições de proliferação celular aumentada, que é característica tanto de tumores quanto de linhagens celulares em geral, e não exclusiva à tumorigênese.

No entanto, também como conseqüência do tumor, alterações no padrão de *splicing* podem ter implicações biológicas importantes. Por exemplo, o processo de *splicing* está relacionado com o processo de *nonsense-mediated mRMA decay* (*NMD*) (LEJEUNE et al. 2005). Este é um processo no qual transcritos de mRNA são degradados quando um códon de terminação está localizado a mais de 50 nucleotídeos *upstream* de uma junção exon-exon. Quando o padrão de *splicing* é alterado por causa de processos tumorigênicos este é apenas um dos processos afetados e pode levar a degradação de transcritos que talvez não devessem ser degradados.

Regulação de *splicing* alternativo associado a tumor

Acredita-se que, *in vivo*, a regulação da expressão de diferentes fatores de *splicing* representa uma forma de regulação do *splicing* alternativo (HANAMURA et

al. 1998). Existem diferentes estudos que têm manipulado a quantidade de fatores de *splicing* para verificar seu efeito na expressão de variantes de *splicing* de genes conhecidos. Através do uso de RNAi, 300 genes de ligação ao RNA de *Drosophila* foram silenciados (total ou parcialmente) e foi demonstrado que a mudança na expressão de de 47 destes genes tem um efeito na regulação do padrão de *splicing* (PARK et al. 2004). Além disso, este estudo revelou que o *splicing* em regiões diferentes dentro de um único transcrito pode ser regulado por grupos diferentes de proteínas. Como o efeito das proteínas foi avaliado em variantes de apenas três genes, provavelmente poderiam ser identificadas outras proteínas com papel na regulação de *splicing* se outros genes também fossem verificados (PARK et al. 2004). O grupo de BLANCHETTE também utilizou a combinação de RNAi e *microarray* para investigar o efeito da redução de fatores de *splicing* específicos (dois do grupo de proteínas SR e dois do grupo de proteínas hnRNP) no padrão de *splicing* através de uma lâmina contendo todas as junções exon/exon de *Drosophila.* Este estudo demonstrou a mesma tendência que foi demonstrada pelo grupo anterior: o silenciamento de cada fator de *splicing* através de RNAi apresentou um efeito diferente. Apenas em poucos casos a redução da expressão de dois fatores diferentes do mesmo grupo (proteínas SR ou proteínas hnRNP) tinha a mesma influência na expressão das variantes (BLANCHETTE et al. 2005).

Um outro estudo, utilizando uma plataforma de *microarray*, analisou expressão de 100 variantes de *splicing* e de alguns fatores de *splicing* em tecido normal e em células de limfoma Hodgkin. Este estudou demonstrou que existem mudanças no padrão de expressão das isoformas analisadas nas células transformadas, no entanto não foi encontrada uma correlação entre estas mudanças e

o padrão de expressão dos fatores de *splicing* investigados (RELOGIO et al. 2005). Além disso, utilizando diferentes bancos de dados de expressão gênica nosso grupo demonstrou que alguns fatores de *splicing* são super-expressos em tumor de cérebro, mama e cólon (KIRSCHBAUM-SLAGER et al. 2004).

Por fim, um grupo investigou a taxa de *splicing* alternativo em diferentes tecidos e analisou a expressão de alguns fatores de *splicing* nos mesmos tecidos utilizando dados de *microarray* (YEO et al. 2004). Este grupo demonstrou que as taxas de *splicing* alternativo mais altas foram encontradas em cérebro, pâncreas e fígado (YEO et al. 2004). Além disso, também foi demonstrado que a expressão das proteínas SR e das proteínas hnRNP em fígado era muito diferente do que a expressão nos outros tecidos, implicando uma possível relação entre a taxa de *splicing* alternativo e a expressão diferencial dos fatores de *splicing*.

