Universidade de São Paulo

Instituto Ludwig de pesquisa sobre o câncer

Suzana Andreoli Marques Ezquina

Estudo da poliadenilação alternativa: Efeito dos polimorfismos nos elementos em *cis* no RNA e de transcritos em antissenso

São Paulo – SP - Brasil
2012

Suzana Andreoli Marques Ezquina

Estudo da poliadenilação alternativa: Efeito dos polimorfismos nos elementos em cis no RNA e de transcritos em antisense

Tese apresentada ao Programa de Pós-Graduação Interunidades em Bioinformática da Universidade de São Paulo, para a obtenção do título de Doutora em Bioinformática

Orientador: Prof. Dr. Sandro José de Souza

São Paulo – SP - Brasil
2012

RESUMO

Estudo da poliadenilação alternativa: Efeito dos polimorfismos nos elementos em *cis* no RNA e de transcritos em antissenso

A clivagem e a poliadenilação são processos essenciais na formação do mRNA, que têm a função de estabelecer a extremidade 3' e garantir a estabilidade do mRNA, sua localização no citoplasma e sua tradução. Os elementos presentes na região 3' não traduzida (3'UTR) dos genes participam na regulação da expressão gênica e da tradução.

A estabilidade dos mRNAs é alterada no câncer. Alterações genéticas observadas em diversos tipos de câncer são capazes de desregular o nível de expressão do mRNA mutado.

Mais de 7000 genes humanos tem sítios alternativos de poliadenilação e a escolha do sítio depende do tecido no qual o gene é expresso, da fase do ciclo celular ou de fatores externos que influenciam a regulação da expressão gênica.

Foram avaliadas as ocorrências relativas dos sinais de poliadenilação nos diversos tipos de transcritos e variantes. Os sinais canônicos AATAAAA e ATTAAA são os mais freqüentes, ocorrendo em 46% e 15% respectivamente, no caso dos transcritos de referência humanos, em todos os genes.

Dentre os genes com eventos de poliadenilação alternativa, a proporção dos sinais é bastante semelhante, sendo que 20% dos variantes curtos não possui sinal contra 11% dos variantes longos.

O estudo da poliadenilação alternativa deve levar em conta as regiões *downstream* aos sítios de clivagem, que têm papel importante na ligação da maquinaria de poliadenilação. Para tanto foi criado um score para estimar a força da ligação da proteína CstF à região *downstream* ao mRNA.

Podemos observar que dentre os genes com poliadenilação alternativa, os sinais canônicos são mais afetados por SNPs em variantes curtos que em variantes longos.

Neste trabalho analisamos a presença de transcritos mais curtos e mais longos no transcriptoma de células de cultura de câncer de mama HCC1954 e demonstramos também que é possível diferenciar os variantes de poliadenilação através de probes de microarray.

Palavras-chave:

- Transcriptoma
- Poliadenilação
- Expressão gênica
- 3'UTR
- SNPs
- Câncer
- Microarray

ABSTRACT

Alternative polyadenylation study: Effects of polymorphisms in RNA's *cis* elements and antisense transcripts

Cleavage and polyadenylation are essential processes involved in mRNA formation, establishing 3' end and assuring mRNA stability, its cytoplasmic location and translation. The sequence elements in 3'UTR play a crucial role in the regulation of gene expression and translation.

The stability of mRNAs is altered in cancer. Genetic alterations seen in several types of cancer can disrupt the regulation of expression levels of mutated mRNA.

More than 7000 human genes have alternative polyadenylation sites and site choice depends on the tissue in which it is expressed, cell cycle or exogenous factors that may influence the regulation of gene expression.

We evaluated the relative occurrence of polyadenylation signals in several types of transcripts and variants. The canonical signals AATAAA and ATTAAA are the most frequent, occurring in 46% and 15% respectively, in human reference transcripts, in all genes.

Among genes with alternative polyadenylation events, signal proportion is similar, nonetheless these is no signal in 20% of shorter variants against 11% of longer variants.

While studying alternative polyadenylation one should also take care about cleavage sites downstream regions, which have an important role in the assembly of the polyadenylation machinery. So forth we came up with a score in order to estimate the binding strength of CstF protein to the downstream region.

We have seen that among alternative polyadenylated genes, canonical signals are more affected by SNPs in shorter than in longer variants.

In this work we analyzed the existence of shorter and longer transcripts in a breast cancer cell culture transcriptome of HCC1954 and we also shown it is possible to differentiate polyadenylation variants through microarray probes.

Keywords:

- Transcriptome
- Polyadenylation
- Gene expression
- 3'UTR
- SNPs
- Cancer
- Microarray

Agradecimentos

Gostaria de agradecer à minha avó e aos meus pais por todo o apoio que me deram, durante todos esses anos. Tenho certeza que sem esse suporte, em todos os aspectos possíveis, a execução desta tese não teria sido possível. Obrigada por todo o carinho, amor, atenção, respeito, abrigo e patrocínio.

Agradeço ao Dr. Sandro por me receber de imediato como aluna em seu laboratório, por acreditar na minha capacidade e independência.

Gostaria de agradecer a Profa Dra Suely Marie e toda a equipe de seu laboratório, na faculdade de Medicina da USP, pelos dados de microarray da Codelink, e pela colaboração na interpretação dos dados e dos resultados.

Agradeço a meus colegas de laboratório, pelas diversas discussões científicas, apoio e ensinamentos. Agradeço pela paciência que todos tiveram ao início de meu doutorado, ao me ensinar inúmeras coisas sobre biologia, computação e sobre a própria pós-graduação.

À Capes pelo suporte financeiro.

Ao instituto Ludwig pelo acolhimento e suporte tecnológico.

Às secretárias Patrícia, Ana Cláudia, Renata e Eliane, pela eficiência e pela ajuda com as questões burocráticas e psicológicas.

À Profa. Dra. Marília Seelaender, minha orientadora de iniciação científica, que me despertou o interesse pela pesquisa, e por sua persistência e determinação, que tanto me inspiram.

Agradeço a Deus por todas as grandes oportunidades que a vida já me ofereceu.

Índice

Índice de Ilustrações

1 - Introdução

O processo de transcrição leva à formação de um pré-mRNA, que sofre uma série de modificações em sua seqüência para dar origem ao RNA mensageiro maduro. As modificações ocorrem de forma concomitante à transcrição, mas são estudadas como modificações pós-transcricionais. O pré-mRNA sofre *capping*, que é a adição de uma guanosina metilada em sua extremidade 5'. Em seguida sofre *splicing*, a retirada de introns e junção dos exons. Em sua extremidade 3', ocorre a clivagem e a poliadenilação (Neugebauer, 2002).

A clivagem e a poliadenilação são processos essenciais na formação do mRNA, que têm a função de estabelecer a extremidade 3' e garantir a estabilidade do mRNA, sua localização no citoplasma e sua tradução.

O processamento do pré-mRNA em sua extremidade 3' tem enorme importância funcional, sendo que quando esse processo não ocorre ou ocorre de forma errônea, a célula sofre conseqüências catastróficas em sua viabilidade e crescimento (Mandel C.R., 2007).

A cauda de poliadenina (poli(A)) é importante para o transporte do RNA mensageiro do núcleo para o citoplasma. Experimentos que substituíram o sítio de poliadenilação por um sítio presente em RNA ribossômico demonstraram que o RNA era clivado, mas não poliadenilado. Isso reduziu o transporte do mRNA para o citoplasma e diminuiu a expressão protéica (Huang & Carmichael, 1996).

O processamento 3' do pré-mRNA também é importante para a estabilidade do RNA, devido às proteínas de ligação à cauda poli(A). A ligação das PABP à cauda poli(A) previnem a degradação do RNA no citoplasma de células de mamíferos (Mandel C.R., 2007). No citoplasma, os RNAs são degradados a partir de suas extremidades 3' - que são mais instáveis - por exossomas recrutados por elementos da região 3´UTR dos mRNAs (van Hoof & Parker, 2002).

A presença da cauda poli(A) também é importante para a tradução eficiente do RNA mensageiro. Estudos em leveduras mostraram que a presença da cauda poli(A) é capaz de iniciar a tradução e que o 5' *cap* interage com a cauda poli(A) para promover a tradução de maneira mais eficiente (Mandel C.R., 2007).

A maquinaria de poliadenilação é acoplada à de *splicing* e de transcrição. Os fatores de poliadenilação interagem com os fatores de transcrição e com o domínio C terminal da RNA polimerase II (Pol II). Alterações nessas interações

levam à poliadenilação errônea e à degradação do mRNA. O sinal de poliadenilação é necessário para o término da transcrição (Mandel C.R., 2007).

A transcrição não termina num local definido ao final do gene; a RNA polimerase II continua a transcrever aproximadamente 1,5 kb após o sítio de poliadenilação. Um modelo propõe que a clivagem do RNA nascente no sítio de poliadenilação leva a uma mudança conformacional no complexo da RNA polimerase II, que induz à pausa e à liberação do complexo alongador do molde de DNA. A formação da extremidade 3' é parte do processo de término da transcrição para a maioria dos genes codificantes de proteínas (Pandit, Wang, & Xiang-Dong, 2008).

Os elementos presentes na região 3' não traduzida (3'UTR) dos genes participam na regulação da expressão gênica e da tradução (Kozak, 2004). Um dos modelos é a circularização do mRNA a partir da ligação simultânea de uma proteína na cauda poli(A) e no 5'cap (Prévôt, Darlix, & Ohlmann, 2003).

1.1 - Processo de poliadenilação

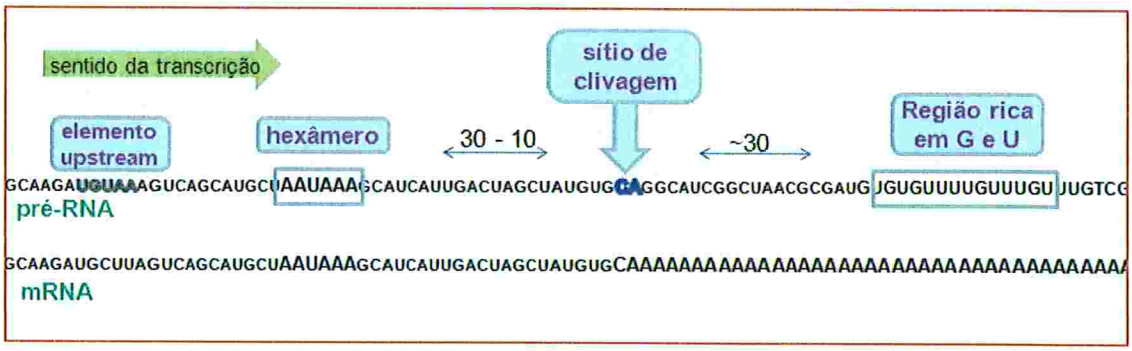


Figura 1: Elementos no RNA importantes para o processo de poliadenilação

Durante a transcrição, os fatores de poliadenilação permanecem ligados ao domínio carboxiterminal (CTD) da RNA polimerase II. Um desses fatores é o CPSF (*cleavage and polyadenylation specificity factor*), que reconhece o hexâmero sinal de poliadenilação. O CPSF se desliga do CTD da RNA polimerase II após a transcrição do sinal. A maior afinidade leva à ligação do CPSF ao hexâmero sinal.

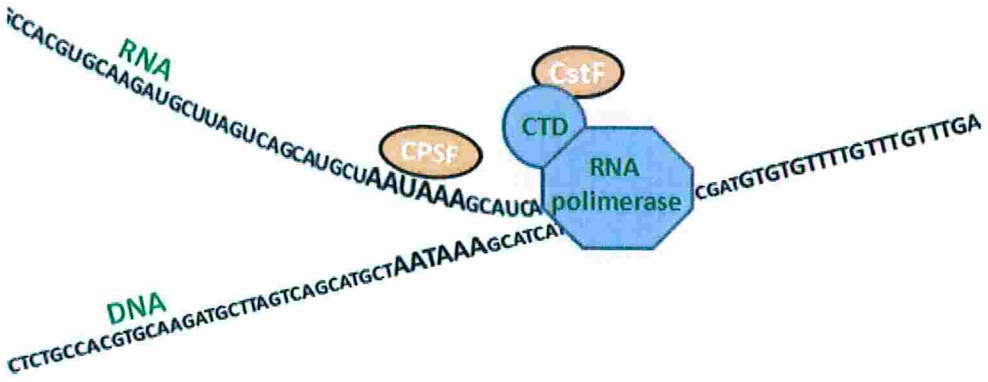


Figura 2: Ligação do CPSF mRNA durante a transcrição. A fita superior representa o mRNA e a inferior o DNA.

Em seguida o mesmo acontece com o CstF (Cleavage stimulation factor), que se desliga da RNA pol II e reconhece uma região *downstream*, rica em U e G.

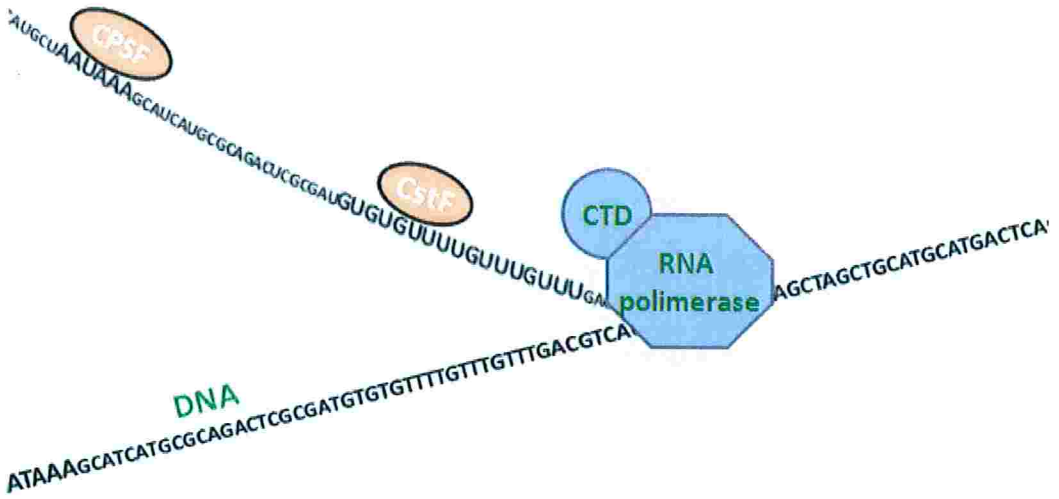


Figura 3: Ligação do CPSF e do CstF no mRNA durante a transcrição. A fita superior representa o mRNA e a inferior o DNA.

Os dois complexos protéicos CPSF e CstF também interagem e o complexo completo é estabilizado pela ligação de duas outras proteínas CF I e CF II (fatores de clivagem).

Após a montagem desse complexo, a ligação da *poly(A) polymerase* (PAP) estimula a clivagem do pré-mRNA no sítio de clivagem, 10 a 35 nucleotídeos após o hexâmero sinal e cerca de 30 nucleotídeos antes da região rica em G e U.

Os fatores de clivagem são liberados e o RNA *downstream* restante é rapidamente degradado. Uma enzima de atividade exonucleolítica 5'-3' inespecífica degrada o RNA clivado (Zhao J, 1999). Diversos trabalhos mostraram que a presença da subunidade CPSF-73kDa no sítio de clivagem sugere que esta também seja a exonuclease que inicie a degradação do produto *downstream* (Gilmartin, 2005).

A PAP inicia lentamente a síntese da cauda poli(A), adicionando aproximadamente 12 resíduos de adenina à extremidade 3' recém formada. Essa pequena cauda poli(A) é estabilizada por proteínas de ligação à poliadenina (PABP II – poly(A) binding protein II), que também acelera a adição de As pela PAP. A *poly(A) polymerase* é diferente das outras polimerases pois não necessita de um molde para a síntese da cauda poli(A). Após a adição de 200 a 250 resíduos de adenina, a reação se torna mais lenta e é finalizada através da

interferência da PABP II. O complexo de poliadenilação é então dissociado do mRNA final (Scorilas, 2002).

1.2 - Elementos no RNA reconhecidos pela maquinaria de poliadenilação

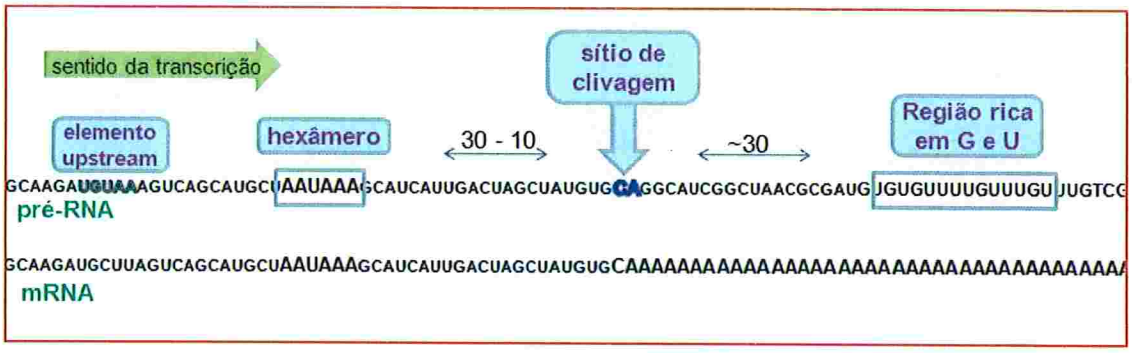


Figura 4: Elementos em *cis* que fazem parte do sinal de poliadenilação: o hexâmero e a região rica em G e U.

O CPSF reconhece o sinal AAUAAA, que é essencial para as reações de clivagem e de poliadenilação. Os seis nucleotídeos são necessários para a ligação e cadeias de RNA de pelo menos 10 bases podem ser reconhecidos com boa especificidade pelo CPSF (Zhao J, 1999).

Na literatura, alguns grupos consideram que o hexâmero AAUAAA é encontrado em 80 a 90% dos mRNAs seqüenciados e que um variante desse hexâmero, AUUAAA, que também promove a poliadenilação eficiente, é encontrado em aproximadamente 10% dos mRNAs (Zhao J, 1999) (Colgan & Manley, 1997).

O Hexâmero sinal de poliadenilação (AAUAAA)

O sinal de poliadenilação é muito conservado nos genes dos mamíferos. Em Ara et al. (2006), foi observado que 22% dos genes humanos, ou 4807 sítios de poliadenilação com o hexâmero AATAAA são conservados em genes ortólogos em camundongos. Dentre os sítios poli(A) conservados entre humanos e camundongos, 20% eram sítios únicos, e 2,5% eram sítios múltiplos, ou seja,

vários no mesmo exon 3' terminal. A partir desse resultado, os autores sugerem que os sítios únicos são evolucionariamente mais conservados (Ara, Lopez, Ritchie, Benech, & Gautheret, 2006). Deve-se notar que o critério adotado foi o último sinal do último exon 3' mapeado em um gene Ensembl e sua região genômica nos 10kb *downstream*, o que pode ser caracterizado como um critério muito abrangente. É também pouco realista, pois não considera a presença de transcritos nessa região.

Além disso, a configuração dos sítios poli(A) também é conservada, sítios únicos tendem a se manter únicos entre genes ortólogos de humanos e camundongos; o mesmo ocorre com sítios múltiplos (Tian, Hu, Zhang, & Lutz, 2005).

Embora não tenham podido explicar a conservação do sítio poli(A) através dos elementos em *cis*, foi observado que os sítios conservados eram mais ricos em "U" na região *downstream*, um indício de um elemento *downstream* mais forte (Ara, Lopez, Ritchie, Benech, & Gautheret, 2006).

O hexâmero sinal AATAAA é muito comum nas seqüências genômicas, pode aparecer uma vez a cada 4096 bases no caso de cada nucleotídeo ocorrer aleatoriamente (Tabaska & Zhang, 1999). Dessa forma, o reconhecimento da extremidade 3' dos mRNAs dependeria da identificação dos elementos *upstream* e *downstream*.

Upstream elements

A região 3'UTR também possui elementos *upstream* ao sinal de poliadenilação AAUAAA que influenciam a eficiência da poliadenilação.

O CF I reconhece o *upstream sequence element* (USE) e essa ligação colabora com a montagem da maquinaria de poliadenilação. Ainda não há uma seqüência consenso estabelecida, apenas se conhece para alguns genes. No colágeno humano, COL1A1, COL1A2 E COL2A1, o consenso é similar a $UAU_{2\text{-}5}GUNA$ (Natalizio, Muniz, Arkin, Wilusz, & Lutz, 2002). Foi demonstrado nesses três genes humanos de colágeno que a mutação dos USEs diminui a eficiência da poliadenilação *in vivo* e *in vitro*, e que a inclusão de oligorribonucleotídeos competidores na posição das USEs inibe a poliadenilação (Natalizio, Muniz, Arkin, Wilusz, & Lutz, 2002).

Downstream elements

O CstF reconhece uma região *downstream* ao sítio de clivagem, rica em "G" e "U". Essa região é menos conservada e mais difusa que o hexâmero sinal, mas geralmente está localizada a partir de 30 nucleotídeos *downstream* ao sítio de clivagem. O sítio de clivagem em geral ocorre ao lado 3' de um dinucleotídeo CA, sendo que o "C" ocorre em aproximadamente 59% dos mRNAs seqüenciados e o "A" em 70% dos casos (Mandel C.R., 2007).

O *downstream sequence element* (DSE) é composto por um elemento rico em "G" e "U", com seqüência consenso YGUGUUYY (Y=pirimidina), e mais adiante, *downstream*, um elemento composto por UUUUU (Cañadillas & Varani, 2003) (Zarudnaya, Kolomiets, Potyahaylo, & Hovorun, 2003).

O domínio C-terminal do CstF se desdobra durante a interação com o RNA e forma um complexo com os fatores da maquinaria de poliadenilação. A ligação mais forte do CstF com o RNA ocorre através dos Us consecutivos e os diferentes elementos dentro da região rica em GU conferem as diferentes afinidades da ligação. Essa distinção estrutural entre seqüências de ligação estável e instável proporciona a diferença entre sítios poli(A) mais fortes e mais fracos. (Zarudnaya, Kolomiets, Potyahaylo, & Hovorun, 2003)

Além do hexâmero sinal AAUAAA e da região *downstream* rica em "G" e "U", alguns pré-mRNAs têm outros elementos gênicos que influenciam a eficiência do uso dos sinais de poliadenilação. Alguns exemplos são o gene de complemento humano C2 e de calcitonina em camundongos, que foram identificados como genes que contém sinais de poliadenilação não canônicos e necessitam de elementos auxiliares para promover a utilização eficiente (Zhao J, 1999).

1.3 - Proteínas envolvidas no processo de poliadenilação

O hexâmero sinal AAUAAA é necessário durante as duas fases, a clivagem e a poliadenilação, que são altamente ligadas *in vivo*. No entanto, essas duas fases são separáveis no estudo *in vitro*. O CPSF reconhece a seqüência AAUAAA independentemente de estrutura secundária (Zhao J, 1999), mas a ligação do CPSF purificado é fraca, somente aumentada pela interação cooperativa do CstF

à fita de RNA (Colgan & Manley, 1997). O complexo ternário RNA+CPSF+CstF torna-se estável e pode funcionar para recrutar outros componentes da maquinaria de poliadenilação para o sítio de clivagem. Dessa forma ele teria participação na especificação do sítio de clivagem e poliadenilação, através de sua interação com o CstF e a *poly(A) polymerase*. Outras evidências sugerem que o CPSF também ajuda na coordenação da poliadenilação nuclear com a transcrição, participa da poliadenilação citoplasmática, ajuda a limitar o tamanho da cauda poli(A) e tem interação com proteínas associadas à maquinaria de *splicing* (Colgan & Manley, 1997).