É difícil prever as conseqüências finais de alterações nos padrões de expressão de um número grande de fatores de *splicing* (MATLIN et al. 2005). Os níveis relativos de diferentes fatores entre si determinarão o efeito final no padrão de *splicing* (HANAMURA et al. 1998).

Para provar uma verdadeira relação causal entre o padrão de expressão diferencial e a taxa de *splicing* alternativo utilizando ferramentas computacionais são necessárias análises adicionais com bancos de dados de expressão gênica atualizados, a análise da concentração final de cada uma das proteínas fatores de *splicing* na célula e uma análise do padrão de *splicing* de todas as variantes de *splicing* do *transcriptoma* em tecidos diferentes. Uma análise destas demonstraria se a super-expressão dos fatores de *splicing* causa um aumento na taxa de *splicing* alternativo, permitindo a expressão aumentada de transcritos novos ou se ela aumenta a

expressão de transcritos processados constitutivamente. No entanto, é importante notar que padrões de expressão de isoformas de mRNA não representarão todos os eventos de *splicing* alternativo, uma vez que certas moléculas de RNA podem ser transportadas para fora do núcleo ou degradadas rapidamente por terem uma estabilidade baixa ou pelo processo de NMD, portanto não seriam considerados todos os eventos possíveis causados por todos os fatores de *splicing* (MATLIN et al. 2005).

Futuras investigações

Muitas questões ainda precisam ser abordadas. Qual fração das variantes associadas a tumor resulta de 'erros' no processamento de pré-mRNA e qual fração é funcional? Como o processo de *splicing* é regulado no momento de proliferação celular? O aumento na expressão gênica geral em câncer, necessitando um processamento aumentado de pré-mRNA, é responsável por uma mudança na regulação de *splicing*? Quais processos celulares têm um efeito no padrão de *splicing* alternativo? Como o *splicing* é regulado por estímulos externos e caminhos de transdução de sinais?

Futuras investigações devem abordar as seguintes questões:

- Quais são todos os exons associados a tumor da nossa lista? A abordagem de *microarray* ajudará definir todos os candidatos utilizando amostras de linhagens celulares e de pacientes juntamente com a ampliação da validação por *Real Time PCR*.

- Quais são os exons que demonstram uma diferença de expressão mais alta entre um tumor específico e tecido normal? Estes exons podem ser juntados em um painel para um eventual uso como ferramenta de diagnóstico.

- Quais são todos os eventos de *splicing* associados a tumor? Neste trabalho investigamos apenas o evento de inclusão de exons, no entanto as outras formas de *splicing* podem ser associadas a tumor da mesma forma.

- Qual é a função dos exons positivamente validados? Ensaios funcionais demonstrarão qual será o efeito do bloqueio ou da ativação da ação destes exons. Estes ensaios serão importantes para a utilização dos exons de ponto de visto terapêutico e para o entendimento do processo de tumorigênese.

- Qual é o processo que causa a expressão diferencial de variantes de *splicing* em tumor? As variantes de *splicing* são conseqüências de erros no processamento de pré-mRNA ou resultam de um processo regulado? Estes questões podem ser abordadas analisando o comportamento dos fatores envolvidos no processamento de pré-mRNA e no spliceossomo. Métodos funcionais como o bloqueio de expressão por RNAi ou manipulações genéticas ajudarão a analisar estes questões funcionais.

Neste trabalho demonstramos a utilização de uma combinação de ferramentas computacionais com validações experimentais para a seleção de exons pertencendo a de variantes super-expressas em tumores excluindo genes super-expressos em tumor. Variantes associadas a tumor podem fornecer informação importante para futuras investigações sobre a regulação de *splicing* associado alternativo tumor. Análises experimentais mais amplas de larga escala validarão até que nível nossos exons candidatos diferencialmente expressos possuem um potencial diagnóstico e/ou terapêutico.