O CPSF purificado é constituído por 4 subunidades, cujas massas moleculares são: 160, 100, 73 e 30 kDa. A subunidade maior (160kDa) é bem caracterizada, contém um sinal de localização nuclear (NLS) bipartido, e seqüências similares aos motivos RNP1 e RNP2 encontrados em proteínas de ligação ao RNA (Zhao J, 1999). A subunidade maior liga-se preferencialmente ao RNA que contém AAUAAA, embora sua ligação seja mais fraca do que observado para o CPSF intacto, sugerindo que a participação das outras subunidades do CPSF facilita o reconhecimento do AAUAAA (Zhao J, 1999). A subunidade de 160kDa do CPSF faz contatos com a CstF e com a PAP, sugerindo que essa subunidade maior tem papel importante na coordenação da clivagem e da síntese da cauda poli(A). No entanto, a subunidade CPSF-160 e a PAP, purificadas, não são suficientes para reconstituir a adição de poli(A) dependente de AAUAAA (Zhao J, 1999).

As reações são concomitantes: no início da transcrição e fosforilação do domínio carboxi terminal da RNA Polimerase II, o CPSF dissocia do fator de transcrição TFIID e se associa à Pol II elongadora (Dantonel 1997) (Colgan & Manley, 1997)

O CPSF e o CstF permanecem associados à Pol II até que encontrem os elementos de poliadenilação em *cis* no RNA e definam o sítio de clivagem. Assim, o domínio carboxi terminal da Pol II é essencial para a poliadenilação e o *splicing* eficientes (Colgan & Manley, 1997).

O CstF (cleavage stimulatory factor) é uma proteína heterotrimérica, cujas subunidades têm pesos moleculares de 77, 64 e 50 kDa. Na montagem do CstF, a subunidade CstF-77 fica no meio, ligando as duas outras, de forma linear (Zhao J, 1999).

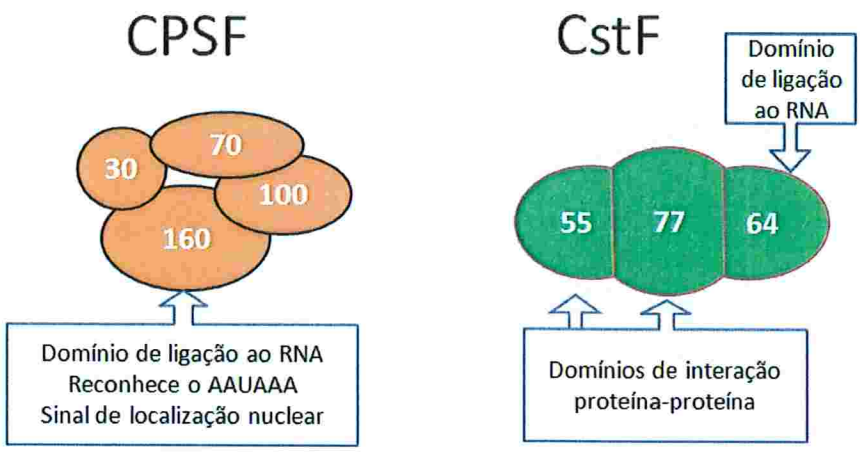


Figura 5: As duas principais proteínas de reconhecimento dos elementos de poliadenilação no RNA. Cleavage and Polyadenylation Specificity Factor (CPSF) e Cleavage Stimulatory Factor (CstF). Elas interagem entre si e com o RNA, formando um complexo estável.

Na presença do CPSF, o CstF se liga fortemente ao RNA, principalmente através da subunidade CstF-64, que reconhece a região rica em G e U e possui um clássico domínio de ligação ao RNA em sua porção aminoterminal (Colgan & Manley, 1997).

O CstF é necessário para a clivagem, mas não para a adição da cauda poli(A). Os fatores de clivagem CF I e CF II são necessários para a clivagem, mas não para a adição de poli(A) (de Vries, Rüegsegger, Hübner, Friedlein, Langen, & Keller, 2000).

O fator de clivagem I (CF I) é um heterodímero, composto por um polipeptídeo de 25 kDa, combinado a outro peptídeo de 59, 68 ou 72 kDa. Sua estrutura protéica, composta por diversos domínios de ligação ao RNA e semelhante a proteínas do spliceossomo, sugere um papel potencial do CF I na coordenação do processamento 3' com o *splicing* do pré-mRNA (Venkataraman, Brown, & Gilmartin, 2005).

Análises da cinética da reação de clivagem indicam que a interação do CF I com o RNA substrato é um passo inicial na montagem do complexo de processamento 3', que facilita o recrutamento dos outros fatores (Rüegsegger U, 1996).

O CF I tem maior afinidade por RNAs com sinais de poliadenilação que com RNAs não relacionados. O CF I também ajuda a estabilizar o CPSF no RNA

substrato. As principais funções do CF I são o reconhecimento adicional do pré-mRNA substrato e ajudar na definição do sítio de poliadenilação. Estudos de *UV cross-linking* entre o CF I e o RNA revelaram que a ligação ocorre preferencialmente em regiões que contêm a seqüência UGUAN (onde a preferência para a base N é A > U > C/G) (Brown & Gilmartin, 2003) (Mandel C.R., 2007). Além disso, o CF I melhora o reconhecimento de seqüências que contêm tanto o hexâmero canônico, como o não canônico (Venkataraman, Brown, & Gilmartin, 2005).

Em seqüências que não contêm o hexâmero canônico, o CF I é capaz de direcionar a clivagem para um sítio *downstream* a uma região rica em A, utilizando-a como se fosse um hexâmero. Nesse modelo, o CF I recrutaria os outros fatores de clivagem para a região correta, devido à sua ligação às seqüências UGUAN upstream à região rica em A (Venkataraman, Brown, & Gilmartin, 2005).

Mesmo no contexto do hexâmero canônico AAUAAA, mutações nos elementos UGUAN proximais diminuem a eficiência da clivagem (Venkataraman, Brown, & Gilmartin, 2005). Mutações pontuais nos elementos UGUAN têm efeito comparável na adição de poli(A) em RNAs pré-clivados.

A ligação específica de uma das subunidades (hClp1) do fator de clivagem II (CF II) com CF I e CPSF foi comprovada por imunoprecipitação, seguida de análise por *Western blotting*. As proteínas CstF e PAP bovina purificadas não foram imunoprecipitadas com a subunidade de CF II. Esses dois resultados indicam que o fator de clivagem II de mamíferos (CF IIm) é essencial para a clivagem, mas não para a poliadenilação (de Vries, Rüegsegger, Hübner, Friedlein, Langen, & Keller, 2000).

A poli(A) polimerase (PAP) é o componente mais conhecido e estudado da maquinaria de poliadenilação. Ela é composta por um único polipeptídeo e são conhecidas pelo menos isoformas provindas de *splicing* alternativo. A PAP II é a forma mais conhecida, composta por 740 aminoácidos (Colgan & Manley, 1997) (Scorilas, 2002).

Como a PAP é necessária na formação de um complexo de clivagem ativo, é possível que RNAs mais curtos sejam provenientes de um mecanismo de *feedback* negativo que regule a expressão de PAP. Se a PAP "*full length*" estiver presente em altos níveis, ela pode reconhecer sítios de poliadenilação mais fracos

e *upstream*, em seu próprio mRNA, ocasionando a expressão de transcritos de PAP mais curtos, instáveis ou inativos (Colgan & Manley, 1997). Os dois terços amino-terminais da PAP são altamente conservados em eucariontes e contêm um domínio catalítico com homologia a uma família de nucleotidil transferases, incluindo muitas DNA e RNA polimerases (Zhao J, 1999).

A poli(A) polimerase (PAP) contém um domínio catalítico e uma região regulatória C terminal rica em serina e treonina, que contém sítios cdk (cyclin-dependent kinase) consenso e não consenso. A PAP é fosforilada pela cdc2-ciclinaB nesses sítios *in vitro* e *in vivo* e é inativada pela hiperfosforilação nas células na fase M da mitose e meiose, quando a cdc2-ciclinaB está ativa. Assim, ela é alvo de regulação temporal, durante a progressão do ciclo celular (Zhao & Manley, 1998).

Em experimentos com células que expressavam menores quantidades do tipo selvagem da PAP ou a forma cdk- PAP (incapaz de ser fosforilada) foram analisados o crescimento celular e a progressão do ciclo. Ambos os tipos celulares apresentaram defeitos em relação às células selvagens, sendo que a mais afetada foi a cdk- PAP. Esses resultados indicam que os níveis da PAP devem ser altamente regulados durante o ciclo celular e suportam a hipótese de que a inibição da PAP pela fosforilação da cdc2-cyclinB é importante para o crescimento normal da célula (Zhao & Manley, 1998).

As células toleram bem níveis baixos de PAP, sem interferência no crescimento celular, mas o limite máximo tem que ser muito controlado, pois pode ser tóxico. Os níveis de CstF-64 podem ser reduzidos 10 vezes sem que haja efeito significante no crescimento celular.

Mesmo em células heterozigóticas, a expressão de PAP nunca ultrapassa muito os níveis endógenos. O mesmo não ocorre com CstF-64, que pode ser superexpresso até 50 vezes os níveis endógenos sem afetar o crescimento celular. Esse fato indica que a superexpressão de CstF-64 não é tóxica para a célula e é consistente com o fato de que os níveis de CstF-64 são regulados durante a diferenciação das células B e podem modular a escolha do sítio poli(A) da IgM de cadeia pesada.

Podemos observar na literatura que os níveis de CstF-64 podem flutuar livremente, enquanto que os níveis de PAP devem ser estritamente controlados.

Na fase M do ciclo celular, há uma repressão da síntese de RNA e proteínas, e a inibição da PAP contribui para isso. Em experimentos de superexpressão de PAP na fase M, observou-se que o crescimento celular diminui drasticamente, e há maior número de células na fase G0/G1. Possivelmente, com a ruptura do envelope celular, um excesso de PAP no citoplasma pode estabilizar alguns mRNAs indevidamente, atrasando a entrada na fase S ou mesmo fazendo com que as células saiam do ciclo celular e permaneçam na fase G0 (Zhao & Manley, 1998).

A quantidade de CstF-64 aumenta 5 vezes durante a transição da fase G0 para a fase S, quando a proliferação é estimulada. Esse aumento é concomitante com o aumento de RNAs poliadenilados entre 5 a 8 vezes no citoplasma (Martincic, Campbell, Edwalds-Gilbert, Souan, Lotze, & Milcarek, 1998).

Em condições fisiológicas e na ausência de outros fatores, a PAP tem muito pouca atividade. A PAP é específica para a utilização de ATP, mas não para o RNA substrato. Na poliadenilação dependente de AAUAAA, a PAP é recrutada para o complexo de processamento pela interação com o CPSF-160 e também é necessária para a clivagem da maioria dos pré-mRNAs (Zhao J, 1999).

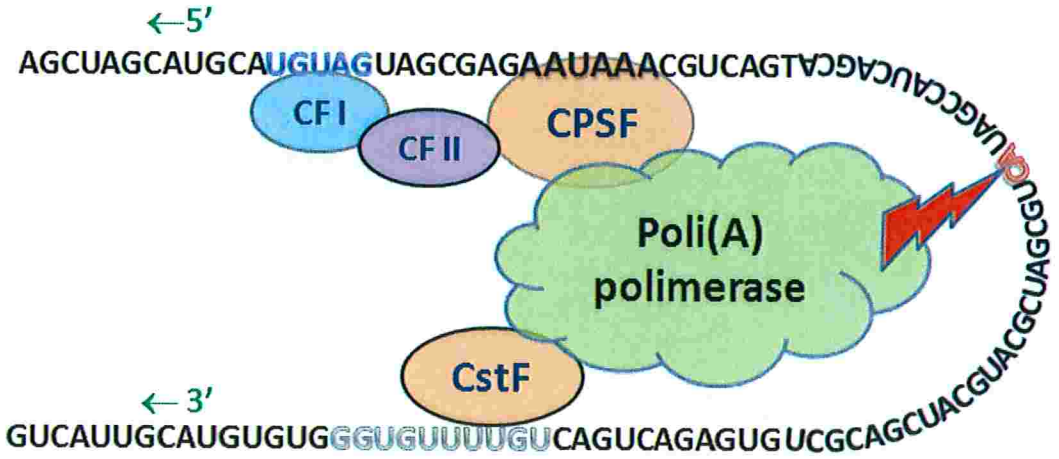


Figura 6: Montagem do complexo de clivagem do mRNA nascente. Pode-se observar a ligação do CF I, do CPSF e do CstF no pré-mRNA, além da ligação entre as proteínas.

O tamanho máximo das caudas poli(A) encontradas em RNA recentemente sintetizados *in vivo* é de 200 a 300 resíduos. Essa restrição de tamanho é mediada pela *poly(A) binding protein II* (PABP II), uma proteína nuclear com alta afinidade por poli(A). Após a síntese de um pequeno trecho de cauda poli(A), a PABP II é ligada e forma um complexo quaternário com a PAP e o CPSF. Esse complexo estabiliza a ligação da PAP à extremidade 3' do RNA, e ajuda a síntese rápida da cauda poli(A) longa. O controle do tamanho é feito quando se perde a interação entre o RNA, o CPSF e a PAP (Colgan & Manley, 1997).

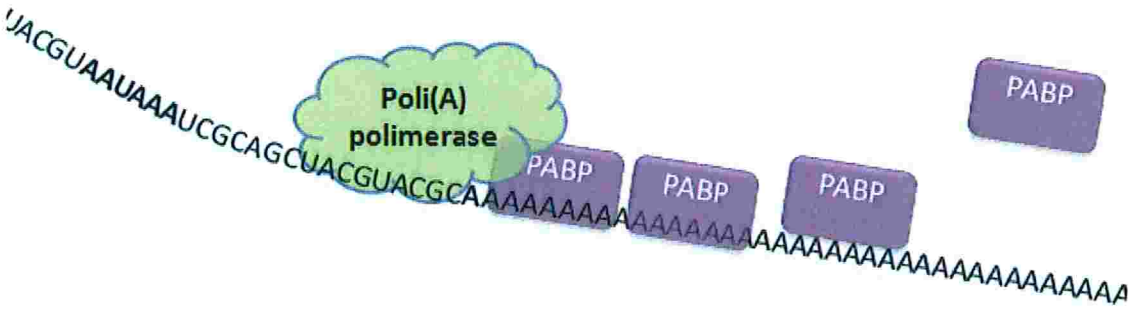


Figura 7: Ligação da poly(A) binding protein (PABP) à cauda poli(A).

A presença da cauda poli(A) fornece ao mRNA nascente um sítio de ligação para uma classe de fatores de regulação, as *poly(A) binding proteins* (PABP) (Mangus, Evans, & Jacobson, 2003). Elas não têm atividade catalítica, mas medeiam a expressão gênica. As PABPs estabilizam o mRNA no citoplasma, evitando sua degradação. Alem disso, a associação com as PABPs é necessária para que alguns mRNAs passem pelo poro nuclear.

As PABPs facilitam a formação da estrutura circular do mRNA que promove a tradução (Mangus, Evans, & Jacobson, 2003).

Em células de mamíferos a PABP se liga aos primeiros 11 a 14 resíduos de adenina juntamente com o CPSF e estimula a *poly(A) polymerase* (PAP), até então em modo distributivo, a se tornar processiva, ou seja, a adicionar os nucleotídeos continuamente em alta velocidade. As PABPs continuam a se ligar à cauda poli(A) nascente até que esta atinja 200 nucleotídeos ou mais (em mamíferos). Em seguida a poli(A) polimerase volta ao modo distributivo, no qual é

solta do mRNA após a adição de mais alguns nucleotídeos (Mangus, Evans, & Jacobson, 2003).

1.4 - Importância da região 3'UTR

A região 3'UTR é importante para o direcionamento dos transcritos para compartimentos subcelulares específicos e para o controle da tradução. A tradução localizada de mRNA é uma forma rápida e eficiente de reação ao estímulo extracelular com benefício de resolução espacial da resposta celular.

Diversos organismos utilizam o mecanismo de localização dos mRNAs em subcompartimentos celulares e isso é evolutivamente conservado (Andreassi & Riccio, 2009).

A localização dos mRNA é relevante em células somáticas altamente polarizadas, como neurônios diferenciados, pois o sítio de transcrição pode ser muito distante de onde a proteína será realmente utilizada e o transporte de uma proteína pode demorar mais que sua meia-vida (Andreassi & Riccio, 2009).

A importância das seqüências da região 3'UTR

A maioria dos elementos regulatórios dos mRNAs estão situados no 5' e 3'UTRs, onde agem como plataformas para a montagem de complexos protéicos nos mRNAs, gerando ribonucleoparticulas (mRNPs). O 5'UTR está principalmente envolvido com o controle da tradução do mRNA e o 3'UTR regula múltiplos aspectos do metabolismo dos mRNAs, como exportação do núcleo, localização citoplasmática, eficiência da tradução e estabilidade dos mRNAs.

A localização assimétrica dos mRNAs é basicamente determinada pelos elementos em *cis* localizados no 3'UTR, com raras exceções. O tamanho dos elementos de localização pode variar desde alguns nucleotídeos até mais de 1kb e é possível encontrar diversas cópias do mesmo elemento ou uma combinação de diferentes elementos (Andreassi & Riccio, 2009).

Proteínas que agem em *trans* reconhecem os elementos de localização na seqüência do mRNA ou na estrutura secundária. No entanto, há pouca conservação entre as seqüências 3'UTR e não se conhece quais proteínas fazem esse tipo de ligação.

As proteínas (fatores em *trans*) podem reconhecer os elementos de localização baseadas na seqüência de mRNA ou na estrutura secundária. Embora as proteínas que se ligam nos elementos de localização regulem o transporte de mRNPs, poucas foram identificadas até agora. Somado ao fato de que as seqüências na 3'UTR são pouco conservadas, a identificação dos elementos de localização é muito difícil por predição.

Segundo Wang et al. (2008) mais de 90% dos genes humanos sofrem *splicing* alternativo. Além de alterar a seqüência da proteína, as UTRs também sofrem *splicing* alternativo e essa variabilidade pode afetar a tradução, estabilidade ou localização dos mRNAs. O maior grau de variabilidade das isoformas está no uso de sítios de poliadenilação alternativa, resultando numa 3'UTR mais longa ou mais curta. Importante: o *splicing* pode ser acoplado à poliadenilação pelo uso de elementos em *cis* conservados, reconhecidos por ambas as maquinarias de *splicing* e de poliadenilação. Diversos motivos relacionados ao *splicing* foram identificados na 3'UTR de transcritos que sofrem poliadenilação alternativa (Wang, et al., 2008).

Wang et al. (2008) revelaram que seqüências consenso que são alvo dos fatores de *splicing* STAR também estão presentes em transcritos sujeitos à poliadenilação alternativa.

As proteínas STAR são ligantes de mRNA altamente conservadas, fatores de *splicing*, envolvidas no desenvolvimento das células musculares e germinativas em *C. elegans* e *Drosophila*. No cérebro humano, os motivos de ligação das proteínas STAR estão enriquecidos ao redor dos sítios alternativos de poliadenilação e em seqüências intrônicas localizadas *upstream* ao exon 21, que possui *splicing* alternativo. Além disso, camundongos que não tinham a proteína STAR-Quakin1 apresentaram hipomielinização devido a defeitos no processamento do mRNA da MBP e sua localização nos oligodendrócitos, sugerindo assim que esses fatores de *splicing* têm papel no direcionamento dos mRNAs (Wang, et al., 2008).

1.5 - Localização dos mRNAs no citoplasma

Os mRNAs podem ser "marcados" no núcleo para determinar sua localização citoplasmática. Essa marcação consiste no reconhecimento dos elementos chamados *zipcodes*, localizados na região 3'UTR dos mRNAs, por proteínas componentes da maquinaria de localização. Isso foi demonstrado por Kress et al. (2004) em oócitos de *Xenopus*, pela ligação de RNA *binding proteins* específicas, que reconhecem os elementos na seqüência do mRNA no núcleo. Após o transporte ao citoplasma, esse complexo ribonucleoprotéico é remodelado e fatores de transporte adicionais são recrutados. Assim, o RNA pode ser levado à região apropriada no citoplasma por motores moleculares ligados a microtúbulos (Kress, Yoon, & Mowry, 2004).

Alguns RNAs contêm toda a informação necessária para o transporte citoplasmático em um único *zipcode*. Um exemplo é o mRNA do gene MAP2, um transcrito localizado no dendrito, que contém um elemento de 640 nucleotídeos em sua 3'UTR, que é necessário e suficiente para o transporte. Como esse elemento é grande e (segundo a predição estrutural) deve conter diversos domínios estruturais, é possível que sub-elementos distintos no *zipcode* efetuem passos individuais no processo de localização (Jambhekar & Derisi, 2007).

Um *zipcode* de 54 nucleotídeos na 3'UTR do mRNA de beta-actina de galinha também contém múltiplos motivos que direcionam a localização sinergicamente para a extremidade principal de fibroblastos de galinha. Dois motivos, GGACT e AATGC foram encontrados no *zipcode* de 54 nucleotídeos e numa região separada, de 43 nucleotídeos, que tem uma fraca habilidade de localização. Foi demonstrado que os dois motivos juntos têm uma maior eficiência na localização, e que, além disso, a região rica em AC entre os dois motivos do *zipcode* é essencial para sua atividade (Kislauskis, Zhu, & Singer, 1997). Duas seqüências ACACCC nessa região se ligam a ZBP1, uma proteína envolvida no transporte de RNA (Jambhekar & Derisi, 2007). O mesmo padrão de seqüência ocorre também em integrina alfa3 e é um motivo necessário para a localização de complexos de adesão na periferia de células humanas em cultura (Adereth, Dammai, Kose, Li, & Hsu, 2005).

Para estudar o papel das seqüências no 3'UTR da integrina alfa3 na localização da proteína, Adereth et al. (2005) fizeram um experimento em que

colocaram a seqüência completa do 3'UTR *downstream* à ORF (open reading frame) da proteína fluorescente GFP. Como controle, a ORF de GFP foi associada às 3'UTR do vírus SV40, um clássico controle em experimentos com a região 3'UTR. Após a transfecção, as células mostraram um padrão pontuado de localização da quimera GFP-integrina enquanto a quimera GFP-SV40 mostrou um padrão de expressão difusa por toda a célula. A proteína GFP provinda da quimera GFP-integrina co-localizou-se com a MLP1 e foi demonstrado também que sua localização depende de MLP1 (myosin-like protein 1) (Adereth, Dammai, Kose, Li, & Hsu, 2005).

Outro mecanismo para localização dos mRNAs no citoplasma é a estabilização seletiva. O exemplo mais característico é a proteína *heat shock 83* (hsp83), que se localiza no pólo posterior de embriões de *Drosophila*. O mRNA de hsp83 é degradado em todo o citoplasma, exceto no pólo posterior. Ambos os mecanismos de proteção e de degradação foram atribuídos a elementos da 3'UTR do mRNA de hsp83. Elementos de proteção que agem em *cis* estão localizados na 3'UTR dos mRNAs. A deleção desses elementos leva à degradação desses mRNAs no embrião inteiro (Bashirullah, Cooperstock, & Lipshitz, 2001).

1.6 - Influência da cauda poli(A) na tradução

A formação da extremidade 3' dos mRNAs através da poliadenilação tem um papel crítico na expressão gênica porque mRNAs que não são propriamente processados não serão transportados para fora do núcleo e não serão traduzidos no citoplasma.

As diferentes espécies de RNA são discriminadas por fatores de exportação distintos de acordo com o seu tipo. A cauda poli(A) funciona como um elemento de identificação para a exportação do mRNA (Fuke & Ohno, 2008).

Em um estudo comparativo entre mRNAs poliadenilados e não poliadenilados, de mesma seqüência, observou-se que os não poliadenilados têm menor capacidade de serem traduzidos, não somente por serem degradados mais rapidamente, mas pela menor eficiência em se associarem aos polissomos (Munroe & Jacobson, 1990).

A falta da cauda poli(A) afeta uma etapa tardia do início da tradução. A redução na eficiência da formação do complexo de iniciação 80S dos mRNAs sem a cauda poli(A) é consistente com a redução da eficiência da tradução e do recrutamento do ribossomo. Foi observado que a eficiência da tradução é maior quando há a presença da cauda e do 5'cap. Quando um dos dois não está presente a eficiência diminui, e é ainda menor quando ambos estão ausentes, em mRNAs que são traducionalmente ativos (Munroe & Jacobson, 1990).

Após a entrada do mRNA no citoplasma, a associação da PABP com a cauda poli(A) promove a interação das extremidades 3' e 5' do mRNA, o que estimula o início da tradução. A formação dessa estrutura circular promove o recrutamento da subunidade 40S do ribossomo, que forma o complexo de iniciação da tradução, juntamente com a interação entre a PABP, o fator de iniciação eIF4G e a proteína de ligação ao 5'cap eIF4E (Mangus, Evans, & Jacobson, 2003). A PABP, uma vez ligada à extremidade 3' do mRNA, facilita a ligação do fator de iniciação da tradução eIF4G.

A combinação desses efeitos proporciona uma garantia de que a maquinaria de tradução preferencialmente traduza mRNAs com cauda poli(A) e 5'cap, além de promover a reciclagem do ribossomos do 3' para o 5' do mesmo mRNA. (Mangus, Evans, & Jacobson, 2003)

O mesmo motivo de reconhecimento da cauda poli(A) pela proteína PABP é usado para a interação com o eIF4G. Experimentos com a ligação de PABP em RNAs repórteres mostraram que ele estimula a tradução mesmo de RNAs não poliadenilados. Esse tipo de constatação demonstra que no caso do início da tradução, a cauda poli(A) fornece apenas um sítio de ligação para a PABP (Mangus, Evans, & Jacobson, 2003).

Há mais de 10 anos, as evidências genéticas e bioquímicas mostraram que a PABP pode estar envolvida na etapa inicial da tradução (Prévôt, Darlix, & Ohlmann, 2003). A evidência da ligação entre as extremidades 5' e 3' foi descrita primeiramente em *S. cerevisiae* e envolve a interação entre PABP e eIF4G, dependente de RNA (Tarun Jr, Wells, Deardoff, & Sachs, 1997). A prova formal da circularização 5'-3' veio com a visualização por microscopia de força atômica do complexo eIF4E/eIF4G/PABP formado por proteínas recombinantes sobre um mRNA poliadenilado e com 5'*cap* (Wells, Hillner, Vale, & Sachs, 1998). O domínio de interação da PABP foi mapeado em 114 amino ácidos na região de N-terminal

da proteína eIF4G expressa a partir do gene Tif4632p em leveduras. Em mamíferos o sítio de interação na seqüência N-terminal estendida do eIF4G I e conservada em eIF4G II, mas não tem homologia com o domínio de interação da PABP em leveduras (Tarun Jr, Wells, Deardoff, & Sachs, 1997).

Em leveduras, o fator de tradução eIF4G se associa com ambas as proteínas de ligação ao *cap* eIF4E e com a proteína de ligação ao poli(A) Pab1p. A associação de Pab1p com eIF4G media a habilidade da cauda poli(A) estimular a tradução *in vitro*, e essa associação não é essencial *in vivo*, a não ser que a função da proteína eIF4E de ligação ao *cap* esteja comprometida (Tarun Jr, Wells, Deardoff, & Sachs, 1997).

1.7 - Estabilidade dos mRNAs: Influência da região 3'UTR na expressão dos mRNAs

Os mRNAs têm tempo de vida diferentes, desde alguns minutos até vários dias. A duração da vida de cada mRNA determina o tempo em que o mesmo pode ser transcrito e o nível de expressão do gene correspondente (Nguyen-Chi & Morello, 2008).

O motivo de instabilidade mais estudado é chamado de A-U *rich element* (ARE), uma seqüência rica em adeninas e uridinas, presente na região 3'UTR dos RNAs instáveis. Essas seqüências são reconhecidas pelas AU *binding proteins* (AUBP), que participam do transporte entre o núcleo e o citoplasma, do controle da estabilidade e da tradução no citoplasma (Nguyen-Chi & Morello, 2008).

A comparação de seqüências ARE de diversos mRNAs de oncogenes e citocinas levou à identificação dos motivos presentes na 3'UTR, 50 a 150 nucleotídeos ricos em A-U. AREs geralmente contem repetições do pentâmero AUUUA.

Atualmente está estabelecido que a degradação dos mRNAs não é um processo padrão, que utiliza nucleases não específicas para degradar o substrato indiscriminadamente. Ao invés disso, a degradação é um processo estreitamente regulado, que emprega fatores específicos em trans atuando em seqüências em cis. O primeiro relato de uma seqüência consenso rica em A-U é de 1986, observado na 3'UTR do fator de necrose tumoral de murinos e humanos (Caput,

Beutler, Hartog, Thayer, Brown-Shimer, & Cerami, 1986), assim como em outros mRNAs que sugerem funções regulatórias, como linfotoxina, fator de estímulo de formação de colônia, interleucina e fibronectina (Bevilacqua, Ceriani, Capaccioli, & Nicolin, 2003).

A deadenilação é o primeiro passo da degradação dos mRNAs. A inclusão das seqüências ARE na 3'UTR nos mRNAs acelera sua degradação, pois levam à rápida remoção da cauda poli(A) e à perda do 5'cap.

As seqüências ARE são consideravelmente diferentes em tamanho, conteúdo AU e número de motivos AUUUA, e ainda não se sabe quais características representam os elementos funcionais, nem como é controlada a afinidade e a cinética da ligação.

Para que proteínas específicas formem complexos estáveis com mRNAs contendo AREs, aparentemente são necessárias diversas iterações com o motivo AUUUA, promovidas por elementos estruturais (*stem-loop*) formados pelo mRNA (Bevilacqua, Ceriani, Capaccioli, & Nicolin, 2003).

No entanto foi criado um banco de dados chamado ARED, que classifica os mRNAs de acordo com o número de iterações do motivo AUUUA (Bakheet, Williams, & Khabar, 2006). Foram compilados mRNAs de diversas fontes, resultando em 2500 genes não redundantes, contendo ARE. Esse banco de dados classifica as ARE específicas da 3´UTR segundo o número de pentâmeros na região rica em U. A Classe I contém somente um pentâmero e a Classe II contém dois ou mais, sendo que 70% dos mRNAs pertencem à Classe I. Os genes que contém seqüências ARE pertencem a categorias funcionais de processos regulatórios intracelulares, sinalização e metabolismo de ácidos nucléicos (Bakheet, Williams, & Khabar, 2006).

A estabilidade dos mRNAs é alterada no câncer. Alterações genéticas observadas em diversos tipos de câncer são capazes de desregular o nível de expressão do mRNA mutado. (Nguyen-Chi & Morello, 2008)

Uma proteína em particular, HuR, está envolvida na estabilização de mRNAs que contém ARE e participa de processos de sinalização como o da inflamação.

A proteína HuR contém três motivos de reconhecimento de RNAs, através dos quais se liga a mRNAs específicos que contém seqüências ricas em U e AU, afetando sua estabilidade e tradução. O aumento da expressão de HuR foi observado em diversos tipos de câncer. Quando a proteína HuR está super

expressa, ela estabiliza transcritos contendo ARE e promove sua tradução. No câncer coloretal, HuR aumenta a estabilidade e tradução da ciclo-oxigenase2 (COX2) (Anant, Houchen, Pawar, & Ramalingam, 2010).

No início da tumorigênese do câncer coloretal há um aumento na expressão de COX2. A COX2 catalisa a formação da prostaglandina em estados patológicos, bem como se observa a síntese elevada de prostaglandinas em sítios tumorais. Nas células normais a COX2 é regulada no nível pós-transcricional através de elementos na seqüência de sua região 3'UTR. Um sinal de poliadenilação alternativo resulta em uma região 3'UTR reduzida, com a perda de elementos de regulação (Young & Dixon, 2010).

O gene PTGS2, que codifica a COX2, contém sua 3'UTR inteira no exon 10, uma região com diversos sinais de poliadenilação, capaz de produzir transcritos entre 2.8kb e 4.6kb. Nas células normais o sinal de poliadenilação canônico (AAUAAA) mais distal é usado, produzindo mRNAs com a região 3'UTR íntegra. No entanto, observou-se que em células de câncer coloretal *in vitro*, o sinal de poliadenilação proximal (AUUAAA) é utilizado, resultando num mRNA que não possui alguns elementos regulatórios em seu 3'UTR encurtado. Assim este mRNA escapa da regulação pós-transcricional e permanece estável no citoplasma das células tumorais.

Os AREs estão entre os fatores em *cis* mais predominantes na região 3'UTR dos mRNAs e regulam sua estabilidade.

Os mRNAs mais conhecidos que contém ARE são os que codificam para Interferon, como o IFN- alfa e beta, que são responsáveis pela defesa precoce contra vírus, e citocinas, como a interleucina, que são produzidas em resposta a um estímulo inflamatório. Uma análise recente mostrou que os mRNAs que contém AREs representam 8% dos genes humanos transcritos e codificam diversas proteínas importantes para o crescimento celular, hematopoiese, transdução de sinal e apoptose, entre outros (Khabar, 2005).

Muitas mudanças observadas na estabilidade dos mRNAs são refletidas nos níveis protéicos. O decaimento regulado dos produtos gênicos é crucial para a homeostase normal de vários processos biológicos.

1.8 - Poliadenilação alternativa

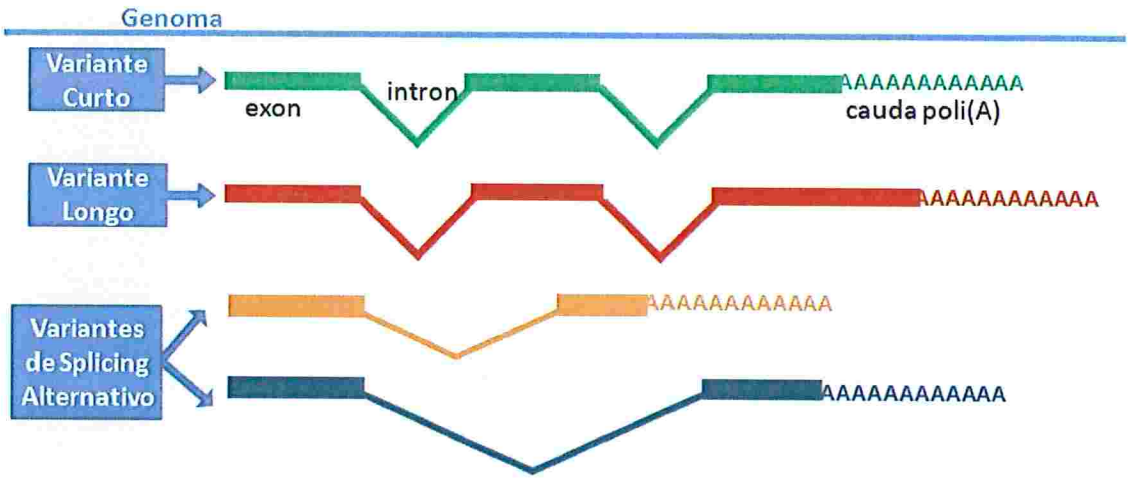


Figura 8: Exemplo de transcritos obtidos a partir de um mesmo cluster através da poliadenilação alternativa

O uso de sítios alternativos de poliadenilação produz isoformas de mRNA com diferentes regiões 3' UTR e algumas vezes com diferentes regiões codificantes. Metade dos genes humanos tem sítios alternativos de poliadenilação e a escolha do sítio depende do tecido no qual o gene é expresso, da fase do ciclo celular ou de fatores externos que influenciam a regulação da expressão gênica.

A poliadenilação alternativa pode ocorrer de duas formas: Dependente de *splicing*, num processo chamado de 3'exon *switching* ou independente de *splicing*, através do uso de diferentes sítios de poliadenilação no exon terminal, também chamado *tandem* UTRs (Zlotorynski & Agami, 2008).

Segundo Edwalds-Gilbert, Veraldi, & Milcarek (1997), ao menos metade dos genes humanos estão sujeitos ao processamento alternativo da região 3', que leva a mRNAs com extremidades 3' de tamanho variável. Conforme mencionado anteriormente, a região 3'UTR contém elementos em *cis* importantes para a estabilidade e localização dos mRNAs no citoplasma e a eficiência da tradução. A poliadenilação alternativa é controlada por elementos em *cis* e por fatores em *trans* e já foi demonstrado que ocorre de acordo com o tipo celular e com algumas doenças (Edwalds-Gilbert, Veraldi, & Milcarek, 1997). Segundo Dahary et

al.(2005), quando os transcritos têm vários sinais poli(A), estes competem pela poliadenilação, sendo que em geral, o sinal anterior é escolhido.

Foi estimado que mais de 29% dos genes humanos sofrem poliadenilação alternativa (Beaudoing & Gautheret, 2001). Segundo Zhang et al. (2005) há 13.942 genes com 29.283 sítios de poliadenilação e 6.418 genes têm somente um sítio. Os genes com poliadenilação alternativa somam 7.524, ou seja, 54% dos genes possuem mais de um sítio de poliadenilação (Zhang, Hu, Recce, & Tian, 2005).

Em um estudo mais antigo, foi utilizado um banco de dados de 3'UTRs humanas e um programa que produz o panorama da poliadenilação alternativa de cada tecido baseado em ESTs. Assim, foram encontradas 5.127 UTRs contendo dois ou mais prováveis sítios de poliadenilação, em aproximadamente 13 mil UTRs, ou seja, 39% dos genes (Beaudoing & Gautheret, 2001).

O mesmo trabalho descreve que em dois terços dos tecidos analisados a forma predominante dos genes com poliadenilação alternativa continha a UTR mais curta; a densidade dos elementos ricos em A e U (ARE) não mostrou nenhum viés (Beaudoing & Gautheret, 2001).

Um estudo com 4800 sítios de poliadenilação conservados entre humanos e camundongos mostrou que os sítios conservados possuem maior eficiência de processamento que sítios não conservados, além de que a ordem 5'-3' dos sítios é mais conservada do que o esperado aleatoriamente (Ara, Lopez, Ritchie, Benech, & Gautheret, 2006).

A poliadenilação alternativa não pode ser predita a partir da seqüência genômica, pois o sinal de poliadenilação ou a região rica em G e U não constituem assinaturas (Beaudoing & Gautheret, 2001). Os dados mais confiáveis são os experimentais, como os mRNAs, RefSeqs e ESTs. No entanto, Hu et al. (2005) usaram a informação de que as seqüências *upstream* e *downstream* a um sítio poli(A) são ricas em U, para predizer os sítios de poliadenilação . Foram encontrados 7524 genes com múltiplos sítios poli(A) com um método que identifica elementos em *cis* que têm papel na poliadenilação (Hu, Lutz, Wilusz, & Tian, 2005).

A poliadenilação alternativa é regulada em cada tecido em resposta à necessidade temporal, espacial ou de acordo com o desenvolvimento. Estudos usando bases de dados de ESTs encontraram um viés no uso do sinal de

poliadenilação alternativa em tecidos como testículos, útero, placenta e retina (Zhang, Lee, & Tian, 2005).

O uso de sítios alternativos de poliadenilação produz isoformas de mRNA com 3'UTRs diferentes, que pode estar conectadas à regulação da expressão gênica mediada por micro-RNA, como parte de um programa global para a proliferação celular (Zlotorynski & Agami, 2008).

Sandberg et al. desenvolveram um método quantitativo para comparar o uso de 3'UTRs alternativas e fizeram uma análise genômica da composição das isoformas de 3'UTR durante a ativação das células T. Os padrões de expressão da poliadenilação alternativa dos grupos *splicing*-dependente e o *splicing*-independente (*tandem* UTRs) foram marcadamente diferentes entre células T primárias em repouso e estimuladas.

Em experimentos durante a ativação de células T, o padrão de expressão do grupo de *splicing* não mostrou uma tendência geral após a ativação das células T, no entanto, em 86% dos genes com *tandem* UTRs a estimulação diminuiu o uso de regiões 3' estendidas (Sandberg, Neilson, Sarma, Sharp, & Burge, 2008).

Essa redução no uso de 3'UTRs estendidas não foi associada à mudanças significativas nos níveis de mRNA, indicando que o efeito é devido à uma mudança no uso do sinal de poliadenilação (Zlotorynski & Agami, 2008).

A ativação de células hematopoiéticas está associada freqüentemente com um aumento dramático na proliferação, que leva à hipótese de que a diminuição no uso do sinal de poliadenilação distal, gerando 3'UTRs menores, pode estar associado com a proliferação celular.

Os sítios de poli(A) *upstream* são usados preferencialmente em células que estão proliferando rapidamente, especialmente células tumorais, resultando em mRNAs com 3'UTRs mais curtas. Em alguns casos, esses mRNAs com 3'UTRs truncadas têm estabilidade aumentada ou são traduzidos com maior eficiência, devido à ausência de sítios de ligação de microRNAs. Isso demonstra que mudanças sutis nos mRNAs causadas pelo uso de sítios alternativos de poliadenilação, como mudança no tamanho da 3'UTR, podem ter efeitos drásticos na expressão gênica (Sandberg, Neilson, Sarma, Sharp, & Burge, 2008).

Um estudo em particular deve ser ressaltado devido a sua especificidade. Foi verificado que no gene da COX2 o sinal de poliadenilação proximal é bem

mais fraco que o sinal distal, e que seu uso é específico de acordo com o tecido onde é expresso. Foi demonstrado também que os elementos de eficiência *upstream* (USEs) ao hexâmero canônico são importantes para a definição do sítio usado. (Hall-Pogar, Zhang, Tian, & Lutz, 2005)

1.8 - Fatores de regulação da poliadenilação alternativa

1.8.1 - SNPs

Single nucleotide polymorphisms (SNPs) (Sherry, Ward, Kholodov, Baker, Smigielski, & Sirotkin, 2001) são a forma mais comum de variação genética no genoma humano, pois ocorrem aproximadamente a cada 1200 pares de bases ao se comparar dois cromossomos. Pesquisas recentes sugerem que a maioria dos SNPs atinge as regiões não-codificantes, que abrangem 95% do genoma. SNPs que atingem regiões regulatórias podem alterar a transcrição, os mecanismos de processamento do RNA e a tradução. Como as regiões 3' UTR possuem muitas seqüências regulatórias e são atingidas por SNPs, nós investigaremos o efeito dos SNPs localizados nos sinais de poliadenilação.

Para avaliar os casos em que o SNP altera a função de uma proteína, é necessário que se faça um estudo pontual, com a avaliação de cada polimorfismo sobre a atividade da proteína em questão.

Um estudo funcional com a N-acetiltransferase 1 foi realizado para investigar os efeitos funcionais dos polimorfismos e haplótipos. Dois SNPs na região codificante reduziram a atividade catalítica da proteína NAT1, bem como a quantidade de mRNA e proteína produzidos. No entanto, com a associação de uma deleção de 9 pares de bases (TAATAATAA) em sua região 3'UTR, seu nível protéico e atividade catalítica voltam a se igualar com a NAT 1 de referência (Zhu, States, Wang, & Hein, 2011).

Dois polimorfismos (1088T>A and 1095C>A) na região 3´UTR da proteína NAT1 reduzem sua atividade catalítica e diminuem seus níveis protéicos e de seu mRNA. Essa observação confere suporte biológico à associação desses polimorfismos com relatos de que causam defeitos congênitos (Zhu, States, Wang, & Hein, 2011).

1.8.2 - Antisense

Natural antisense transcripts (NATs) (Kumar & Carmichael, 1998) são seqüências de RNA complementares a outros RNAs endógenos; podem ser transcritos em *cis* a partir da fita oposta dos mRNAs codificantes, ou em *trans*, a partir um lócus distinto. Eles representam uma forma de regulação da transcrição que tem sido alvo de pesquisas nos últimos anos. Como eles têm complementaridade de seqüência, têm potencial regulatório inerente. Esses transcritos podem ser classificados como NATs se forem poliadenilados, sofrerem *splicing* ou *capping*, pois dessa forma é possível determinar sua posição genômica e orientação.

Os transcritos antissenso naturais podem regular a expressão gênica através de três principais mecanismos: Interferência transcricional, mascaramento de RNA e mecanismos desencadeados por RNA dupla fita. Neste último, o processamento dos transcritos em antissenso envolve sua hibridização com o transcrito na fita senso, quando ambos são expressos na mesma fita ao mesmo tempo. É possível que transcritos em antissenso que sobreponham genes com poliadenilação alternativa possam influenciar a regulação da expressão dos variantes de poliadenilação.

Recentemente, a análise computacional de projetos de seqüenciamento de transcriptomas, revelou que a expressão dos pares *sense-antisense* é abundante no genoma (Werner & Sayer, 2009).

No entanto, a quantidade de pares senso-antissenso no genoma humano é controversa, pois depende do critério utilizado. Yelin et al.(2003) analisaram ESTs em bancos de dados públicos e identificaram 2667 *loci* genômicos com evidência de transcrição em ambas fitas de DNA. Após análises experimentais, eles sugerem que existam aproximadamente 1600 pares de genes *sense-antisense.* Entretanto, a ocorrência deve ser ainda maior que essa estimativa, pois os métodos dependem das seqüências depositadas nos bancos de dados públicos, que ainda não estavam completos. Além disso, os genes antissenso previstos devem se sobrepor a regiões exônicas e ter tamanho suficiente para que a sobreposição seja observada pelos métodos de detecção (Yelin, et al., 2003). Segundo Galante et al. (2007), 50% dos genes humanos e de camundongos

estão envolvidos no pareamento *sense-antisense* (Galante, Vidal, de Souza, Camargo, & de Souza, 2007).

Diversos estudos independentes mostraram que no genoma humano, 5 a 10% dos genes têm um antissenso natural, em *cis*. Existe certa preferência pela complementaridade dos antissensos naturais pela 3'UTR de seus genes alvo (Sun, Hurst, Carmichael, & Chen, 2005).

Há muita heterogeneidade nas terminações 3' e 5' dos genes humanos. Muitos genes que se sobrepõe têm estruturas complexas no 5' UTR e na região promotora. Os transcritos em *cis* geralmente têm uma sobreposição maior ou perfeita com o RNA senso, enquanto que os transcritos em *trans* têm sobreposição menor ou imperfeita com o transcrito senso (Lavorgna, Dahary, Lehner, Sorek, Sanderson, & Casari, 2004) (Vanhée-Brossollet & Vaquero, 1998).

RNAs de dupla fita, de diferentes origens, têm papel importante na regulação pós transcricional. Eles são substrato para enzimas que desaminam resíduos de adenosina para inosina, dentro da estrutura polinucleotídica, resultando em desespiralamento total ou parcial. RNAs altamente modificados são degradados rapidamente ou retidos no núcleo, enquanto que os RNAs com poucas alterações são transportados para o citoplasma e produzem proteínas alteradas (Kumar & Carmichael, 1998). As longas cadeias conservadas na região 3'UTR dos mRNAs podem estar envolvidas na regulação da estabilidade do RNA através da formação de longos RNAs de dupla fita, perfeitamente pareados.

Além disso, em muitos casos a sobreposição do antissenso envolve a poliadenilação alternativa, utilizando variantes de um mesmo gene que diferem em seu 3' terminal, com a formação de pares de sense-antisense com sobreposição 3'-3' (Dahary, Elroy-Stein, & Sorek, 2005).

Inicialmente os transcritos naturais antissenso foram descritos em procariontes, onde eles estão envolvidos no controle da expressão gênica, diminuindo a expressão dos transcritos em senso, além de outras funções biológicas, como transposição e controle temporal do desenvolvimento, etc. A regulação pelo antissenso é encontrada em procariontes e até mesmo em archaebacteria, o que sugere que essa regulação também ocorra em eucariontes (Vanhée-Brossollet & Vaquero, 1998). Há muitos estudos que mostram que a introdução de oligonucleotídeos artificiais complementares ao RNA do gene de interesse pode inibir sua expressão. Esses estudos mostraram que ácidos nucléicos antissenso

podem modular a expressão gênica também em eucariontes (Vanhée-Brossollet & Vaquero, 1998).

Há diversos exemplos na literatura em que a alta expressão de transcritos em senso e antissenso não ocorrem concomitantemente: No caso dos transcritos da dopa decarboxilase em Drosophila; nos transcritos do vírus HSV-1 em fase latente e fase infecciosa; no gene eb4-psv durante o desenvolvimento de Dictyostelium; nos transcritos de colágeno α1 em condrócitos sob tratamento farmacológico, entre outros (Vanhée-Brossollet & Vaquero, 1998).