REFERÊNCIAS BIBLIOGRÁFICAS

6 REFRÊNCIAS BIBLIOGRÁFICAS

Adams MD, Celniker SE, Holt RA, et al. The genome sequence of Drosophila melanogaster. **Science** 2000; 287:2185-95.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. **Molecular biology of the cell**. 4th ed. New York: Garland Publishing; 2002. How Cells Read the Genome: From DNA to Protein; p.299-375.

Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res** 1997; 25:3389-402.

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. **Nature** 2000; 408:796-815.

Baudry D, Hamelin M, Cabanis MO, et al. WT1 splicing alterations in Wilms' tumors. **Clin Cancer Res** 2000; 6:3957-65.

Beghini A, Ripamonti CB, Peterlongo P, et al. RNA hyperediting and alternative splicing of hematopoietic cell phosphatase (PTPN6) gene in acute myeloid leukemia. **Hum Mol Genet** 2000; 9:2297-304.

Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. **Bioinformatics** 2004; 20:1464-5.

Berget SM. Exon recognition in vertebrate splicing. **J Biol Chem** 1995; 270:2411-4.

Black DL. Splicing in the inner ear: a familiar tune, but what are the instruments? **Neuron** 1998; 20:165-8.

Black DL. Mechanisms of alternative pre-messenger RNA splicing. **Annu Rev Biochem** 2003; 72:291-336.

Blanchette M, Green RE, Brenner SE, Rio DC. Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. **Genes Dev** 2005; 19:1306-14.

Boggs RT, Gregor P, Idriss S, Belote JM, McKeown M. Regulation of Sexual-Differentiation in Drosophila-Melanogaster Via Alternative Splicing of Rna from the Transformer Gene. **Cell** 1987; 50:739-47.

Bonaldo MF, Lennon G, Soares MB. Normalization and subtraction: two approaches to facilitate gene discovery. **Genome Res** 1996; 6:791-806.

Boon K, Osorio EC, Greenhut SF, et al. An anatomy of normal and malignant gene expression. **Proc Natl Acad Sci U S A** 2002; 99:11287-92.

Brenner S, Johnson M, Bridgham J, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. **Nat Biotechnol** 2000; 18:630-4.

Brentani H, Caballero OL, Camargo AA, et al. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. **Proc Natl Acad Sci U S A** 2003; 100:13418-23.

Brentani RR, Carraro DM, Verjovski-Almeida S, et al. Gene expression arrays in cancer research: methods and applications. **Crit Rev Oncol Hematol** 2005; 54:95-105.

Brett D, Hanke J, Lehmann G, et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. **Febs Letters** 2000; 474:83-6.

Brinkman BM. Splice variants as cancer biomarkers. **Clin Biochem** 2004; 37:584-94.

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

C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. **Science** 1998; 282:2012-8.

Chabot B, Frappier D, La Branch. Differential ASF/SF2 activity in extracts from normal WI38 and transformed WI38VA13 cells. **Nucleic Acids Res** 1992; 20:5197-204.

Classon M, Harlow E. The retinoblastoma tumour suppressor in development and cancer. **Nat Rev Cancer** 2002; 2:910-7.

Cragg MS, Chan HTC, Fox MD, et al. The alternative transcript of CD79b is overexpressed in B-CLL and inhibits signaling for apoptosis. **Blood** 2002; 100:3068-76.

Croft L, Schandorff S, Clark F, Burrage K, Arctander P, Mattick JS. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. **Nat Genetics** 2000; 24:340-1.

Dias Neto E, Correa, RG, Verjovski-Almeida, S et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. **Proc Natl Acad Sci U S A** 2000; 97:3491-6.

Diehn M, Sherlock G, Binkley G, et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. **Nucleic Acids Res** 2003; 31:219-23.