Esses transcritos podem estar envolvidos no silenciamento gênico, quando hibridizam como DNA codificante ou com a regulação gênica, envolvendo a degradação do transcrito senso (RNA de interferência). (Lavorgna, Dahary, Lehner, Sorek, Sanderson, & Casari, 2004). Podem também ter um papel no *imprinting* genômico, na inativação do X, no *splicing* alternativo e na edição de RNA.

Estudos sobre a regulação dos transcritos em antissenso dos genes supressores de tumor de Wilms, EIF2-alfa e myc mostram que quando o transcrito em senso está mais expresso, a transcrição do antissenso diminui; inversamente quando o antissenso diminui, a transcrição em senso aumenta (Lipman, 1997). A razão senso/antissenso aumenta com a maior expressão do gene e vice-versa. A partir desses exemplos, surgiu um modelo de acoplamento direto entre a regulação transcricional e a estabilidade do mRNA, no caso da rápida degradação dos duplexos senso-antissenso (Lipman, 1997).

1.9 - Doenças mendelianas relacionadas à poliadenilação defectiva

A doença da hemoglobina H está relacionada à perda do sinal canônico de poliadenilação (AATAAA), que leva à redução da alfa-hemoglobina, ocasionando um tipo de anemia.

Algumas mutações que afetam o sinal de poliadenilação associadas a essa doença são AATAAA → AATGAA (Losekoot, et al., 1991) e AATAAA → AATAAG no gene da Globina alfa2 humana (Higgs, Goodbourn, Lamb, Clegg, Weatherall, & Proudfoot, 1983), que causa uma expressão anormal, pois a transcrição prossegue além do sítio poli(A) perdido. Mutações combinadas, afetando o sinal de poliadenilação e grandes deleções também são associadas com essa doença (Prior et al, 2007). (Chatterjee & Pal, 2009)

Pseudodeficiência da arilsulfatase A

http://www.ncbi.nlm.nih.gov/omim/607574

A Leucodistrofia Metacromática é uma desordem metabólica causada pela deficiência da arilsulfatase A (cerebroside-3-sulfate 3-sulfohydrolase). A deficiência dessa enzima já foi observada em indivíduos saudáveis, uma condição na qual o termo pseudodeficiência foi introduzido. A incidência dessa doença herdada recessivamente é estimada em 1:40.000. A deficiência da arilsulfatase A (ARSA) leva ao acúmulo intra-lisossômico de sulfato cerebrosídio, principalmente no sistema nervoso central, causando uma desmielinização progressiva, eventualmente levando à morte do paciente.

O alelo da pseudodeficiência da ARSA contém duas alterações de seqüência: um defeito na poliadenilação e uma substituição de amino ácido. As duas são transições de A para G: uma mutação de asn350 para serina no exon 6, causando a perda de um sítio de glicosilação. A outra transição de A para G ocorre no exon 8, na extremidade 3'UTR do gene, causando a perda de um sinal de poliadenilação. Ela muda o primeiro sinal de poliadenilação *downstream* ao *stop codon* de AATAAC para AGTAAC, segundo os autores. Este causa uma severa deficiência de uma espécie de mRNA de 2,1kb, o que explica a síntese

diminuída de ARSA em fibroblastos pseudodeficientes (Gieselmann, Polten, Kreysing, & von Figura, 1989).

Foi demonstrado que a ARSA sintetizada em indivíduos com pseudodeficiência é menos abundante e seu tamanho é menor, em relação ao normal. A diferença em tamanho da ARSA foi atribuída à glicosilação alterada, mas a causa da atividade atenuada é atribuída às alterações na poliadenilação e na glicosilação.

O efeito combinado da redução do mRNA da ARSA devido ao defeito na poliadenilação e a diminuição de sua atividade, além do direcionamento aberrante da proteína ARSA para os lisossomos, reduz sua atividade no homozigoto para aproximadamente 8% do normal (Harvey, Carey, & Morris, 1998).

No entanto a redução na atividade da arilsulfatase foi previamente atribuída ao defeito na poliadenilação, que reduz a quantidade de mRNA da ARSA em aproximadamente 90%. Uma amplificação do DNA genômico e hibridização com oligonucleotídeos alelo-específicos detectou as mutações em quatro indivíduos não relacionados, com pseudodeficiência na ARSA (Gieselmann, Polten, Kreysing, & von Figura, 1989).

Talassemia

http://www.ncbi.nlm.nih.gov/omim/141900

As talassemias possuem diversos fenótipos, determinados pela extensão da hemólise e eritropoiese ineficiente; diversas mutações no sinal de poliadenilação da beta-globina podem causar talassemia.

Um gene de beta-globina clonado de uma pessoa com beta-talassemia continha uma substituição de T para C no sinal de poliadenilação AATAAA, tornando-o AACAAA, com perda de função. A análise por Northern blotting revelou uma espécie de RNA de 1500 bases de comprimento. Foi determinado que a extremidade 3' desse RNA está localizada a 900 nucleotídeos *downstream* ao sítio normal de adição da cauda poli(A) (Orkin, Cheng, Antonarakis, & Kazazian, 1985).

Em pacientes israelenses de beta+talassemia foram encontradas mudança de AATAAA para AATAAG na porção 3'UTR do gene da beta-hemoglobina, assim como uma deleção de 5 pares de bases (AATAAA----A----). Essas mutações no sinal de poliadenilação foram avaliadas quanto ao mecanismo pelo qual levam ao

fenótipo de talassemia. A análise de RNA derivado de sangue periférico demonstrou a presença de espécies elongadas e poliadeniladas de RNA nos pacientes que possuíam essas mutações. No entanto, a deleção dos 5 pares de bases impediu completamente a clivagem no sítio normal e resultou no fenótipo da beta+talassemia (Rund, Dowling, Najjar, Rachmilewitz, Kazazian, & Oppenheim, 1992)

1.10 - Expressão gênica em Gliomas

Os gliomas, um tipo de tumor do sistema nervoso central, são representados em 76% por astrocitomas. A Organização Mundial de Saúde classifica os tumores astrocíticos em 4 graus de malignidade, de acordo com suas características histológicas.

O grau I, astrocitoma pilocítico, não é invasivo, é relativamente circunscrito e tem crescimento lento. O grau II, astrocitoma difuso, tem alta diferenciação celular, mas crescimento lento. O astrocitoma anaplástico, grau III, possui atipia nuclear e uma alta taxa de proliferação. O tumor mais agressivo dentre os astrocitomas é o glioblastoma, grau IV, que possui a maior taxa de proliferação e capacidade de invasão (Louis, Ohgaki, Wiestler, Cavenee, Burger, & al, 2007).

O glioblastoma (GBM) representa 17% de todos os tumores cerebrais primários, e 54% de todos os gliomas.

A transformação das células neuroepiteliais em tumorais é um processo desencadeado pela aquisição de alterações genéticas seqüenciais. O glioblastoma tem um alto número de alterações genéticas, e pode ser diferenciado de outros gliomas pelo seu padrão de expressão (Holmberg, et al., 2011).

A presença de características de células tronco em células de gliomas levanta a possibilidade de que mecanismos que promovem a manutenção e auto-renovação das células tronco exerçam o mesmo papel em células tumorais. Foi demonstrado que em gliomas de alto grau há a expressão de marcadores de células tronco neurais, como SOX2, Oct4 e Nanog, e que essa expressão é correlacionada com o aumento da malignidade (Holmberg, et al., 2011).

As diferenças na expressão gênica podem ser observadas na progressão de gliomas de diferentes graus. Um estudo sobre o fator de transcrição SOX2

mostrou uma correlação positiva entre sua expressão e o grau de malignidade dos gliomas. As áreas com maior expressão de SOX2 coincidiram com as áreas de maior proliferação celular em glioblastomas e com células não diferenciadas (Annovazzi, Mellai, Caldera, Valente, & Schiffer, 2011).

Em um segundo estudo, com cultura de células de glioma C6, o mRNA de cat-1 é superexpresso após a depletação de aminoácidos. Seu mRNA é mantido em altos níveis mesmo quando ele não é mais transcrito, o que sugere uma regulação pós-transcricional. O mRNA *full-length* (7.9 kb) aumentou 5 vezes nas células depletadas de aminoácidos. No entanto um transcrito de 3.4kb resultante do uso de um sítio alternativo de poliadenilação não foi induzido, sugerindo que o mRNA do cat-1 foi estabilizado por seqüências de RNA que agiram em *cis* no 3'UTR. O gene do cat-1 é sujeito à regulação adaptativa de acordo com a disponibilidade de aminoácidos. A depleção de amino ácidos inicia eventos moleculares que levam ao aumento da estabilidade do mRNA de cat-1, através do uso de um sítio alternativo de poliadenilação (Aulak, et al., 1999).

O gene Hiwi, que tem papel na auto-renovação de células tronco, está significativamente super expresso em alguns tipos de câncer. Foi observado que este gene está especificamente super expresso em gliomas, e que sua expressão é aumentada com a malignidade do tumor. Pacientes com alta expressão de Hiwi tiveram um pior prognóstico que pacientes com menor expressão (Sun, et al., 2011).

2 – Objetivos

O objetivo da presente tese é observar a presença da poliadenilação alternativa nos transcritos humanos, provenientes de diferentes tipos celulares e técnicas de seqüenciamento. Avaliaremos e quantificaremos os hexâmeros sinais de poliadenilação nos mRNAs e em suas posições relativas no genoma humano. Rastrearemos a presença de SNPs (single nucleotide polymorphisms) em cada sinal e sua localização preferencial em genes com eventos de poliadenilação alternativa.

Numa segunda etapa, o objetivo é avaliar se existe expressão diferencial entre variantes de poliadenilação e se o uso dos sinais de poliadenilação varia de acordo com o tamanho e posição genômica do sítio utilizado em cada transcrito. Para tanto, foram avaliadas seqüências de referência, bem como seqüências provindas de linhagens celulares de câncer de mama e microarrays de astrocitomas de diferentes graus.

O uso dos diferentes dados tumorais possibilitará também avaliar se há diferença de expressão entre variantes de poliadenilação de alguns genes.

3 - Materiais e métodos

Foram escritos programas na linguagem Perl para utilizar os métodos descritos abaixo.

3.1 - Seqüências primárias

Utilizaremos as seqüências de RefSeqs, mRNAs *full length* e ESTs de bancos de dados públicos, bem como a seqüência genômica, como fontes de dados primárias.

As seqüências de RefSeqs, mRNAs e de ESTs (Boguski MS, 1993) foram obtidas do site da Universidade da Califórnia, em Santa Cruz (http://genome.ucsc.edu/) (Kent WJ, 2002). Elas são atualizadas semanalmente, e neste trabalho utilizamos a versão de 13 de janeiro de 2010. Essa versão possui 32.438 seqüências de RefSeqs, 364.161 seqüências de mRNAs.

Doravante, neste trabalho, essas seqüências serão denominadas apenas de transcritos, exceto quando explicitado.

```
ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/
File: mrna.fa.gz       168522 KB   01/13/2010       02:56:00 PM
File: refMrna.fa.gz    34000 KB    01/13/2010       02:58:00 PM
This directory contains the Feb. 2009 assembly of the human genome
(hg19,GRCh37 Genome Reference Consortium Human Reference 37
(GCA_000001405.1)), as well as repeat annotations and GenBank
sequences.
```

A seqüência genômica referência utilizada foi a versão de fevereiro 2009, montagem hg19 do genoma humano (GRCh37 Genome Reference Consortium Human Reference 37), também obtida através do site do UCSC. (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/)

3.2 - Identificação dos mRNAs poliadenilados

As pesquisas são iniciadas com todas as seqüências atuais 3' *full-lenght* dos mRNAs humanos, ou seja, seqüências completas contendo a cauda poli(A), quando existentes.

A cauda poli(A) é identificada com o seguinte critério, para mRNAs e RefSeqs: ela deve ter 5 ou mais As no final 3' da seqüência, podendo aceitar um erro, ou seja, uma base diferente de A. Nesse caso a menor cauda a ser aceita terá 6 bases.

Inicialmente os transcritos foram separados por tipo (RefSeq, mRNA, EST) e pela presença da cauda poli(A). Os transcritos tinham que ter pelo menos 5 As na extremidade 3'. Para o corte da cauda, upstream aos 5 As, era necessário que tivessem 8 As em cada janela de 10 bases. Toda a seqüência foi cortada, até o ultimo A *upstream* à janela de 10 bases. Era considerado poliadenilado todo transcrito que tivesse pelo menos 5 As cortados, ou seja, 1 erro (C,G ou T) a cada 6 bases (A).

Exemplo de seqüências aceitas como cauda poli(A)

... 16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	← 3'	
... C	T	G	A	C	A	A	A	A	A	A	A	A	A	A	A		aceito
... C	T	G	A	C	T	A	T	C	G	A	A	A	A	A	A		aceito
... C	T	G	A	C	T	A	T	C	G	A	A	A	A	A	C		aceito
... C	T	G	A	C	T	A	T	C	G	A	A	A	A	T	A		aceito
... C	T	G	A	C	T	A	T	C	G	A	A	A	G	A	A		aceito
... C	T	G	A	C	T	A	T	C	G	A	A	T	A	A	A		aceito
... C	T	G	A	C	T	A	T	C	G	A	T	A	A	A	A		aceito
... C	T	G	A	C	T	A	T	C	G	A	A	A	T	A	T		não aceito
... C	T	G	A	C	T	A	T	C	G	G	A	A	T	A	A		não aceito
... C	T	G	A	C	T	A	T	C	G	A	A	A	T	G	A		não aceito

Figura 9: Critério de aceitação das últimas seis bases da extremidade 3' dos mRNAs poliadenilados.

3.3 - Alinhamento dos transcritos com o genoma

As posições de alinhamento dos mRNAs e RefSeqs foram obtidas da seção de tabelas do site do UCSC, e foram feitos através do BLAT. De acordo com as posições obtidas, pudemos verificar quais seqüências de cDNA alinham em apenas uma posição no genoma e quais alinham em duas ou mais posições. A partir dessas posições os transcritos foram realinhados para determinar a posição exata de suas bordas exon-intron e observar as bases alinhadas ao genoma.

Para realizar o alinhamentos dos RefSeqs e mRNAs foi utilizado o programa sim4 (Florea, Hartzell, Zhang, Rubin, & Miller, 1998) com os parâmetros A=4 N=1 P=0, ou seja, o sim4 imprime toda a seqüência alinhada, não corta as caudas poli(A) antes de realizar o alinhamento e reconhece as bordas entre os exons e introns (*splice-site*).

Na Figura 10 podemos observar um exemplo de alinhamento com o programa sim4. É possível observar a delimitação das bordas exon-intron de forma numérica e base a base. Além disso, a cauda poli(A) não se alinha no genoma, o que indica que é uma cauda poli(A) real e não *internal priming*.

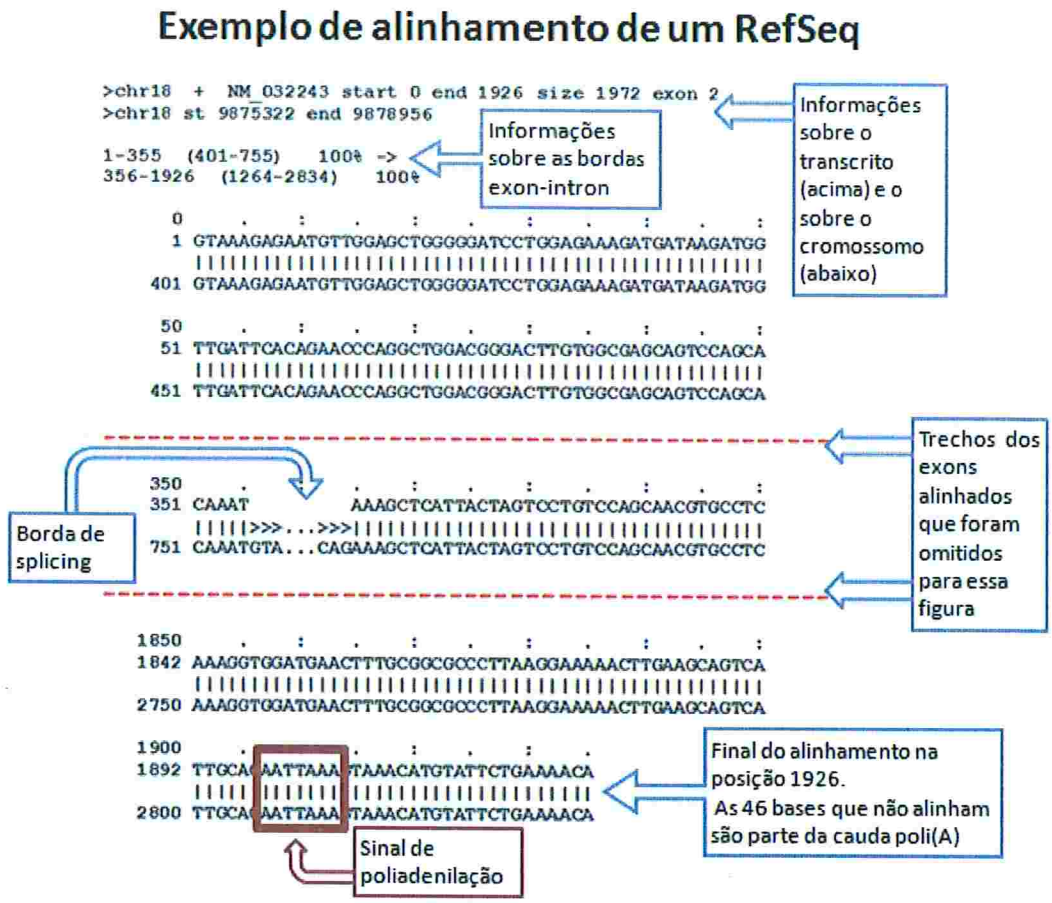


Figura 10: Exemplo de alinhamento com o programa sim4. A cada bloco de texto, os dados sobre o transcrito estão na linha de cima e sobre o cromossomo na linha de baixo.

Depois de encontrados os poliadenilados, é necessário retirar os casos de *Internal Priming*. O *Internal Priming* (Lee, Park, & Tian, 2008) ocorre quando o cDNA seqüenciado tem uma aparente cauda poli(A), mas a mesma está presente no genoma, ou seja, é um trecho rico em A que alinha ao genoma, por razões como a retrotransposição, por exemplo.

A seqüência dos transcritos foi mantida intacta para o alinhamento. A seqüência genômica foi cortada 400 bases *upstream* e 800 bases *downstream* em relação ao alinhamento previsto pelo BLAT. Assim, a eficiência do programa é melhor, pois a seqüência genômica é maior que a do transcrito e é possível observar se as caudas alinham com o genoma, representando casos de *internal priming*. É muito comum que a cauda poli(A) do transcrito alinhe com o genoma

em uma região distante do alinhamento do início da seqüência, o que não representa um caso de *internal priming*.

O próximo passo é verificar a extremidade 3' dos mRNAs selecionados. Para isso cortamos a cauda poli(A), conforme descrito acima, e verificamos se a extremidade 3' obtida é igual à do alinhamento com o genoma, comparamos as posições genômicas das extremidades 3'. Como a clivagem em geral é numa base de adenina, o mRNA que teve sua cauda cortada não terá a última base, por exemplo.

3.4 - Filtros dos alinhamentos

Uma etapa necessária após alinhar as seqüências dos transcritos com o genoma é a seleção dos alinhamentos que apresentam a melhor qualidade. Para isso foi necessário identificar e retirar das análises seguintes os alinhamentos descontínuos, ou seja, casos em que um trecho do mRNA transcrito não é alinhado ao genoma, seja ele no início, meio ou final da seqüência. Foram retirados também alinhamentos cujo valor da identidade de um ou mais exons era menor que 95%.

Os alinhamentos de seqüências que apresentaram *internal priming* também foram retirados das análises posteriores, ou seja, aqueles cuja cauda poli(A) alinha-se com o genoma.

No caso de transcritos que alinham em dois ou mais lugares no genoma, de acordo com o alinhamento do BLAT, foi realizado o alinhamento em todas as posições genômicas com o sim4. Foi calculado um *score* para a escolha do melhor alinhamento. O *score* é dado pela identidade média dividida pela cobertura, onde cobertura é o número de bases alinhadas. Selecionamos então a posição genômica onde cada transcrito obteve o maior *score*.

3.5 - Localização do hexâmero sinal

O hexâmero sinal de poliadenilação está localizado na região entre 10 a 30 nucleotídeos da extremidade 3' do transcrito, sem a cauda poli(A). Os dois

hexâmeros canônicos descritos na literatura são AAUAAA e AUUAAA. Além desses, há substituições de uma base que também são usados como sinais pela maquinaria de poliadenilação. Outros grupos confirmam a existência de sinais alternativos em uma fração significante dos 3'UTRs (Tabaska & Zhang, 1999) (Beaudoing & Gautheret, 2001).

Buscaremos nas seqüências de mRNA os hexâmeros que são utilizados com eficiência pela maquinaria de poliadenilação, baseado em dados experimentais da literatura (Beaudoing, Freier, Wyatt, Claverie, & Gautheret, 2000) (Sheets MD, 1990).

Hexâmero	Eficiência da Poliadenilação
AATAAA	100
ATTAAA	77
AGTAAA	29
CATAAA	18
TATAAA	17
GATAAA	11
ACTAAA	11
AATACA	11
AATATA	10
TTTAAA	-

Tabela 1: Dados experimentais sobre a eficiência da clivagem, obtido de (Sheets MD, 1990), e (Tian, Hu, Zhang, & Lutz, 2005)

A partir da literatura sobre a descrição da localização dos hexâmeros (Beaudoing, Freier, Wyatt, Claverie, & Gautheret, 2000), limitamos a busca de sinais à região entre -10 e -36 bases do final da seqüência transcrita, onde começa a cauda poli(A) (Sheets MD, 1990) e fizemos a busca pela seqüência exata dos 10 hexâmeros citados acima.

Os hexâmeros AAUAAA e AUUAAA, entre os outros que utilizamos neste trabalho podem ser localizados ao longo de toda a extensão dos transcritos. O hexâmero sinal AATAAA é muito comum nas seqüências genômicas, pode aparecer uma vez a cada 4096 bases (Tabaska & Zhang, 1999), caso fosse contabilizado de forma aleatória. Para ilustrar esse fato, foi realizada a busca pelo hexâmero AAUAAA por toda a seqüência dos RefSeqs. Dentre os 16.668 RefSeqs que possuíam sinal de poliadenilação, foram encontrados 21.945 AAUAAA por toda a seqüência, sendo que 13.206 somente na região funcional do sinal de poliadenilação. Assim, temos que aproximadamente 40% dos hexâmeros AAUAAA estão localizados fora da região funcional. Da mesma maneira, foram encontrados 9.881 hexâmeros AUUAAA ao longo da seqüência dos RefSeqs, sendo que aproximadamente 66% fora da região funcional e 3.345 hexâmeros AUUAAA somente na região funcional do sinal de poliadenilação.

3.6 - Busca dos hexâmeros sinais no alinhamento

A partir dos arquivos de alinhamento dos transcritos com o genoma, gerados pelo programa sim4, é possível observar com precisão a posição genômica dos hexâmeros, bem como se há algum *mismatch* nessa região. O hexâmero sinal é mais freqüente na forma de AAUAAA ou de AUUAAA, mas há diversos variantes de uma base que também são usados, *in vitro*, com menor eficiência de clivagem e poliadenilação. Em nossas buscas, usamos também 8 desses outros hexâmeros. (Sheets MD, 1990)

3.7 - Agrupamentos dos transcritos em clusters

Para agrupar transcritos mapeados no mesmo lócus gênico e na mesma fita, utilizamos um método baseado nas posições das bordas dos exons. Esse protocolo foi escolhido após o estudo e teste de outras maneiras de agrupamento.