Ding WQ, Cheng ZJ, McElhiney J, Kuntz SM, Miller LJ. Silencing of secretin receptor function by dimerization with a misspliced variant secretin receptor in ductal pancreatic adenocarcinoma. **Cancer Res** 2002; 62:5223-9.

Dutcher SK. Long-lost relatives reappear: identification of new members of the tubulin superfamily. **Curr Opin Microbiol** 2003; 6:634-40.

Galante PA, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ. Detection and evaluation of intron retention events in the human transcriptome. **RNA** 2004; 10:757-65.

Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. **Nat Biotechnol** 2004; 22:535-46.

Ge K, DuHadaway J, Du W, Herlyn M, Rodeck U, Prendergast GC. Mechanism for elimination of a tumor suppressor: aberrant splicing of a brain-specific exon causes loss of function of Bin1 in melanoma. **Proc Natl Acad Sci U S A** 1999; 96:9689-94.

Chirgwin JM, Przybyla AE, MacDonald RJ, Rutter WJ. Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. **Biochemistry** 1979; 18:5294-9.

Gilbert W. Why genes in pieces? **Nature** 1978; 271:501.

Hanahan D, Weinberg RA. The hallmarks of cancer. **Cell** 2000; 100:57-70.

Hanamura A, Caceres JF, Mayeda A, Franza BR, Jr., Krainer AR. Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. **RNA** 1998; 4:430-44.

Hayes GM, Dougherty ST, Davis PD, Dougherty GJ. Molecular mechanisms regulating the tumor-targeting potential of splice-activated gene expression. **Cancer Gene Ther** 2004; 11:797-807.

Hide WA, Babenko VN, van Heusden PA, Seoighe C, Kelso JF. The contribution of exon-skipping events on chromosome 22 to protein coding diversity. **Genome Res** 2001; 11:1848-53.

Hoffman JD, Hallam SE, Venne VL, Lyon E, Ward K. Implications of a novel cryptic splice site in the BRCA1 gene. **Am J Med Genet** 1998; 80:140-4.

Holmila R, Fouquet C, Cadranel J, Zalcman G, Soussi T. Splice mutations in the p53 gene: case report and review of the literature. **Hum Mutat** 2003; 21:101-2.

Hui L, Zhang X, Wu X, et al. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. **Oncogene** 2004; 23:3013-23.

Ibrahim EC, Schaal TD, Hertel KJ, Reed R, Maniatis T. Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. **Proc Natl Acad Sci U S A** 2005; 102:5002-7.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. **Nature** 2004; 431:931-45.

Johnson JM, Castle J, Garrett-Engele P, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. **Science** 2003; 302:2141-4.

Jongeneel CV, Iseli C, Stevenson BJ, et al. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. **Proc Natl Acad Sci U S A** 2003; 100:4702-5.

Jurica MS, Moore MJ. Pre-mRNA splicing: awash in a sea of proteins. **Mol Cell** 2003; 12:5-14.

Kan Z, States D, Gish W. Selecting for functional alternative splices in ESTs. **Genome Res** 2002; 12:1837-45.

Kirschbaum-Slager N, Lopes GM, Galante PA, Riggins GJ, de Souza SJ. Splicing factors are differentially expressed in tumors. **Genet Mol Res** 2004; 3:512-20.

Kirschbaum-Slager N, Parmigiani RB, Camargo AA, de Souza SJ. Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. **Physiol Genomics** 2005; 21:423-32.

Kornblihtt AR, de la MM, Fededa JP, Munoz MJ, Nogues G. Multiple links between transcription and splicing. **RNA** 2004; 10:1489-98.

Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. **Nature** 2001; 409:860-921.

Le K, Mitsouras K, Roy M, et al. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. **Nucl Acids Res** 2004; 32:e180.

Lee C, Roy M. Analysis of alternative splicing with microarrays: successes and challenges. **Genome Biol** 2004; 5:231.

Lejeune F, Maquat LE. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. **Curr Opin Cell Biol** 2005; 17:309-15.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene index analysis of the human genome estimates approximately 120,000 genes. **Nat Genet** 2000; 25:239-40.

Lim LP, Burge CB. A computational analysis of sequence features involved in recognition of short introns. **PNAS** 2001; 98:11193-8.

Maeda T, Furukawa S. Transformation-associated changes in gene expression of alternative splicing regulatory factors in mouse fibroblast cells. **Oncol Rep** 2001; 8:563-6.

Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. **Nat Rev Mol Cell Biol** 2005; 6:386-98.

Mercatante DR, Mohler JL, Kole R. Cellular response to an antisense-mediated shift of Bcl-x pre-mRNA splicing and antineoplastic agents. **J Biol Chem** 2002; 277:49374-82.

Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. **Genome Res** 1999; 9:1288-93.

Misteli T, Spector DL. RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo. **Mol Cell** 1999; 3:697-705.

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

Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. **Nucleic Acids Res** 2001; 29:2850-9.

Navaglia F, Fogar P, Greco E, et al. CD44v10: an antimetastatic membrane glycoprotein for pancreatic cancer. **Int J Biol Markers** 2003; 18:130-8.

Nilsen TW. The spliceosome: the most complex macromolecular machine in the cell? **Bioessays** 2003; 25:1147-9.

Osorio EC, de Souza JE, Zaiats AC, de Oliveira PS, de Souza SJ. pp-Blast: a "pseudo-parallel" Blast. **Braz J Med Biol Res** 2003; 36:463-4.

Pan Q, Shai O, Misquitta C, et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. **Mol Cell** 2004; 16:929-41.

Park JW, Parisky K, Celotto AM, Reenan RA, Graveley BR. Identification of alternative splicing regulators by RNA interference in Drosophila. **Proc Natl Acad Sci U S A** 2004; 101:15974-9.

Pellegrini F, Budman DR. Review: tubulin function, action of antitubulin drugs, and new drug development. **Cancer Invest** 2005; 23:264-73.

Quackenbush J, Cho J, Lee D, et al. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. **Nucleic Acids Res** 2001; 29:159-64.

Relogio A, Ben Dov C, Baum M, et al. Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. **J Biol Chem** 2005; 280:4779-84.

Ross JS, Fletcher JA, Linette GP, et al. The HER-2/neu Gene and Protein in Breast Cancer 2003: Biomarker and Target of Therapy. **Oncologist** 2003; 8:307-25.

Sakabe NJ, de Souza JES, Galante PAF, et al. ORESTES are enriched in rare exon usage variants affecting the encoded proteins. **Comptes Rendus Biologies** 2003; 326:979-85.

Sambrook J. Adenovirus amazes at Cold Spring Harbor. **Nature** 1977; 268:101-4.

Schmucker D, Clemens JC, Shu H, et al. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. **Cell** 2000; 101:671-84.

Schuler GD, Boguski MS, Stewart EA, et al. A gene map of the human genome. **Science** 1996; 274:540-6.

Schwerk C, Schulze-Osthoff K. Regulation of Apoptosis by Alternative Pre-mRNA Splicing. **Mol Cell** 2005; 19:1-13.

Sigal A, Rotter V. Oncogenic Mutations of the p53 tumor suppressor: the demons of the Guardian of the Genome. **Cancer Res** 2000; 60:6788-93.

Silva AP, de Souza JE, Galante PA, Riggins GJ, de Souza SJ, Camargo AA. The impact of SNPs on the interpretation of SAGE and MPSS experimental data. **Nucleic Acids Res** 2004; 32:6104-10.

Smith CW, Valcarcel J. Alternative pre-mRNA splicing: the logic of combinatorial control. **Trends Biochem Sci** 2000; 25:381-8.

Sneath RJ, Mangham DC. The normal structure and function of CD44 and its role in neoplasia. **Mol Pathol** 1998; 51:191-200.