Para cada transcrito de um mesmo cromossomo, foi observada sua fita, senso ou antissenso, para que não sejam agrupados transcritos de fitas opostas

como se fossem do mesmo *cluster*. Os exons são separados de acordo com sua posição genômica de inicio e fim. O agrupamento foi feito comparando-se a posição genômica dos exons de cada transcrito. Foram agrupados no mesmo *cluster* transcritos que possuíam ao menos um exon em comum.

- **Genes**

Neste trabalho utilizamos as siglas dos genes de acordo com o Mammalian Gene Collection (MGC) (http://mgc.nci.nih.gov/Info/). O MGC fornece acesso a clones de cDNA codificante, de tamanho completo, com validação da seqüência de genes de humanos, camundongos e ratos. Segundo detalhes do projeto do Mammalian Gene Collection, é feito o alinhamento dos mRNAS contendo ORF (open reading frame) completa contra o genoma usando BLAT (Kent, 2002). Somente os alinhamentos com 95% de identidade com a seqüência genômica são mantidos.

Os nomes de genes dados pelo MGC foram obtidos a partir do track do site do UCSC, correspondentes a cada RefSeq a partir da tabela hg19.refGene e as siglas correspondentes aos RefSeqs e mRNAS conjuntamente, a partir da tabela hg19.mgcGenes, obtidas da seção "Tables" do site do UCSC (http://genome.ucsc.edu/cgi-bin/hgTables)

As duas tabelas somam 22.136 nomes de genes distintos segundo o MGC.

3.8 - Identificação dos genes com eventos de poliadenilação alternativa

Após o agrupamento e ordenação dos transcritos pertencentes a cada gene, pode-se prosseguir para a identificação dos genes que possuem variantes de poliadenilação. Cada grupo de transcritos é então ordenado, em ordem crescente, do menor sítio de poliadenilação para o maior. Assim, identifica-se o transcrito mais curto e o mais longo, dentro de um mesmo gene.

Primeiramente são identificados e separados os últimos exons de cada transcrito. Confere-se a fita novamente, para que transcritos de fitas opostas não

sejam agrupados como variantes de poliadenilação. Os últimos exons, de cada gene, de cada fita são então separados em arquivos diferentes, e ordenados, do menor para o maior.

Para separar os transcritos que representam variantes de poliadenilação, são examinados o primeiro e o último de cada grupo. Se a diferença entre as terminações for maior que 25 bases, pode-se afirmar que esses genes apresentam eventos de poliadenilação alternativa.

O critério de selecionar as diferenças maiores que 25 bases foi adotado porque a localização do sinal de poliadenilação é -10 a -36 bases do sítios de poliadenilação, ou seja, se o sinal estiver localizado numa região de 25 bases justaposta à primeira, estará formando um novo sítio de poliadenilação, uma nova posição de clivagem.

3.9 - Identificação dos eventos de poliadenilação alternativa devido ao splicing alternativo

Para identificar quais eventos de poliadenilação eram devidos ao *splicing*, usamos o critério de verificar qual era o exon terminal de cada seqüência. Se a "borda" aceptora do sitio de *splicing* do último exon é diferente de um transcrito para outro, do mesmo gene, está caracterizada a poliadenilação alternativa devida ao *splicing*. Se a última borda aceptora do sítio de *splicing* é a mesma, o que ocorre é um evento característico de poliadenilação alternativa. No entanto, para os fins deste trabalho, os dois casos serão tratados como poliadenilação alternativa, exceto onde explicitamente mencionado.

3.10 - SNPs no sinal de poliadenilação

Para a procura de SNP no sinal de poliadenilação, foi utilizado o genoma referência humano, na montagem 37 (hg.19, fev.2009), modificado de acordo com as regras IUPAC para nucleotídeos ambíguos, em cada substituição de única

base (single-base) anotada pelo dbSNP versão 131. Os arquivos FASTA foram obtidos em:

ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/snp131Mask/.

A Tabela 2 resume a notação estendida IUPAC (International Union of Pure and Applied Chemistry). Os códigos de ambigüidades são inseridos devido a adventos no processo de seqüenciamento, quando não se pode identificar unicamente um monômero. Para os ácidos nucléicos, cada combinação de duas bases recebe um símbolo de acordo com suas propriedades químicas, e de três bases recebe uma notação inversa (ex: B equivale às três bases diferentes de A).

Símbolo	Significado	Origem (ou propriedade química)
A	A	Adenina
C	C	Citosina
G	G	Guanina
T	T	Timina
M	AC	Amino
R	AG	Purina
W	AT	Interação fraca (do inglês:weak)
S	CG	Interação forte (do inglês: strong)
Y	CT	Pirimidina (do inglês: Pyrimidin)
K	GT	Cetona (do inglês: Keto)
V	ACG	Não-T
H	ACT	Não-G
D	AGT	Não-C
B	CGT	Não-A
N	GATC	Qualquer nucleotídeo

Tabela 2: Nomenclatura IUPAC para nucleotídeos.

Fonte: http://www.mit.edu/afs/sipb/project/seven/arch/sun4x_57/lib/python2.0/site-packages/Bio/Data/IUPACData.py

Foi então observada a posição do SNP de cada gene, a fim de verificar se o SNP mapeado no sinal estava no transcrito mais longo ou mais curto do mesmo gene. Foi observado também qual o resultado da mudança de base dada pelo SNP sobre a seqüência do sinal.

3.11 - Sinais da região *downstream* aos transcritos

Fizemos o estudo da região *downstream* dos transcritos para eliminar os demais casos de *internal priming* não identificados no alinhamento e para verificar se existe a região rica em G e U em todos os transcritos poliadenilados.

Primeiramente os sítios poli(A) são uniformizados, retirando-se a heterogeneidade entre 10 bases *upstream* e 10 bases *downstream* ao sítio poli(A) de cada transcrito. Essa estratégia uniformiza o sítio poli(A) provindo de cada sinal para uma única base. Foi observado que a heterogeneidade dos sítios poli(A) é encontrada principalmente em até 10 bases antes ou depois da posição prevista, sabendo-se que o sinal de poli(A) se encontra entre -10 e -36 bases do sítio de clivagem (Iseli, et al., 2002).

Em seguida é feita a análise das 50 bases *downstream* aos sítios poli(A), na seqüência genômica. Nessa etapa é importante separar os genes da fita senso e antissenso de cada cromossomo, para que a seqüência de 50 bases seja realmente *downstream* ao final da transcrição. Na Figura 11 podemos visualizar a freqüência das 50 bases *downstream* aos transcritos.

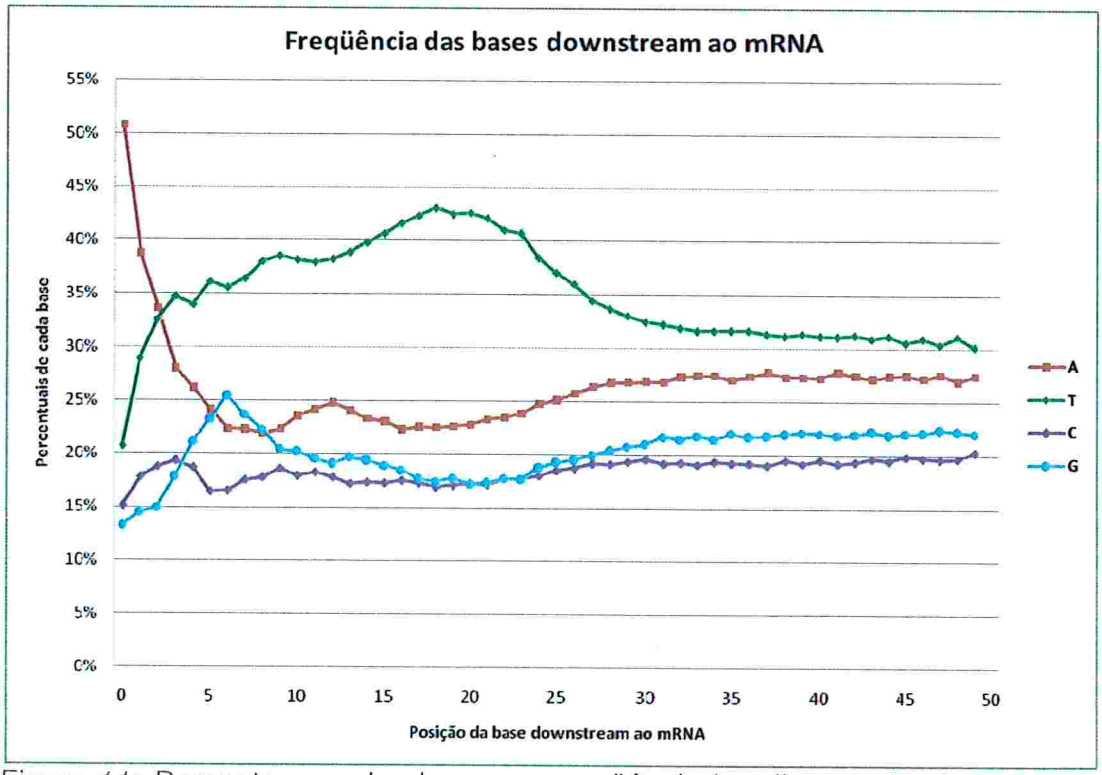


Figura 11: Porcentagens das bases na seqüência imediatamente *downstream* à extremidade 3' dos mRNA

Na ausência de um motivo consenso para o sinal *downstream*, desenvolvemos um *score* baseado na informação da literatura sobre a eficiência da poliadenilação, a partir de informações sobre as seqüências da região *downstream* rica em G e U (Cañadillas & Varani, 2003) (Zarudnaya, Kolomiets, Potyahaylo, & Hovorun, 2003) (Mandel C.R., 2007).

Foi atribuído o *score* de 20 para cada dinucleotídeo TT encontrado na seqüência de 50 bases; e um *score* de 5 para cada dinucleotídeo GT, não sobreposto ao primeiro. Caso a seqüência avaliada possuísse 5 bases A consecutivas ou 6 bases, sendo que 5 dessas são um A, era atribuída a notação de *internal priming*. Assim, atribuímos *score* entre zero e 360 para as seqüências *downstream* ou a notação de *internal priming*.

3.12 - Identificação de pares *sense-antisense*

Para analisar a formação de pares de genes em *antisense* com envolvimento de um gene com poliadenilação alternativa, foram separados todos os genes com eventos de poliadenilação alternativa e em seguida foram comparadas as posições genômicas com o total de genes com transcritos poliadenilados e com o total de genes da tabela UCSC known genes.

Para tanto, foram separados os trechos mais longos de cada gene, baseados na posição genômica do final do transcrito mais curto e o final do transcrito mais longo. Na Figura 12 podemos observar como ocorre a sobreposição entre o transcrito em antissenso e o maior transcrito do cluster em senso (ou vice-versa).

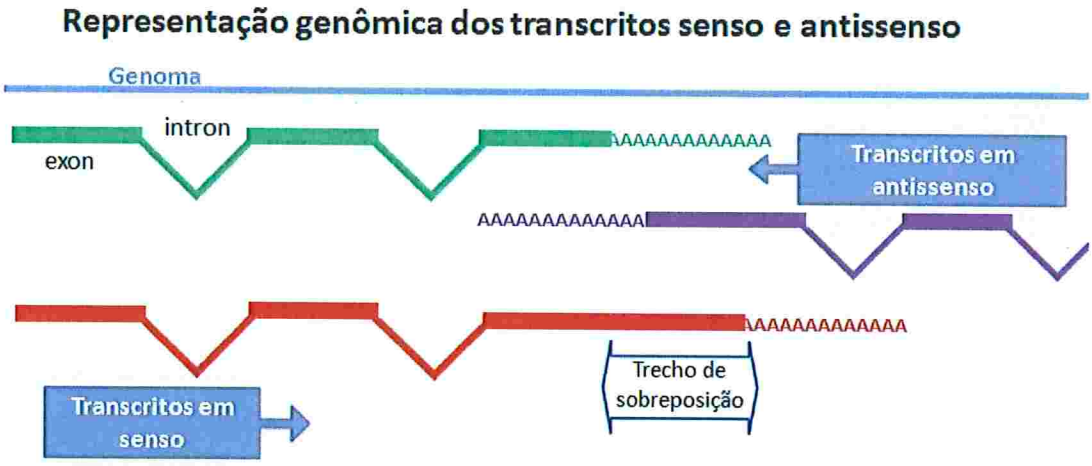


Figura 12: O transcrito em verde representa um variante mais curto e o vermelho representa um variante de poliadenilação mais longo, ambos em senso. O trecho de sobreposição é representado pela diferença entre as terminações 3' dos variantes mais curto e mais longo.

Primeiramente foram identificados todos os possíveis trechos de sobreposição dos genes com poliadenilação alternativa, ou seja, o trecho entre o maior e o menor transcrito de cada gene. Para avaliar a presença de sobreposições, foram avaliados os transcritos em antissenso que terminam na posição genômica do possível trecho de sobreposição.

3.13 - Expressão dos genes com variantes de poliadenilação por microarrays de diferentes graus de gliomas

CodeLink™ Expression Analysis	4.1.0.29054
Analysis Report Name:	Expression Report
Image File Name:	T00339848_160108_635.tif
Analysis Date and Time:	1/16/2008 3:06:07 PM
GEN file name:	EXP294X192-912.22.GEN
Normalization Method:	Median-Normalization
Computation Method for Median:	Spot Based

Com o objetivo de mostrar a expressão diferencial dos variantes de poliadenilação de um mesmo gene, utilizamos um conjunto de 19 microarrays de tecidos cerebrais humanos. Os microarrays foram feitos com a plataforma

Codelink (CodeLink™ Expression Analysis), em *arrays* de 55 mil *probes*, e incluem as seguintes amostras:

Três amostras de córtex de cérebro normal, três amostras de substância branca de cérebro normal, quatro amostras de astrocitoma pilocítico, seis amostras de astrocitomas de grau II e três amostras de glioblastomas.

As amostras de cérebro normal foram extraídas em cirurgias de esquizofrenia e as amostras tumorais foram obtidas na cirurgia de extração do tumor.

- **Identificação da posição genômica das probes:**

Pudemos verificar que a grande maioria das *probes* dos *arrays* da Codelink® estão na região 3' dos genes, o que facilita sua utilização em nossas análises. Utilizamos a posição das *probes* fornecidas pelo Ensembl e selecionamos apenas as que são mapeadas em apenas uma posição genômica e que estão em genes com pelo menos um transcrito poliadenilado.

O método de normalização utilizado foi o fornecido pela plataforma Codelink®, *Median-Normalization*, e o método computacional para a mediana é o *SpotBased*.

As *probes* localizadas em regiões intrônicas foram descartadas, bem como transcritos com extremidade 5' divergente.

Em seguida agrupamos as duas ou mais *probes* de cada gene, para verificar se elas são capazes de identificar dois transcritos distintos, e mais além, se elas são também capazes de diferenciar os variantes de poliadenilação.

3.14 - Procedência de dados

Alguns conceitos de procedência de dados foram utilizados para organização dos dados e da linha de trabalho, assim como para a identificação dos erros nos processos de forma mais eficiente (Tan, 2004; Missier, Embury, & Stapenhurst, 2008).

4 - Resultados

4.1 - Identificação dos mRNAs poliadenilados a partir das seqüências primárias

Todas as 32.438 seqüências de RefSeqs e 364.161 mRNAs obtidas dos bancos de dados públicos foram submetidas às regras descritas anteriormente para a identificação de transcritos poliadenilados. Obtivemos 18.744 RefSeqs e 70.222 mRNAs poliadenilados. Para as análises seguintes, utilizamos somente as seqüências poliadeniladas. Notamos aqui que, embora a cauda poli(A) descrita na literatura possua em média 200 bases, nas seqüências de RefSeqs e mRNAs pudemos observar, em média, 22 bases na cauda poli(A), devido tanto a limitações do processo de seqüenciamento quanto às convenções dos bancos de dados públicos.

4.2 - Alinhamentos dos transcritos com o genoma humano

Para alinhar os transcritos poliadenilados ao genoma, utilizamos as posições de alinhamento dadas pelo site da UCSC (Kent WJ, 2002), na seção de tabelas. A partir dessas coordenadas, utilizamos o programa sim4 para realinhar as seqüências dos RefSeqs e mRNAs à região genômica alvo, de forma a observar cada nucleotídeo alinhado.

Do total de 18.744 RefSeqs poliadenilados, 18.411 (98,2%) possuíam uma única posição de alinhamento com o genoma. Do total de 70.222 mRNAs poliadenilados, 67.679 (96,4%) alinharam-se em apenas uma posição no genoma. As demais 2.876 seqüências de cDNA possuem duas ou mais posições de alinhamento ao genoma. Para decidir qual a posição mais fidedigna, utilizamos o seguinte critério: após o alinhamento com o programa sim4 e atribuímos uma nota para todos os alinhamentos da mesma seqüência. Esta nota foi calculada a partir da identidade média dividida pela cobertura, ou seja, quantas bases são iguais, dividido pelo numero de bases alinhadas. Selecionamos apenas o alinhamento de maior nota cada seqüência.

Ao final dos alinhamentos, obtivemos um total de 128.258 seqüências alinhadas ao genoma, que compreendem tanto as 86.090 seqüências que se alinham em apenas uma posição no genoma, como os diversos alinhamentos das 2.876 seqüências que se alinham em duas ou mais posições no genoma.

4.3 - Filtros dos alinhamentos

Aproximadamente sete mil transcritos foram excluídos das análises subseqüentes, pelos critérios de: qualidade do alinhamento menor que 95% ou alinhamento descontínuo em um ou mais trechos. Foram excluídos também os transcritos cujo número de exons obtido pelo alinhamento com o sim4 era diferente do alinhamento com o BLAT, oferecido pelo site do UCSC.

Além disso, foram excluídos também os transcritos com casos de *Internal priming*, identificados nesta etapa pelo alinhamento da cauda poli(A) com o genoma.

4.4 - Agrupamentos dos transcritos em genes

Os 81.664 RefSeqs e mRNAs poliadenilados, selecionados a partir de seu alinhamento adequado, foram agrupados em 32.136 *clusters*.

Em seguida, utilizamos os nomes de genes de acordo com o MGC, conforme descrito em "Métodos", no qual existem 22.136 nomes de genes distintos. No entanto, os genes que contêm transcritos poliadenilados, segundo os critérios utilizados neste trabalho, totalizaram 15 mil.

Os clusters que tinham dois nomes foram manualmente curados; isso ocorria quando dois genes justapostos compartilhavam um exon, quando transcritos quiméricos se sobrepunham a dois genes ou quando ocorrera a atualização de um nome de gene em apenas alguns transcritos deste. Dessa forma, pudemos atribuir nomes de genes MGC a 15 mil *clusters*, correspondentes a 34.614 dos mRNAs selecionados até esta etapa. Aos demais 17.134 *clusters* que não possuíam nome segundo o MGC, foram atribuídos identificadores arbitrários. Na maioria dos casos esses *clusters* possuíam apenas um mRNA.

4.5 - Identificação dos genes com eventos de poliadenilação alternativa

Após o agrupamento dos 81.664 transcritos em 32.136 clusters, pudemos verificar que apenas 14 mil possuíam dois ou mais transcritos e, assim, eram passíveis de sofrer eventos de poliadenilação alternativa. Verificamos as extremidades 3' dos transcritos de mesmo *cluster* e constatamos que 7070 *clusters* possuem variantes de poliadenilação, cuja diferença entre as terminações é de pelo menos 25 bases.

Desse total, pudemos encontrar nomes de genes segundo o MGC para 4.697 clusters, sendo que 258 possuíam dois nomes de genes e foram manualmente curados, escolhendo-se o nome correspondente ao gene com o maior número de transcritos. Os demais permaneceram com o identificador arbitrário.

Para verificar como ocorre a distribuição do tamanho relativo dos transcritos dos genes com eventos de poliadenilação alternativa, atribuímos ao maior transcrito de cada gene o valor de 100 e aos demais, valores proporcionais (em nucleotídeos), obtendo o gráfico da Figura 13Figura 13: Variação do tamanho relativo dos transcritos em cada gene, em relação ao maior transcrito.. Há pouca variação de tamanho entre os transcritos do mesmo gene, o que é esperado para variantes de poliadenilação, em que a variação ocorre apenas na porção 3' dos transcritos.

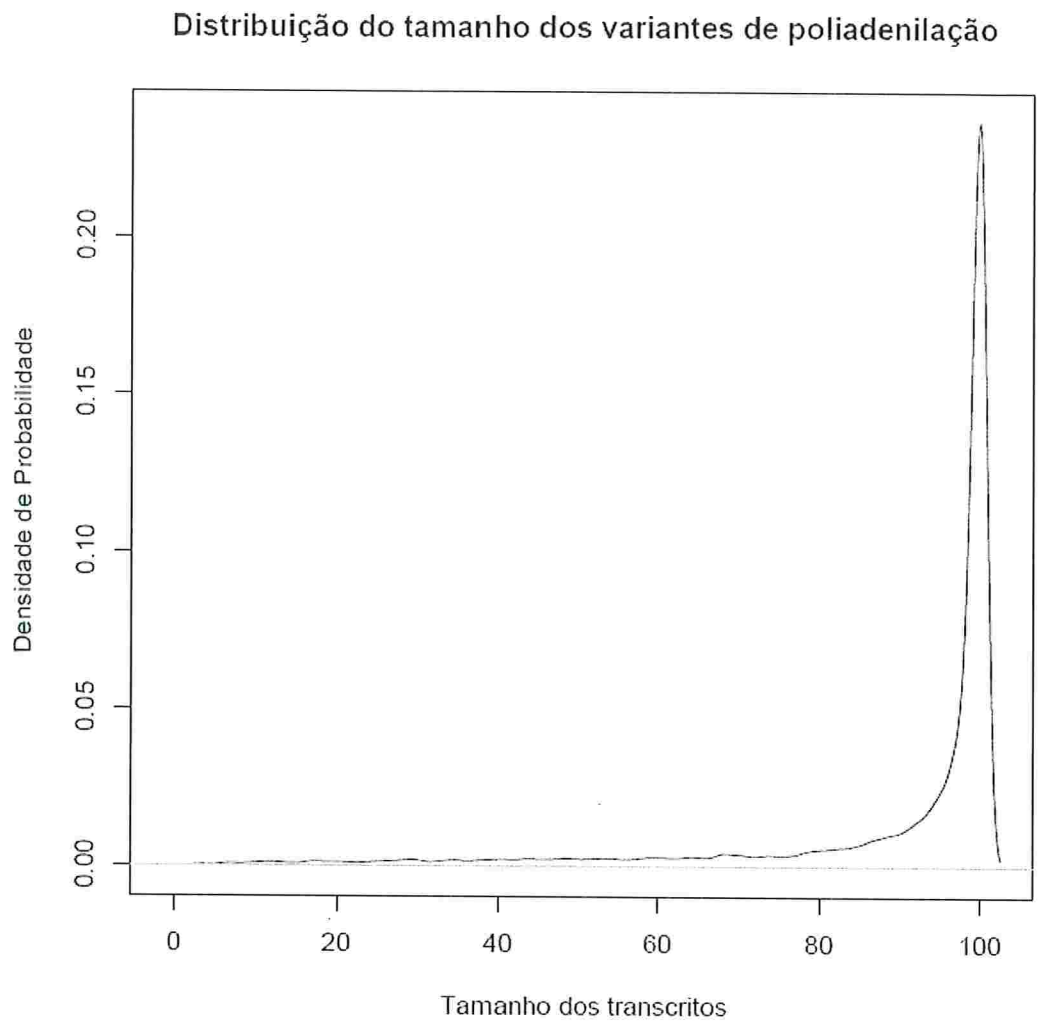


Figura 13: Variação do tamanho relativo dos transcritos em cada gene, em relação ao maior transcrito.

Verificamos o número de transcritos em cada gene com eventos de poliadenilação alternativa e obtivemos o histograma em números absolutos da Figura 14. É possível observar que a maioria dos genes tem menos que cinco transcritos diferentes.

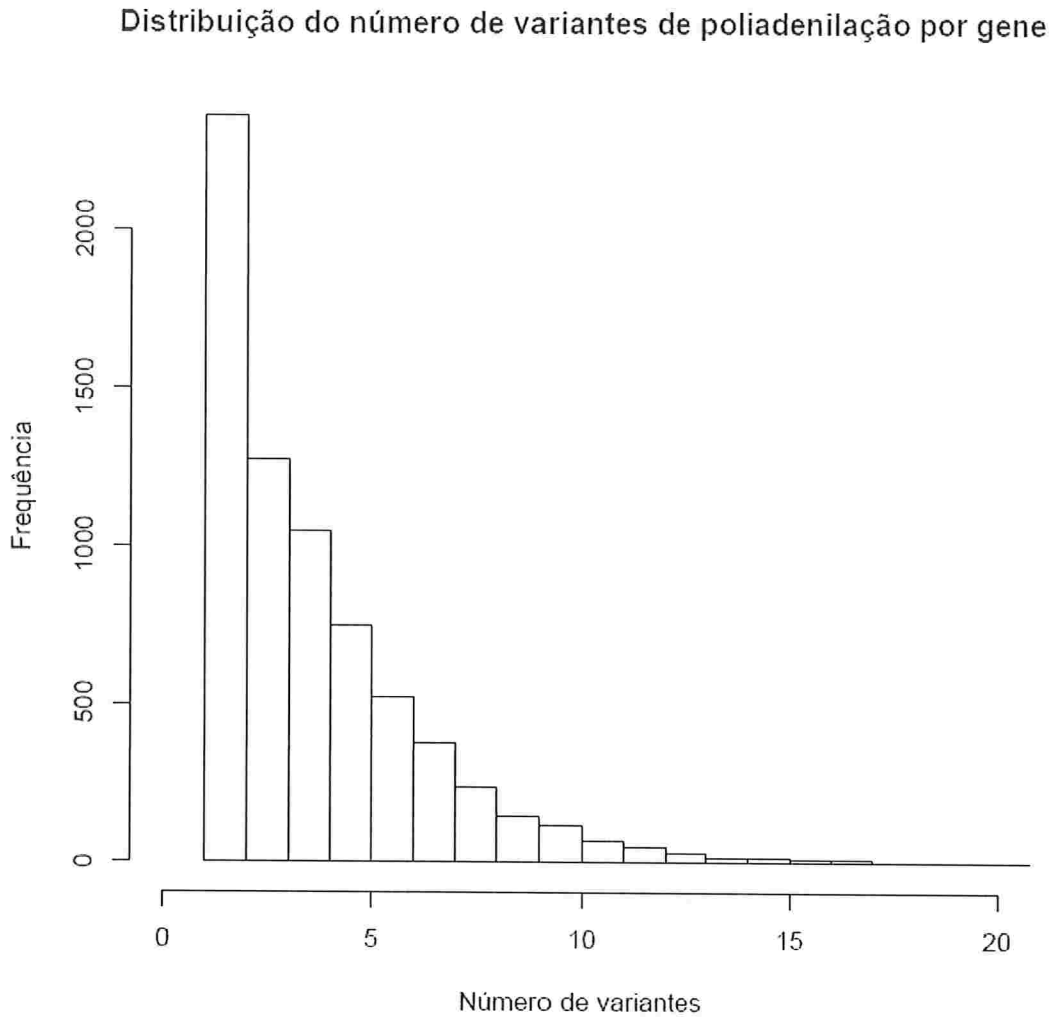


Figura 14: Histograma do número de variantes de poliadenilação por gene. Freqüência representa o número absoluto de genes com o respectivo número de variantes de poliadenilação.

4.6 - Localização do hexâmero sinal

O programa em Perl para o reconhecimento e localização do sinal de poliadenilação foi escrito de forma a procurar 10 hexâmeros descritos na literatura. A busca for restrita à região de -10 a -36 bases do sítio de clivagem, ou seja, 10 a 36 bases *upstream* à inserção da cauda poli(A).

Na primeira etapa de identificação dos transcritos poliadenilados, foi gerada a seqüência de cada transcrito com sua cauda poli(A) cortada. Essas seqüências foram usadas aqui para identificar o sítio de clivagem.

Dentre todos os transcritos com bom alinhamento, 70% apresentaram sinal de poliadenilação, sendo que destes, 94% exibiram um único sinal.

Nas seqüências dos transcritos, foram analisadas a posição preferencial dos hexâmeros sinais AATAAA na Figura 15 e ATTAAA na Figura 16abaixo. Foram utilizadas somente as seqüências que continham cada um dos hexâmeros canônicos: AATAAA, presente em 45.181 transcritos e ATTAAA, presente em 12.051 transcritos. Para o caso dos dois hexâmeros, podemos observar que em pouco mais de 10% das seqüências, o hexâmero encontra-se a 15 bases *upstream* em relação à extremidade 3´ dos transcritos.

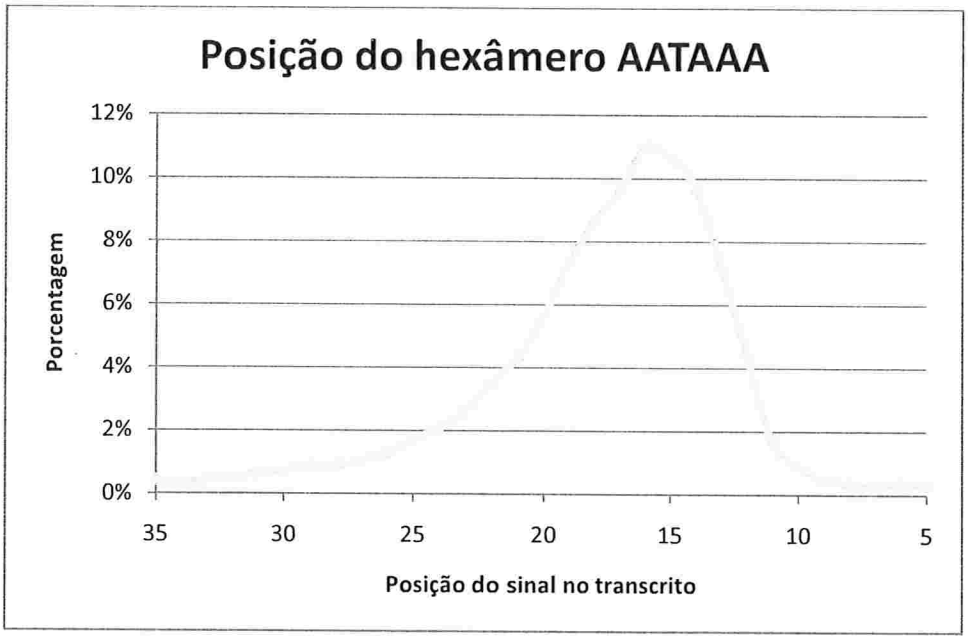


Figura 15: Posição do hexâmero AATAAA em relação à extremidade 3' dos transcritos poliadenilados. As porcentagens se referem ao total de 45.181 transcritos que possuem o sinal AATAAA.

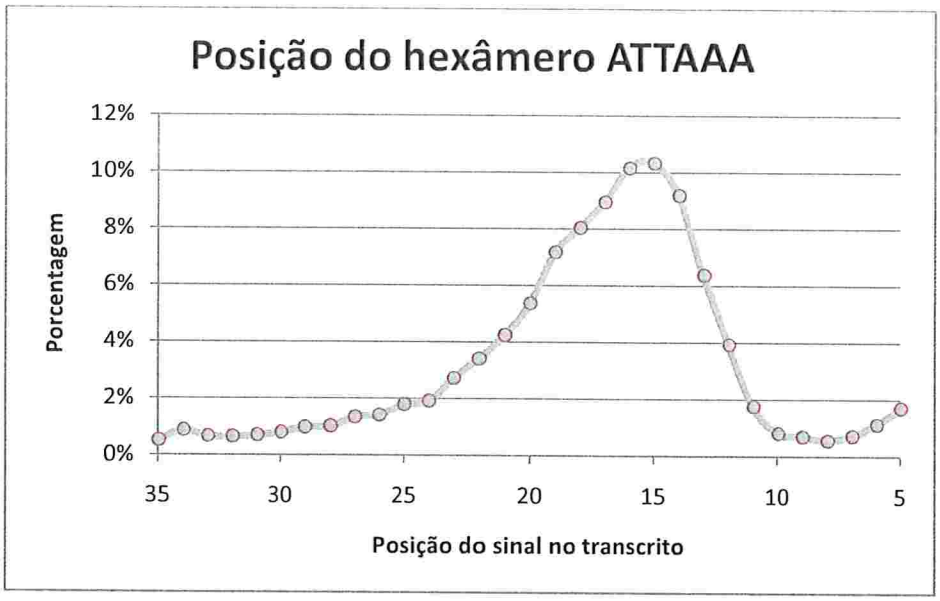


Figura 16: Posição do hexâmero AATAAA em relação à extremidade 3' dos transcritos poliadenilados. O eixo y se refere à porcentagem de seqüências que possuem o sinal ATTAAA em determinada posição, do total de 12.051 transcritos.

4.7 - Busca dos hexâmeros sinais nos transcritos alinhados

Após o alinhamento dos mRNAs poliadenilados contra o genoma humano, observamos novamente a porcentagem de uso dos sinais de poliadenilação.

Podemos observar na Figura 17 a porcentagem dos hexâmeros que ocorrem em toda a região do sinal, 10 a 36 bases *upstream* ao sítio de clivagem. Diversos transcritos possuem mais de um sinal. Na Figura 18 podemos observar a porcentagem do sinal mais *downstream*, ou seja, o sinal mais próximo ao sítio de clivagem. Neste gráfico há somente um sinal por transcrito.

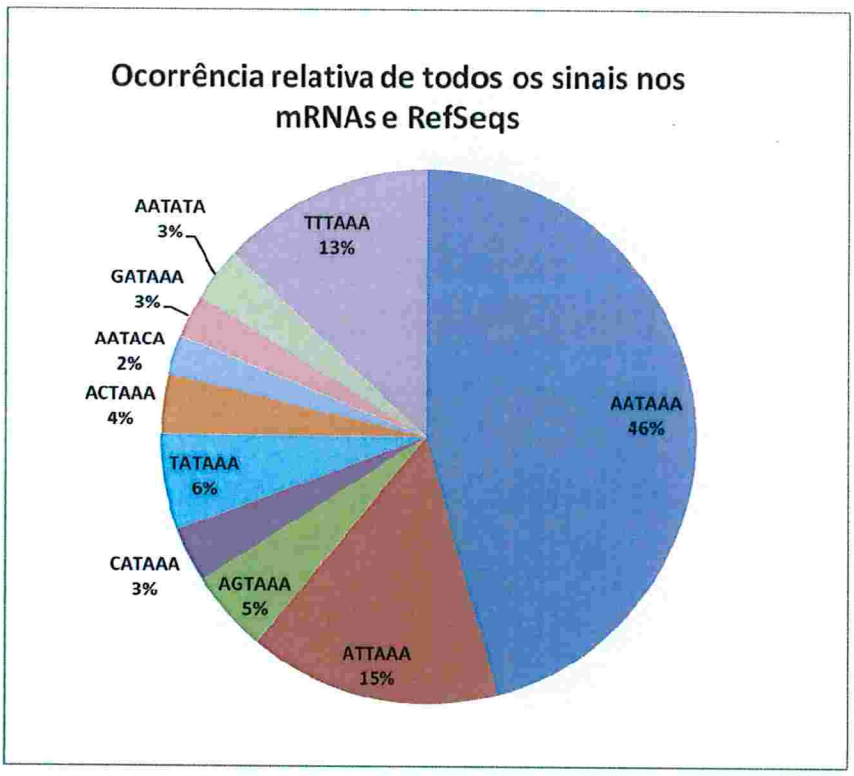


Figura 17: Ocorrência relativa de sinais de poliadenilação encontrada na região de 10 a 36 bases *upstream* ao sítio de clivagem, de todos os mRNAs e RefSeqs estudados.

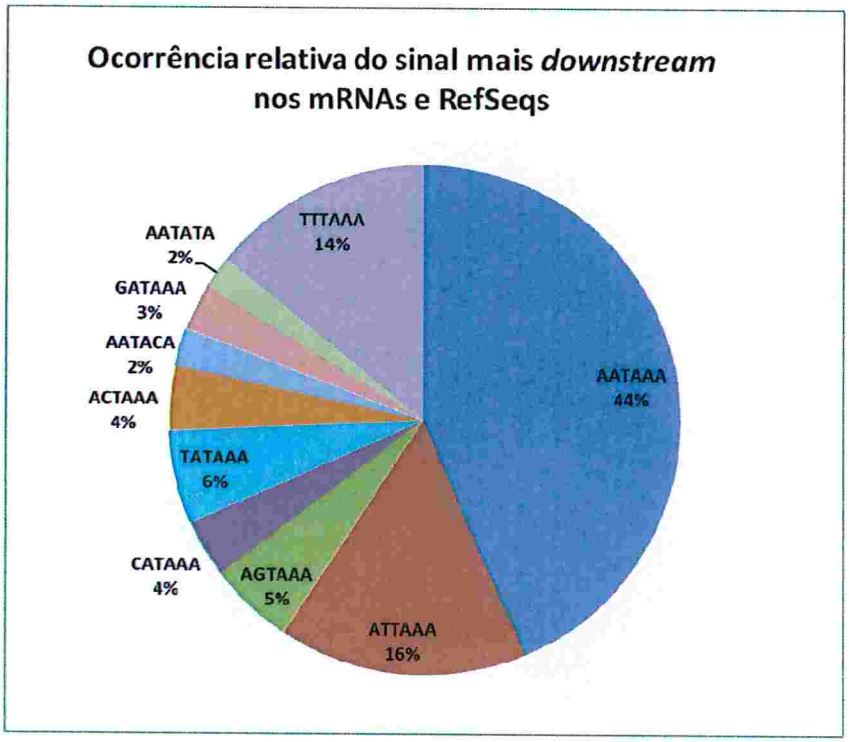
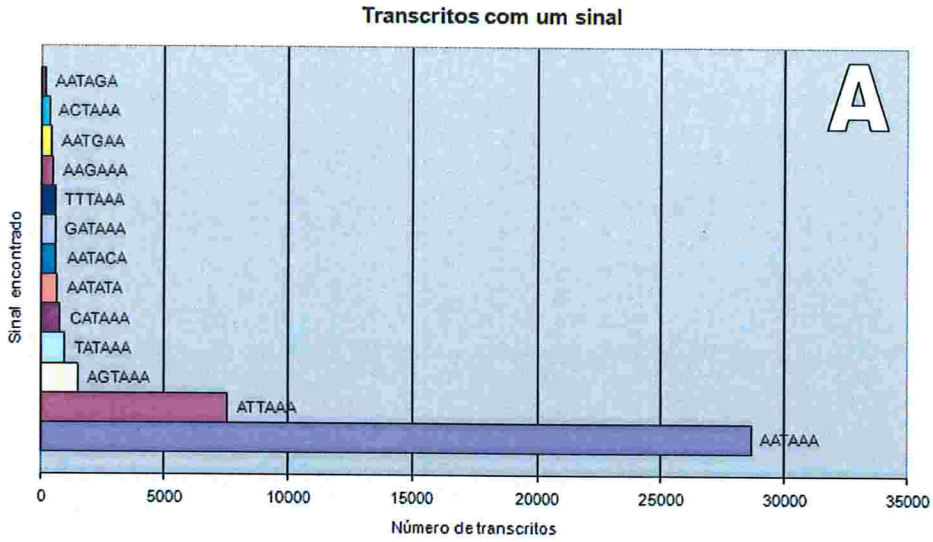


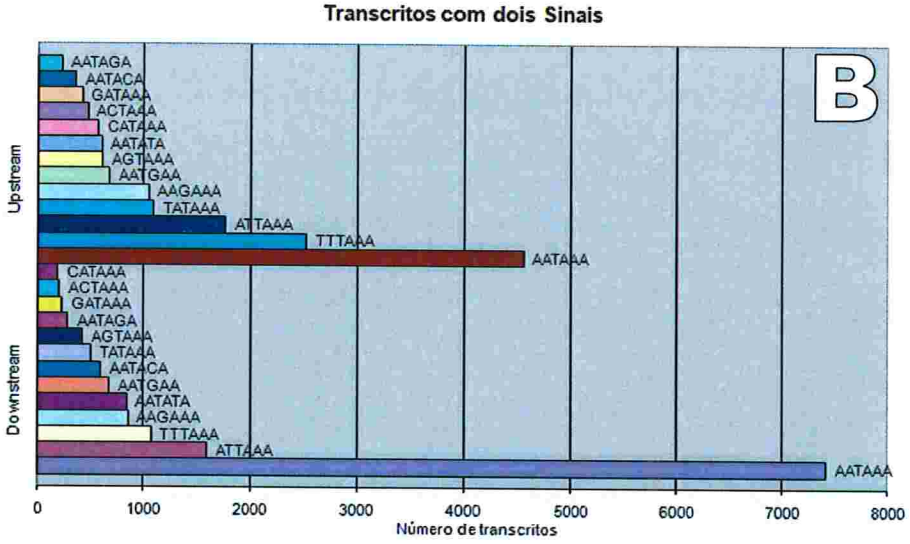
Figura 18: Ocorrência relativa do sinal *downstream*, ou seja, o sinal mais próximo à extremidade 3' (também chamado distal).

Na região de localização do hexâmero sinal, há a possibilidade de encontrarmos até oito sinais, se estiverem concatenados e sobrepostos uns aos outros. Esse tipo de arquitetura pode ser encontrada em diversos genes, como o da esfingomielina sintase 1 (SGMS1) e miociclina (MYOC), que contêm três sinais de poliadenilação em sua região 3'UTR. Um exemplo extremo pode ser observado na Figura 19, encontrado em um RefSeq do gene QPCTL, no cromossomo 19. O RefSeq NM_017659 possui 6 hexâmeros na região onde pode ser encontrado o sinal de poliadenilação.

Figura 19: Exemplo de transcrito com seis sinais, cuja região é mostrada em vermelho. Abaixo, podemos observar os seis hexâmeros AATAAA encontrados na região no sinal.

Mostramos na Figura 20 a freqüência de sinais simples(A), duplos(B) e triplos(C) encontrados no conjunto de RefSeqs e mRNAs. Nestes casos, consideramos apenas os transcritos que continham um único sinal (A), dois (B) e três (C) sinais na região de -10 a -36 bases da extremidade 3´ de sua seqüência.


Transcritos com um sinal
A
AATAGA
ACTAAA
AATGAA
AAGAAA
TTTAAA
GATAAA
AATACA
AATATA
CATAAA
TATAAA
AGTAAA
ATTAAA
AATAAA
Sinal encontrado
0
5000
10000
15000
20000
25000
30000
35000
Número de transcritos


Transcritos com dois Sinais
B
Upstream
AATAGA
AATACA
GATAAA
ACTAAA
CATAAA
AATATA
AGTAAA
AATGAA
AAGAAA
TATAAA
ATTAAA
TTTAAA
AATAAA
Downstream
CATAAA
ACTAAA
GATAAA
AATAGA
AGTAAA
TATAAA
AATACA
AATGAA
AATATA
AAGAAA
TTTAAA
ATTAAA
AATAAA
0
1000
2000
3000
4000
5000
6000
7000
8000
Número de transcritos

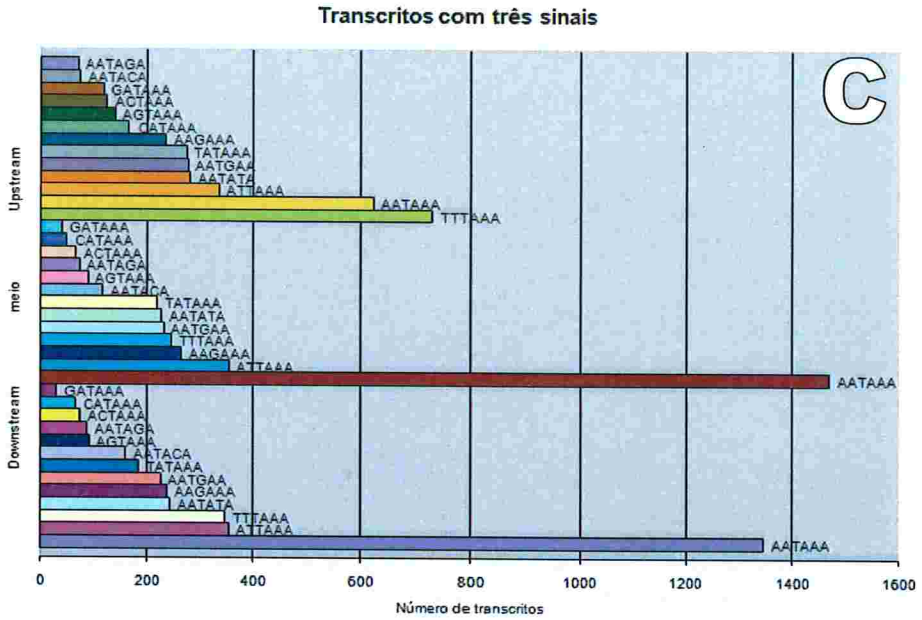


Figura 20: Freqüência dos sinais de poliadenilação nos transcritos com (A) um, (B) dois e (C) três sinais. Podemos observar proporções diferentes conforme o número de sinais encontrados em cada transcrito.

Verificamos a proporção dos sinais em transcritos mais curtos e mais longos dos genes com variantes de poliadenilação, conforme a Figura 21. No grupo dos variantes mais curtos, 20% dos transcritos não possuíam nenhum dos 10 sinais procurados, enquanto que no grupo dos variantes mais longos, 11% dos transcritos não possuíam sinal.

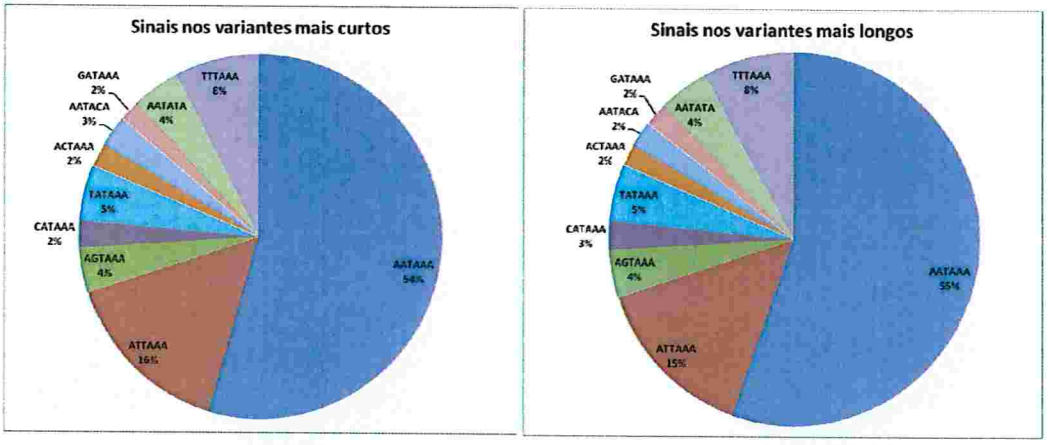


Figura 21: Ocorrência relativa dos sinais nos variantes de poliadenilação.

4.8 - Sinais da região *downstream* aos transcritos

Após a retirada da heterogeneidade dos sítios de clivagem, conforme descrito em "Métodos", obtivemos 39.734 sítios de poliadenilação distintos, nos quais procedemos à identificação de *internal priming*.

Nossa avaliação da região *downstream* (vide "Métodos") é capaz de identificar seqüências caracterizadas como *internal priming*, o que ocorre em 7.789 sítios. Para os restantes 31.945 sítios, foi calculado um *score*, entre zero e 360, correspondente à riqueza de dinucleotídeos TT e GT, importantes para a ligação do CstF.

Na Figura 22, temos as médias dos *scores* relativas a cada sinal de poliadenilação, agrupadas por tamanho de variante.