Sogayar MC, Camargo AA, Bettoni F, et al. A transcript finishing initiative for closing gaps in the human transcriptome. **Genome Res** 2004; 14:1413-23.

Stickeler E, Kittrell F, Medina D, Berget SM. Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis. **Oncogene** 1999; 18:3574-82.

Vandenbroucke II, Vandesompele J, Paepe AD, Messiaen L. Quantification of splice variants using real-time PCR. **Nucleic Acids Res** 2001; 29:E68.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. **Science** 1995; 270:484-7.

Venables JP. Aberrant and alternative splicing in cancer. **Cancer Res** 2004; 64:7647-54.

Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. **Science** 2001; 291:1304-51.

Wang L, Duke L, Zhang PS, et al. Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. **Cancer Res** 2003a; 63:4724-30.

Wang Z, Lo HS, Yang H, et al. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. **Cancer Res** 2003b; 63:655-7.

Xie H, Zhu WY, Wasserman A, Grebinskiy V, Olson A, Mintz L. Computational analysis of alternative splicing using EST tissue information. **Genomics** 2002; 80:326-30.

Xu Q, Lee C. Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. **Nucleic Acids Res** 2003; 31:5635-43.

Yakushijin Y, Steckel J, Kharbanda S, et al. A directly spliced exon 10-containing CD44 variant promotes the metastasis and homotypic aggregation of aggressive non-Hodgkin's lymphoma. **Blood** 1998; 91:4282-91.

Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. **Genome Biol** 2004; 5:R74.

Zerbe LK, Pino I, Pio R, et al. Relative amounts of antagonistic splicing factors, hnRNP A1 and ASF/SF2, change during neoplastic lung growth: implications for pre-mRNA processing. **Mol Carcinog** 2004; 41:187-96.

Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. **BMC Bioinformatics** 2004; 5:16.

Zhou Z, Licklider LJ, Gygi SP, Reed R. Comprehensive proteomic analysis of the human spliceosome. **Nature** 2002; 419:182-5.

Zhu J, Shendure J, Mitra RD, Church GM. Single molecule profiling of alternative pre-mRNA splicing. **Science** 2003; 301:836-8.

NATANJA SARA KIRSCHBAUM - SLAGER

Curriculum vitae

Informações pessoais

E-mail: natanja@compbio.ludwig.org.br
natanja@gmail.com

Data de nascimento: 06/04/1975
País de nascimento: Holanda

Formação acadêmica

2002-presente **Doutorado em Oncologia**
Fundação Antônio Prudente, São Paulo, Brasil. Laboratório de Biologia Computacional, sob a orientação do Dr. Sandro J. de Souza e Dra. Anamaria A. Camargo. Tese: "Desenvolvimento de um sistema em larga escala para análise da associação entre variantes de splicing alternativo e o câncer e sua validação experimental".

1998-2000 **Mestrado em Neurobiologia**
School of Medicine of the Hebrew University, Hadassa Hospital, Jerusalem, Israel. Institute of Medical Sciences. Sob a orientação do Prof. M. Rosin e Prof. P. Lazarovici. Departamento de Farmacologia, Escola de Farmácia. Tese: "O efeito neuroprotetor dos inibidores da acetilcolinesterase e da oxidase monoamínica em células PC12 diferenciadas por NGF sob condições isquêmicas". Concluído com excelência.

1995-98 **Bacharelado em Ciências Médicas**
School of Medicine of the Hebrew University, Hadassa Hospital, Institute of Medical Science, Jerusalem, Israel.

1994-95 **Primeiro ano de bacharelado em Psicologia e Sociologia**
Hebrew University, Jerusalem, Israel.

1993-94 **Mechina Academit mada'e hateva:** Programa preparatório para alunos imigrantes na área das ciências exatas da Hebrew University, School for Overseas Students, Jerusalém, Israel.