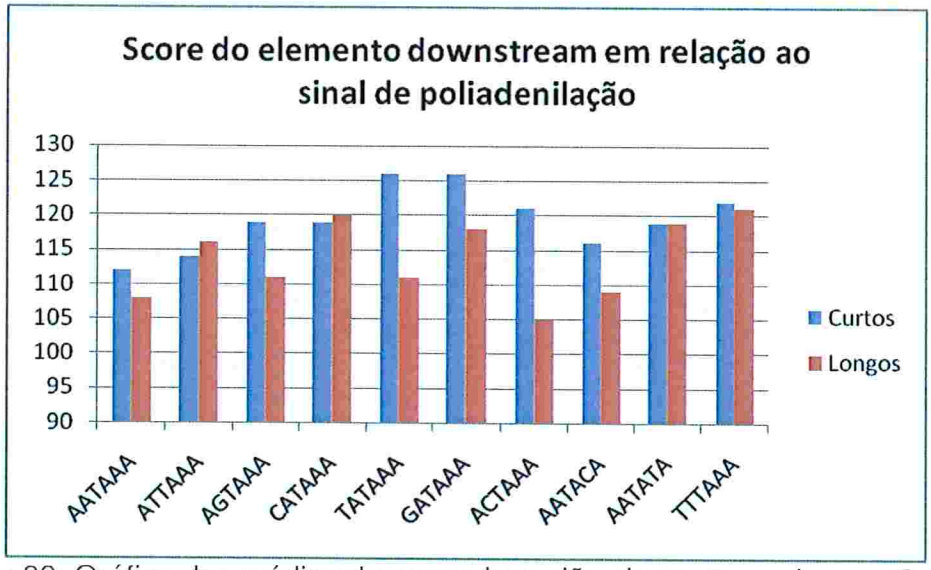


Figura 22: Gráfico das médias do *score* da região *downstream* rica em G e U em relação ao hexâmero sinal, para variantes de poliadenilação curtos e longos.

No gráfico acima, ocorre que quanto mais alto o *score* do elemento *downstream*, maior a chance de ocorrer clivagem e poliadenilação naquele sítio, pois mais forte será a ligação do elemento CstF na seqüência *downstream*. Em contrapartida, quanto mais baixo o score, menor a chance de ocorrer clivagem e poliadenilação, pois o CstF terá uma ligação mais fraca, ou não se ligará à seqüência *downstream*.

4.9 - SNPs no Sinal de Poliadenilação

Os polimorfismos podem ocorrer em qualquer posição do genoma, então verificamos sua coexistência com os sinais de poliadenilação.

Foram procuradas as posições genômicas de todos os 128.205 sinais de poliadenilação, mesmo quando havia mais de um sinal por transcrito. Devido aos *gaps* no alinhamento, 1269 sinais não foram encontrados nas posições genômicas previstas. Utilizamos o genoma modificado de acordo com o dbSNP131, que representa as bases de acordo com a nomenclatura IUPAC, conforme descrito em "Métodos", na Tabela 2: Nomenclatura IUPAC para nucleotídeos.

Foi possível encontrar SNPs na posição genômica de 1860 sinais, ou seja, no total, 1860 transcritos possuíam SNP no sinal de poliadenilação. Considerando que um ou mais desses transcritos pertencem ao mesmo *cluster*, pudemos contabilizar 930 *clusters* com SNP no sinal, no total de genes com transcritos poliadenilados.

Dentre o total de 7070 clusters com eventos de poliadenilação alternativa, pudemos identificar 378 SNPs distintos. Desses, 230 clusters tinham o SNP no sinal do transcrito mais curto do gene com poliadenilação alternativa.

Na Figura 23 podemos observar a ocorrência relativa dos sinais de poliadenilação, para os casos onde foi encontrado um SNP na mesma posição genômica.

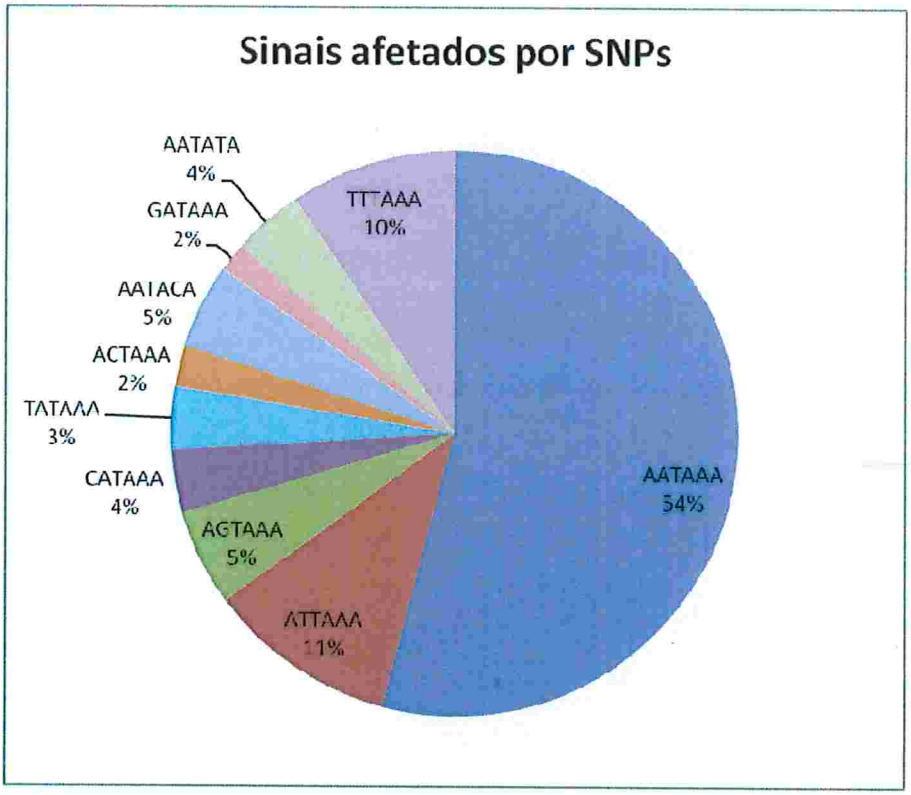


Figura 23: Ocorrência relativa de todos os sinais de poliadenilação colocalizados com SNPs no genoma humano.

Para avaliar como os polimorfismos afetariam as seqüências, foram testadas as modificações de bases dadas pelos SNPs. A Figura 24 ilustra o que ocorreria nesses casos. Em azul, observamos as proporções originais da Figura 23, em vermelho, os sinais que manteriam sua funcionalidade e em verde, os hexâmeros que não se assemelhariam a sinais de poliadenilação.

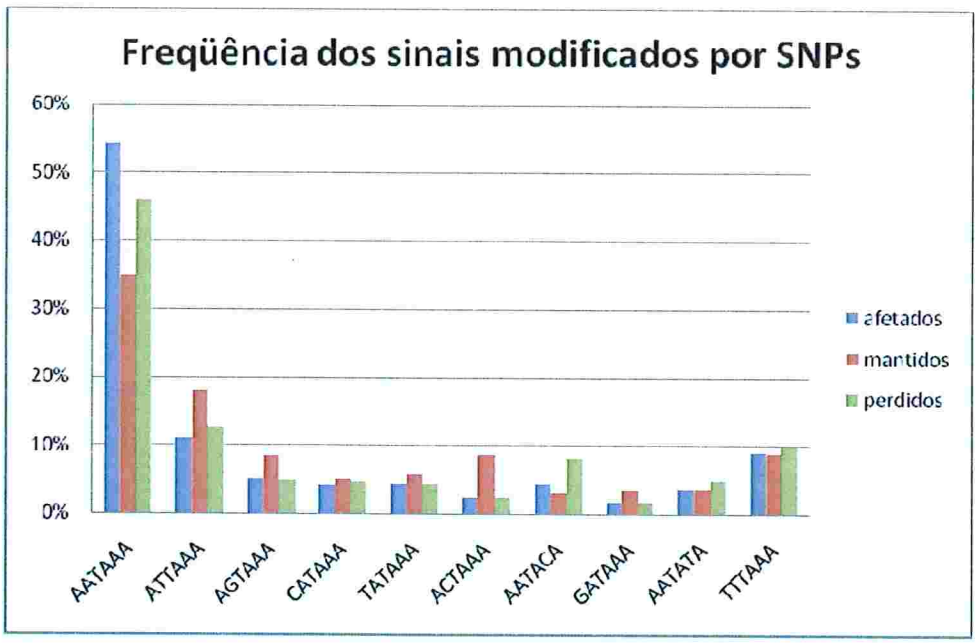


Figura 24: Freqüências relativas aos sinais localizados nas mesmas posições genômicas de SNPs, em azul. Em vermelho estão representados os sinais que, após a modificação dada pelo SNP, ainda mantêm um hexâmero semelhante a um sinal. Em verde, os sinais que são completamente perdidos após a modificação dada pelo SNP.

Como exemplo do efeito da presença de polimorfismo em uma seqüência transcrita, podemos observar na Figura 25 um alinhamento múltiplo do gene da desidrogenase DHRS7 realizado pelo software ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/), utilizando seqüências de RefSeqs, mRNAs e *reads* de HCC1954 (vide item 4.13).

Figura 25: Exemplo de alinhamento múltiplo de diversos transcritos do mesmo gene, realizado pelo ClustalW. Podemos observar a presença do SNP no sinal de dois transcritos. No entanto, o sítio de poliadenilação não é perdido, devido à existência de sinais concatenados.

Procuramos avaliar se os SNPs encontrados estão distribuídos diferencialmente em genes com eventos de poliadenilação alternativa. Na Figura 26 podemos observar a ocorrência relativa dos SNPs colocalizados na posição genômica de sinais de variantes de poliadenilação.

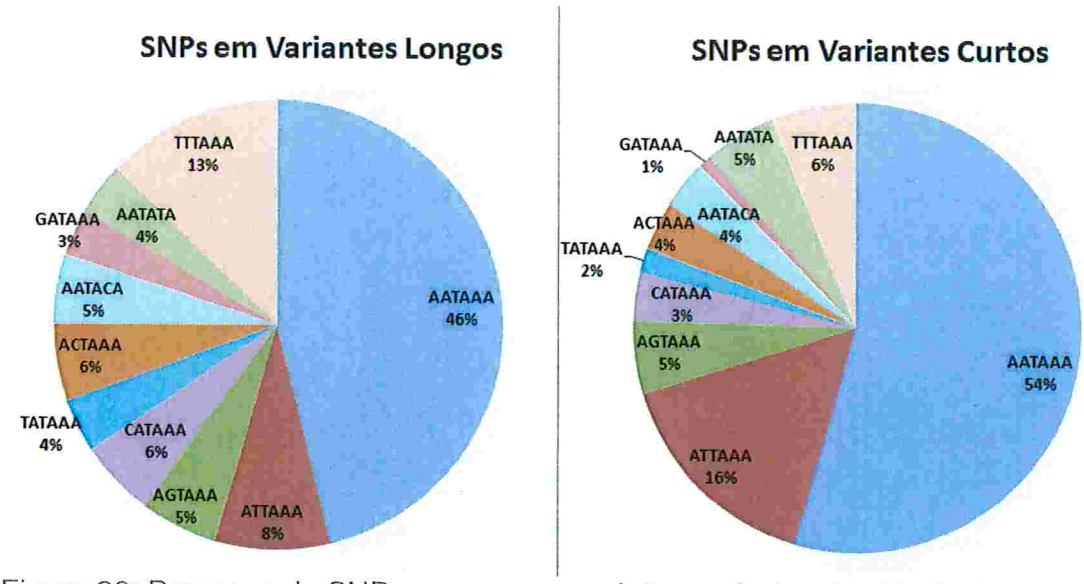


Figura 26: Presença de SNPs na mesma posição genômica dos sinais de variantes de poliadenilação mais curtos e mais longos.

Para a montagem dos gráficos acima, foram selecionados apenas os SNPs em sinais das extremidades mais curta e mais longa dos genes com variantes de poliadenilação.

É possível observar nos gráficos acima que, embora mantenham a proporção da própria presença de sinais no transcriptoma, a comparação entre eles mostra que os SNPs em variantes curtos afetam em maior proporção os sinais canônicos (AATAAA e ATTAAA) que em variantes longos.

4.10 - Formação de pares *sense-antisense* devido à poliadenilação alternativa

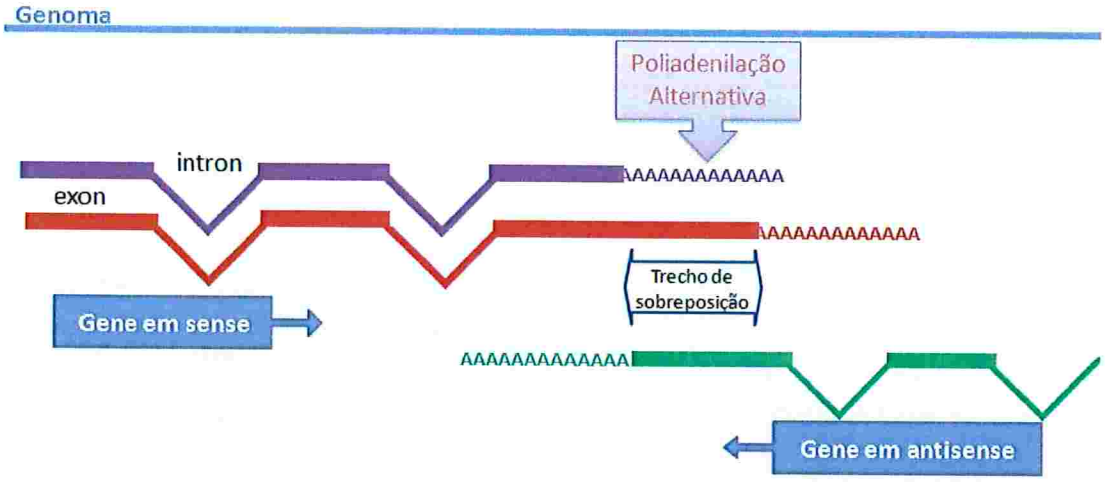


Figura 27: Esquema da disposição genômica dos casos de sobreposição entre genes com poliadenilação alternativa e genes em antissenso.

O gene em senso é aquele no qual identificamos o evento de poliadenilação alternativa, no qual o maior transcrito sobrepõe o gene em antissenso. Para os genes com poliadenilação alternativa consideramos as posições genômicas dos transcritos de tamanhos máximo e mínimo. Identificamos transcritos em antissenso cuja extremidade 3' está no intervalo entre estas posições genômicas. Analisamos a sobreposição entre genes com poliadenilação alternativa em senso e genes em antissenso. A sobreposição é restrita ao trecho estendido do maior transcrito do gene com poliadenilação alternativa, para que possamos encontrar variantes de poliadenilação cuja expressão sofre regulação por antissenso.

A comparação das posições genômicas dos genes com poliadenilação alternativa com o total de genes com transcritos poliadenilados revelou um total de 435 pares de genes de fitas opostas com sobreposição entre o trecho estendido do gene com poliadenilação alternativa e o gene em antissenso.

Com os genes poliadenilados:

São 215 sobreposições da fita "mais" alternativa sobre a fita "menos".

São 220 sobreposições da fita "menos" alternativa sobre a fita "mais".

Por outro lado, a comparação das posições genômicas dos genes com variantes de poliadenilação com os genes obtidos da tabela "Known Genes" do site do UCSC (site: http://genome.ucsc.edu/cgi-bin/hgTables) revelou um total de 636 genes nos quais um variante mais longo de poliadenilação se sobrepõe à terminação 3' dos genes conhecidos.

Com os genes do UCSC *known* genes:

São 124 sobreposições da fita "mais" alternativa sobre a fita "menos".

São 512 sobreposições da fita "menos" alternativa sobre a fita "mais".

Dentre os transcritos envolvidos em sobreposições *sense-antisense*, provindos dos genes poliadenilados, foi possível encontrar 500 sinais de poliadenilação, cuja ocorrência relativa pode ser observada na Figura 28.

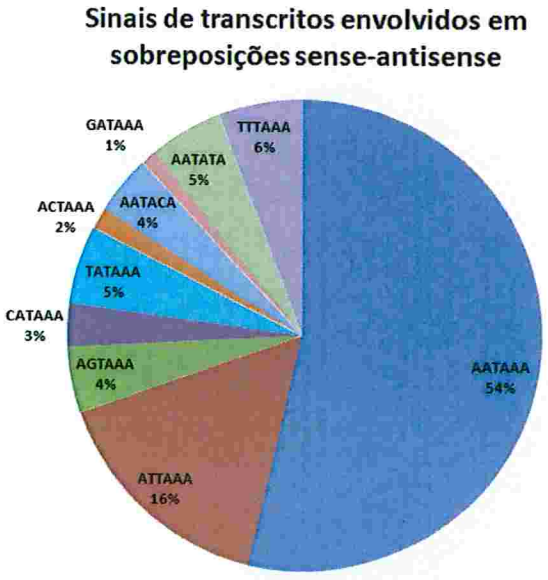


Figura 28: Ocorrência relativa dos sinais dos genes poliadenilados envolvidos em sobreposições senso-antissenso.

No entanto, somente podemos observar diferenças significativas entre o uso de cada sinal se separarmos os genes em senso e em antissenso.

Utilizando somente o sinal mais próximo à extremidade 3' temos as ocorrências relativas observadas na Figura 29, na qual o esquema em vermelho reflete os genes com eventos de poliadenilação alternativa e o esquema em verde o total de genes poliadenilados, em antissenso.

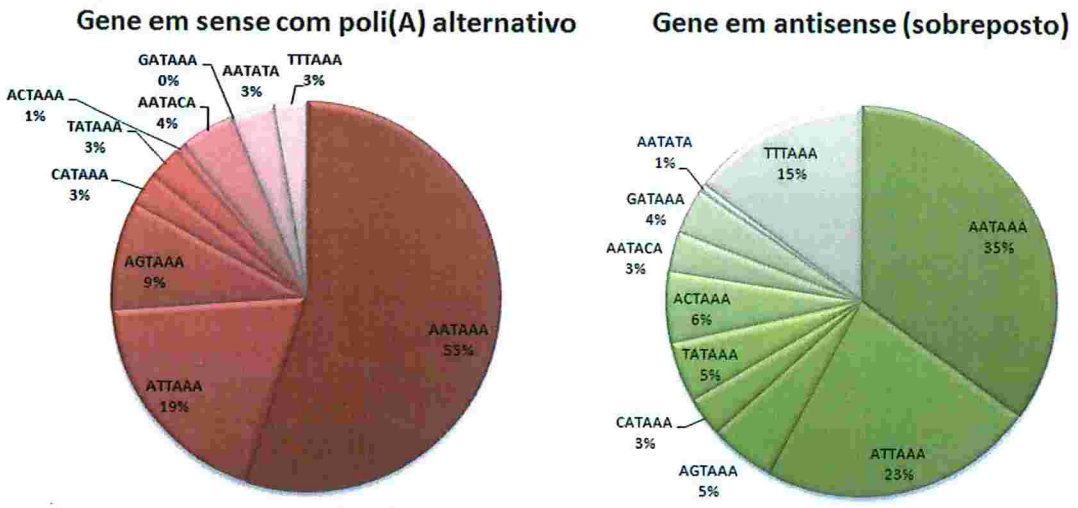


Figura 29: À esquerda temos as porcentagens de ocorrência dos sinais de poliadenilação dos genes mais longos em senso. À direita, podemos ver os sinais dos transcritos que são sobrepostos pelos genes com poli(A) alternativo e estão em antissenso; representados em verde na figura esquemática acima.

Na Figura 29, é possível observar que as freqüências de sinais do gene em senso obedecem aproximadamente às freqüências observadas para o total de genes com poliadenilação alternativa. No entanto, para os genes em antissenso (os sobrepostos), semelhantes aos genes sem eventos de poliadenilação alternativa, é possível observar um aumento dos sinais não canônicos, cuja soma das freqüências sobe de 25% para 42%. Como exemplo, o sinal ACTAAA tem freqüência de 1% entre os genes em senso, com poliadenilação alternativa, e 6% entre os genes em antissenso.

4.11 - Expressed Sequence Tags (ESTs)

Iniciamos as análises de todas as 8.301.458 ESTs humanas obtidas do UCSC, versão de fevereiro de 2009 (GRCh37/hg19) com a procura pelas seqüências poliadeniladas.

Utilizamos o procedimento anteriormente descrito para procurar pelos RefSeqs e mRNAs poliadenilados, mas o critério foi modificado para que fosse um pouco menos estringente, pois as ESTs são seqüências mais curtas; seu comprimento médio é de 513 bases. São seqüenciadas aleatoriamente e de única passagem a partir de bibliotecas de cDNA e, por isso, possuem mais erros de seqüenciamento (Nagaraj, Gasser, & Ranganathan, 2006).

As ESTs foram consideradas poliadeniladas se possuíssem pelo menos quatro adeninas nas cinco últimas bases da extremidade 3'. Dessa maneira foram selecionadas 299.363 ESTs poliadeniladas.

Para encontrar a posição de início da cauda poli(A), utilizamos uma janela deslocada base a base, que admite seis adeninas a cada de dez bases. Ao encontrar mais uma base diferente de adenina, encontra-se o suposto sítio de clivagem.

No entanto, segundo o mesmo site do UCSC, apenas 278.469 dessas ESTs poliadeniladas estavam mapeadas no genoma humano e serão utilizadas nas próximas etapas.

A partir das posições genômicas dos 14.440 clusters (com ao menos dois transcritos poliadenilados cada), foram localizadas ESTs mapeadas nas mesmas posições dos genes conhecidos, ou seja, 179.927 ESTs representam 12.676 genes.

Metade dessas ESTs representam variantes de poliadenilação já existentes em RefSeqs e mRNAs e a outra metade (89.705 ESTs) representam novos sítios de poliadenilação, com 40 bases de diferença entre um e outro.

A mesma análise, feita com 25 bases de diferença os variantes de poliadenilação revelou que a partir do total de 179.927 ESTs, 92.360 ESTs representam novos sítios e 87.567 ESTs representam os mesmos sítios de poliadenilação observados em mRNAs e RefSeqs.

4.12 - Seqüenciamento em larga escala (HCC1954)

HCC1954 é uma célula tetraplóide, derivada do ducto da glândula mamária. Essa célula é derivada de um carcinoma ductal invasivo, de grau 3, de estágio primário IIA, sem metástases no linfonodo.

Foram utilizados neste trabalho um total de 510.693 *reads* de cDNA do transcriptoma das células de HCC1954, seqüenciados com a tecnologia de piroseqüenciamento 454 da Life Sciences® (Zhao, et al., 2009). Foi possível encontrar a localização genômica de 499.998 *reads*.

Foi utilizado também o transcriptoma da linhagem linfoblastóide obtida da mesma paciente e denominada doravante de HCC1954BL. Para essa linhagem foi obtido um total de 510.376 *reads* provenientes das células linfoblastóides de HCC1954BL. A localização genômica de 486.875 *reads* foi determinada através do alinhamento com o programa BLAT; em caso de alinhamento em mais de uma posição no genoma humano, foi selecionada a posição com o maior número de identidade. Para a identificação da presença da cauda poli(A) foram utilizados os mesmos critérios da análise de ESTs e assim 34.983 *reads* possuíam cauda poli(A).

- **HCC1954 e HCC1954BL**

Para a análise dos transcritos de HCC1954, utilizamos como referência do transcriptoma humano os 81.672 transcritos poliadenilados (RefSeqs e mRNAs). Esses transcritos formam 14.440 genes, com pelo menos dois transcritos em cada, sendo que 7070 genes possuíam eventos de poliadenilação alternativa.

As posições genômicas dos *reads* de HCC1954 foram comparadas às dos RefSeqs e mRNAs poliadenilados do transcriptoma humano. Somente 64.149 *reads* tinham posições genômicas equivalentes a 2802 genes com variantes de poliadenilação. Estes não eram *reads* poliadenilados e não podem representar o final 3' dos transcritos conhecidos.

- **Identificação de *reads* poliadenilados**

Utilizamos os 510.693 *reads* de HCC1954 e utilizamos o procedimento anteriormente descrito para procurar pelos *reads* poliadenilados, com critérios um pouco menos estringentes, a seguir.

Verificamos quatro bases no final 3´ do *read*, admitindo, dentre essas quatro, uma base diferente de adenina. Em seguida são verificadas as bases seguintes no sentido *upstream*, e em cada janela de 10 bases, são aceitas como cauda poli(A) 6 adeninas entre 10 bases.