Prêmios e bolsas

2003-presente Bolsa de doutorado da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

2003 Auxílio a participação em evento científico para curso em bioinformática do Laboratório Nacional de Computação Científica e do Centro Internacional para Engenharia Genética e Biotecnologia.

2002-2003 Bolsa de doutorado da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

1998 – 2000 Bolsa para alunos com excelente desempenho acadêmico da Faculdade de sciencas medicas, Hebrew University.

1993 – 1995 Bolsa para estudantes do exterior do ministério de imigração.

1993 Prêmio por excelência em Programa Preparatório para Ciências Exatas. Hebrew University, Jerusalem.

Atuação Profissional

2001-2002 **Compugen Ltd., empresa de bioinformática e Evogene Ltd, empresa 'start up' de bioteqnologia e genômica de plantas**
Operadora e administradora geral de laboratório de genômica de plantas.

1997 **Hadassah Hospital, Jerusalem, Depto de Medicina Nuclear**
Iniciação científica em neurofisiologia, laboratório do Dr. G. Goelman

- ***Habilidades computacionais:*** Sistema Linux, Perl, Java, C++, uso e administração de banco de dados MySQL, MS office, Pascal.

- ***Idiomas:*** Holandês (lingua nativa), Inglês (fluente), Hebráico (fluente), Português (fluente).

Apresentação de trabalhos em eventos científicos

1. **Natanja Kirschbaum-Slager**, Raphael Parmigiani, Graziela M. P. Lopes, Pedro A. F. Galante, Gregory J. Riggins, Anamaria A. Camargo, Sandro J. de Souza The use of expression data to study the association between alternative splicing and cancer. Trabalho apresentado como pôster no *Ninth Annual International Conference on Research in Computational Molecular Biology*, Camebridge, Maio 2005.

2. **Kirschbaum-Slager, N**; Lopes, G.M.P.; Galante, P.A.F; Riggins, G.J.; de Souza, S.J. Splicing factors are differentially expressed in tumors. Trabalho apresentado como pôster e escolhido para apresentação oral no *International Conference on Bioinformatics and Computational Biology*, Angra dos Reis, Brasil, Outubro, 2004.

3. **Natanja Kirschbaum-Slager**, Raphael Bessa Parmigiani, Helena Brentani, Anamaria A. Camargo, Sandro J. de Souza. A bioinformatics pipeline for the experimental validation of tumor specific alternative splicing variants. Trabalho apresentado como pôster no *International Conference on Bioinformatics and Computational Biology*, Riberão Preto, Brasil, Maio 2003.

4. **Kirschbaum-Slager N.**, Abu-Raya, S., Lazarovici, P. and Weinstock, M. (2000). TV3326, a cholinesterase and monoamine oxidase inhibitor, protects NGF-differentiated PC12 cells against oxygen-glucose deprivation. (abs). Neurosci. Lett. Suppl. 55, S29. Trabalho apresentado como pôster no *International meeting of the Israeli Society of Neuroscience*, Eilat, Israel, 2000.

Publicações

1. **Kirschbaum-Slager N**, Parmigiani RB, Camargo AA, de Souza SJ. Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. Physiol Genomics. 2005; 21(3):423-32.

2. **Kirschbaum-Slager N**, Lopes GM, Galante PA, Riggins GJ, de Souza SJ. Splicing factors are differentially expressed in tumors. Genet Mol Res. 2004; 30;3(4):512- 20.

3. Galante PA, Sakabe NJ, **Kirschbaum-Slager N**, de Souza SJ. Detection and evaluation of intron retention events in the human transcriptome. RNA. 2004 May;10(5):757-65.

4. Weinstock M, **Kirschbaum-Slager N**, Lazarovici P, Bejar C, Youdim MB, Shoham S. Neuroprotective effects of novel cholinesterase inhibitors derived from rasagiline as potential anti-Alzheimer drugs. Ann N Y Acad Sci. 2001 Jun;939:148-61.