Com esse critério obtivemos 21.190 *reads* poliadenilados. Desses, 20.559 estavam mapeados no genoma.

- ***Reads* envolvidos em poliadenilação alternativa**

HCC1954

Os *reads* poliadenilados também foram analisados quanto às posições genômicas equivalentes às dos genes com eventos de poliadenilação alternativa.

Segundo os critérios detalhados acima, identificamos 21.031 *reads* como poliadenilados, sendo que, desses, apenas 9.088 não foram caracterizados como *internal priming*, ou seja, sua cauda poli(A) não alinha no genoma.

A partir desses 9.088 *reads*, pudemos verificar que eles estão mapeados em 2.712 genes com eventos de poliadenilação alternativa.

Separamos também em dois grupos: 5.023 *reads* que representam novos variantes de poliadenilação e 4.065 *reads* que têm a mesma terminação que os variantes já existentes entre mRNAs e RefSeqs.

Dentre esse último grupo, 850 têm a terminação idêntica ao maior transcrito dos genes com poliadenilação alternativa e 2.577 representam os variantes mais curtos desses genes. As proporções obtidas podem ser observadas na Figura 30.

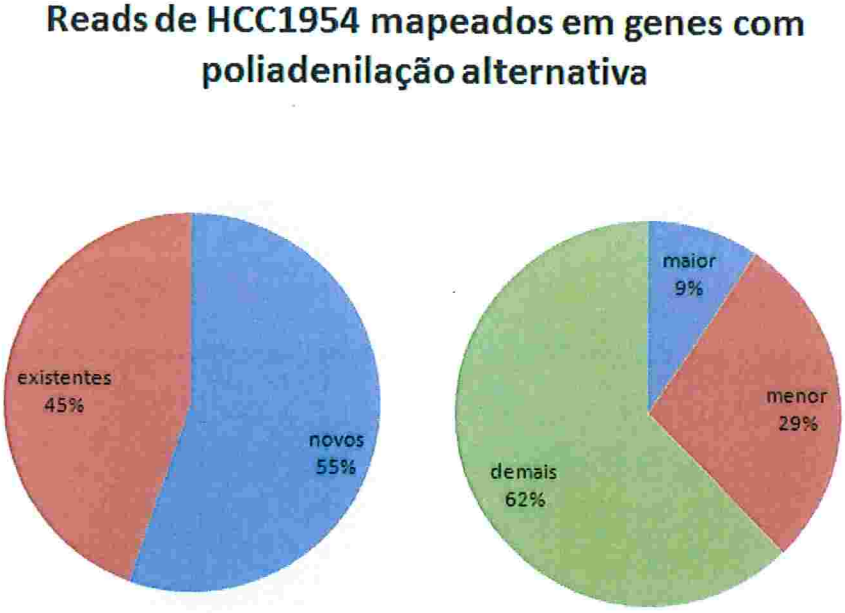


Figura 30: À esquerda, observa-se a maioria de variantes de poliadenilação que não são observados no conjunto de mRNAs e RefSeqs. À direita, distribuição dos reads de HCC1954BL entre variantes de poliadenilação mais curtos, mais longos e intermediários.

- Sinais de poliadenilação em HCC1954

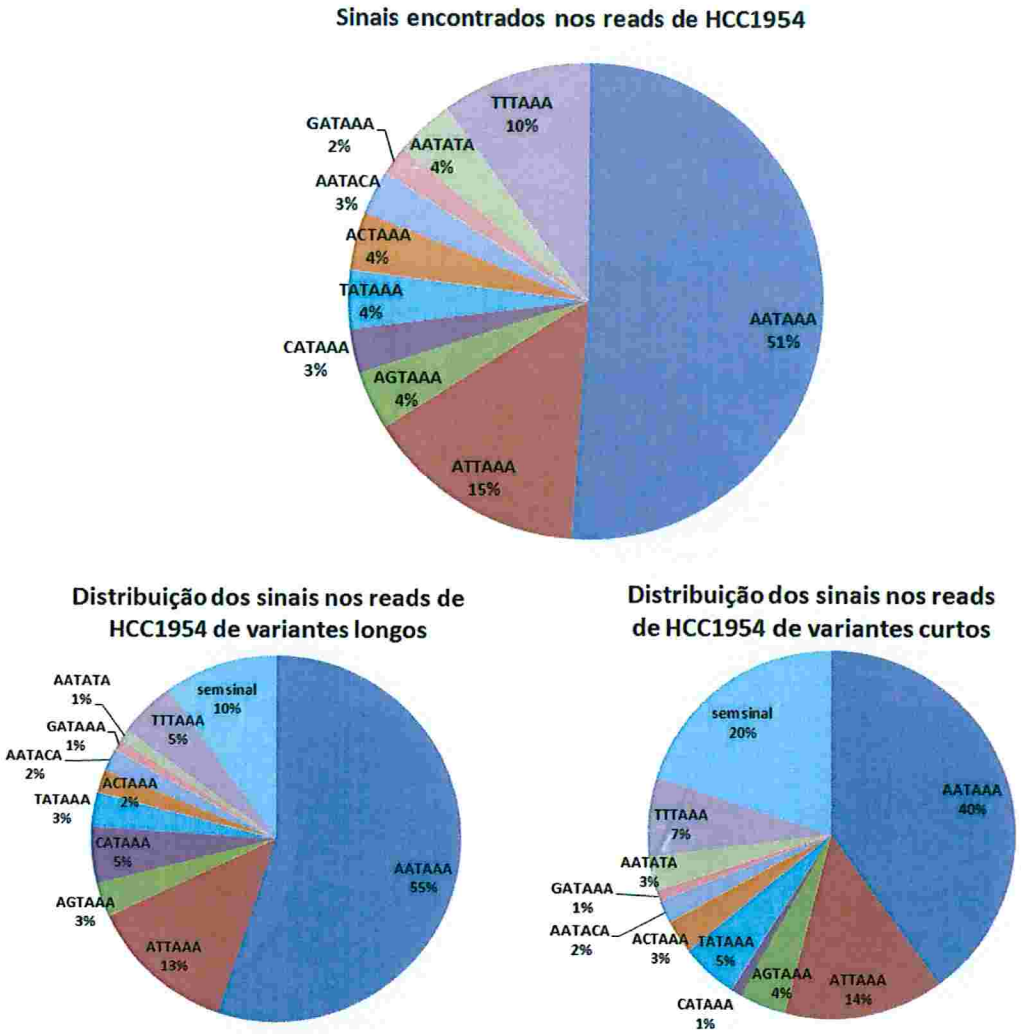


Figura 31: Proporção de sinais de poliadenilação encontrados nos reads de HCC1954. Em 43% dos reads não foi encontrado hexâmero sinal. Abaixo, pode-se observar a proporção dos sinais em reads que representam variantes de poliadenilação mais curtos e mais longos.

HCC1954BL

A mesma análise foi realizada para os 510.376 *reads* provenientes das células linfoblastóides de HCC1954BL. A localização genômica de 486.875 *reads* foi determinada e 34.983 *reads* possuíam cauda poli(A) segundo os critérios adotados. Os *reads* que representaram *internal priming* foram retirados das análises posteriores, restando assim 13.631 *reads* que estão mapeados em 4.163 genes com eventos de poliadenilação alternativa.

Desses 13.631 *reads*, 9.365 formam novas terminações 3' e 4.266 *reads* representam variantes de poliadenilação já existentes no grupo de mRNAs e RefSeqs. Além disso, verificamos que 1512 *reads* têm terminação idêntica ao maior variante e 4.745 *reads* representam os variantes menores dos genes com poliadenilação alternativa. Os demais *reads* representam variantes intermediários desses genes.

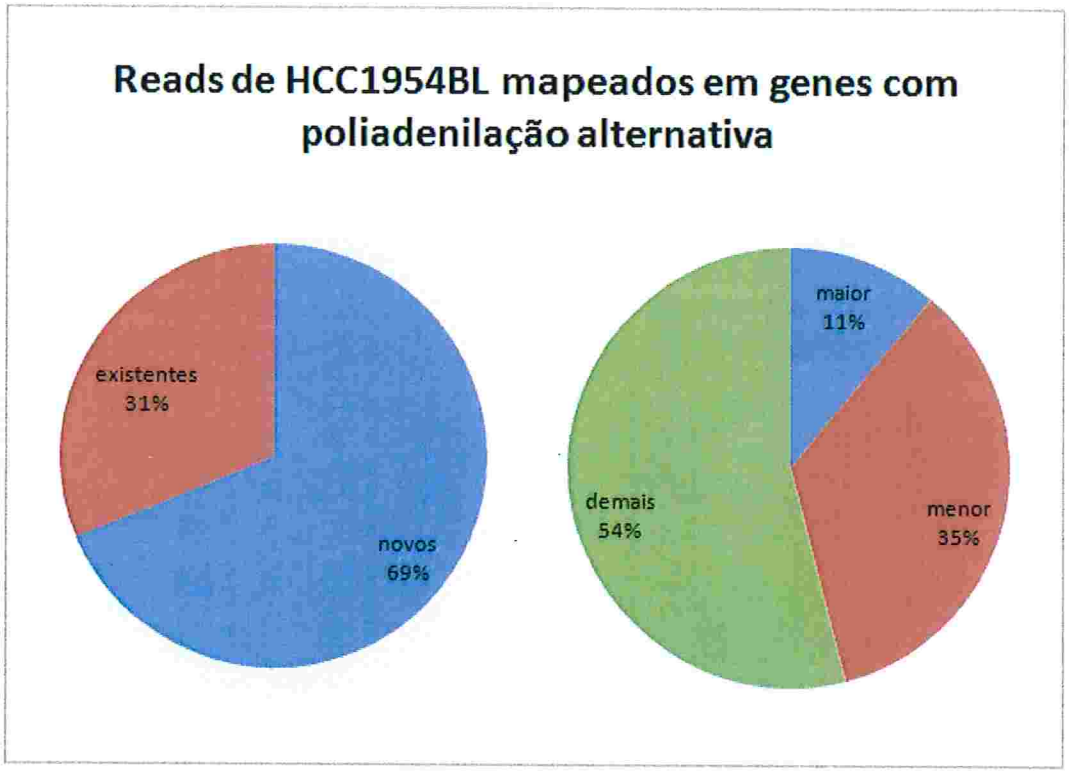


Figura 32: À esquerda, observa-se a maioria de variantes de poliadenilação que não são observados no conjunto de mRNAs e RefSeqs. À direita, distribuição dos reads de HCC1954BL entre variantes de poliadenilação mais curtos, mais longos e intermediários.

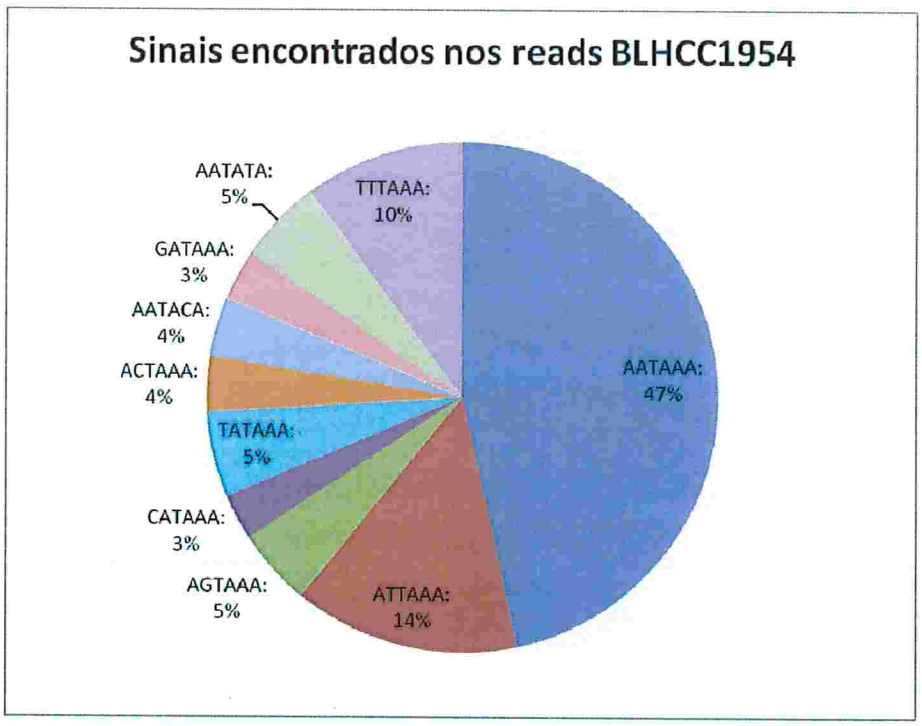


Figura 33: Proporção de sinais de poliadenilação encontrados nos reads de HCC1954BL. Interessantemente, em 49% dos reads não foi encontrado sinal. As porcentagens acima correspondem a aproximadamente 51% dos reads com bom alinhamento no genoma.

- **Antisense em HCC1954 e em HCC1954BL**

Para avaliar se os *reads* provindos da cultura dos dois tipos celulares estavam envolvidos na região de sobreposição dos genes envolvidos em *sense-antisense*, separamos a região de sobreposição de acordo com os resultados anteriores.

Assim utilizamos os 329 trechos genômicos de sobreposição selecionados anteriormente. Analisamos então a posição genômica dos *reads* poliadenilados e verificamos que 203 *reads* de HCC1954 e 437 *reads* de HCC1954BL estão envolvidos em sobreposições *sense-antisense*. Em seguida foram retirados os *reads* que poderiam ser provenientes de *internal priming*, restando assim 142 *reads* de HCC1954 e 310 *reads* de HCC1954BL envolvidos em sobreposições sense-antisense.

Dentre os 142 *reads* de HCC1954 selecionados acima, 44% possuíam um hexâmero sinal de poliadenilação. Sua proporção pode ser observada na Figura 34 a seguir:

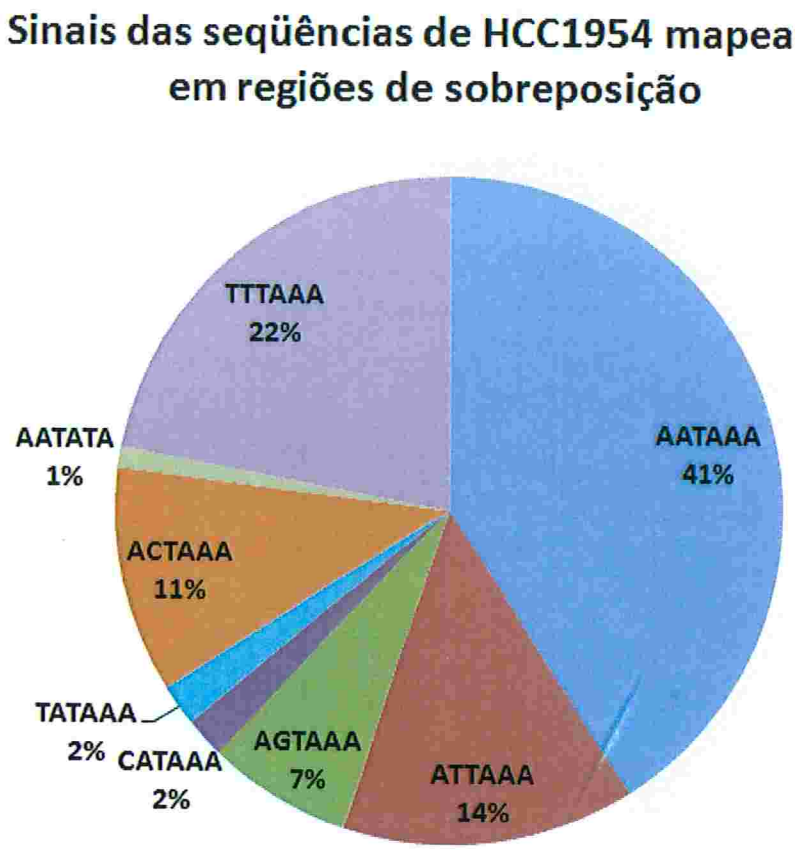


Figura 34: Ocorrência relativa dos sinais das seqüências de HCC1954 em regiões de sobreposição *sense-antisense*.

Numa outra análise, consideramos somente o caso dos variantes de poliadenilação definidos por reads de HCC1954 e calculamos o score de sua seqüência downstream. Na Figura 35 podemos observar o gráfico das médias do score em relação aos sinais encontrados nos variantes de poliadenilação.

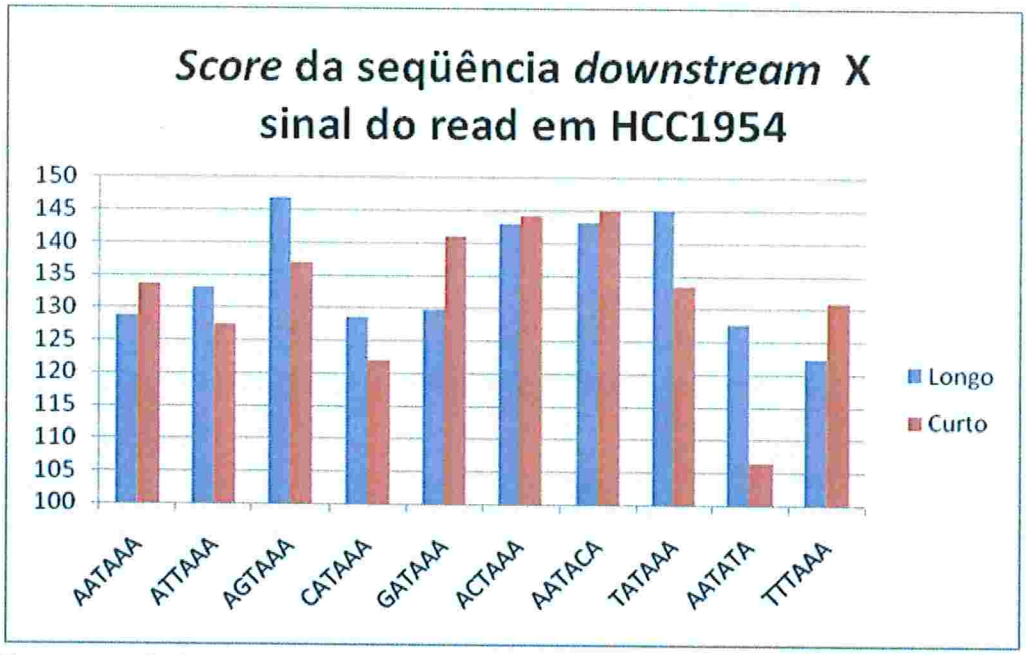


Figura 35: Gráfico das médias do score da região *downstream* rica em G e U em relação ao hexâmero sinal, para variantes de poliadenilação curtos e longos, definidos por reads de HCC1954.

No gráfico acima, assim como para o caso do total de genes poliadenilados, ocorre que quanto mais alto o *score* do elemento *downstream*, maior a chance de ocorrer clivagem e poliadenilação naquele sítio, pois mais forte será a ligação do elemento CstF na seqüência *downstream*, e vice-versa. No gráfico observamos que há pouca variação entre o *score* mais alto e mais baixo, mesmo dividido entre cada sinal de poliadenilação, pois estes já estão selecionados entre os que estão em posição 3' equivalente aos genes de referência e entre os que possuem os dois sinais de poliadenilação: o hexâmero e o sinal *downstream*.

Na Figura 36 realizamos a comparação dos *scores downstream* obtidos a partir das seqüências de referência (apresentados na Figura 22) e dos *scores* obtidos a partir dos reads de HCC1954 (apresentados na Figura 35).

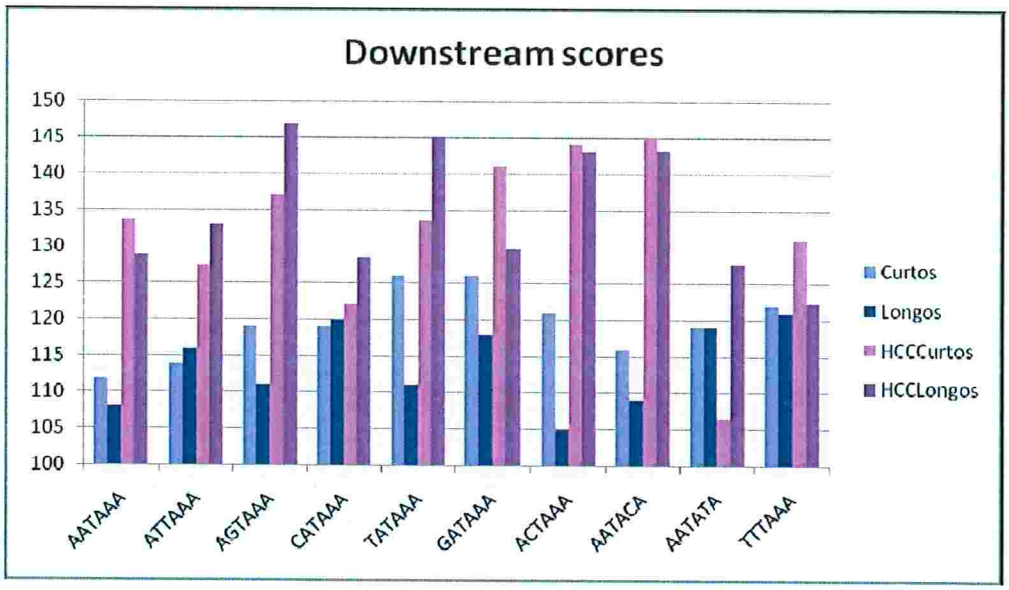


Figura 36: *Downstream scores* dos 10 hexâmeros sinais de poliadenilação em transcritos de referência e em reads de HCC1954, diferenciando entre variantes de poliadenilação curtos e longos.

Neste gráfico temos a comparação entre os hexâmeros versus os *scores* da região *downstream* entre os transcritos de referência e entre os *reads* de HCC1954. É possível observar que os *scores* dos *reads* de HCC1954 são em geral mais altos que os *scores* dos transcritos de referência, o que pode ser explicado pela própria seleção de um número muito pequeno de *reads* de HCC1954 pode ser o motivo dos altos *scores*, pois são os melhores casos observados nesse conjunto de *reads* obtidos de uma linhagem celular.

- **ESTs x HCC1954**

Devido ao fato de que quase todos os *reads* de HCC1954 representam novos variantes de poliadenilação, decidimos comparar suas posições genômicas com ESTs, por que esses novos sítios de poliadenilação podem já estar descritos sob a forma de EST.

Dentre 4151 *reads* de HCC1954 que formam novos variantes de poliadenilação, mais curtos que o maior variante encontrado, 1389 não são descritos por ESTs. Isso significa que estes representam novos variantes, presentes apenas nessa linhagem celular.

Nas seqüências provindas de HCC1954BL, foram utilizadas para a comparação todas as 15.327 seqüências poliadeniladas, sem *internal priming* e sem redundância de posição genômica.

Comparando com os sítios de poliadenilação formados pelo conjunto de RefSeqs, mRNAs e ESTs, 9435 *reads* representam sítio de poliadenilação igual aos já conhecidos. E assim observamos que 5.892 formam novos sítios poli(A), que somente ocorrem na linhagem HCC1954BL.

- **Microarray de HCC1954**

Para confirmar os variantes de poliadenilação encontrados no transcriptoma, também utilizamos o microarray da Affymetrix (modelo U133A), usado para analisar a expressão de 22 mil genes bem caracterizados na amostra. No grupo de 8366 *probesets* que corresponde à terminação dos transcritos envolvidos na poliadenilação alternativa, nós identificamos 970 genes representados por dois ou mais *probesets* e somente 553 estão presentes no conjunto de genes com eventos de poliadenilação alternativa.

Neste subconjunto, 157 desses genes tem pelo menos uma isoforma nos *reads* seqüenciados de HCC1954, mas somente 27 genes eram representados por pelo menos um variante mais curto e um mais longo.

Mostramos na Figura 37 alguns dos genes que tem uma expressão qualitativamente similar no microarray e nos *reads* de 454. A expressão obtida no microarray da Affymetrix é em intensidade RMA normalizada e a do seqüenciamento mostra o número de *reads*.

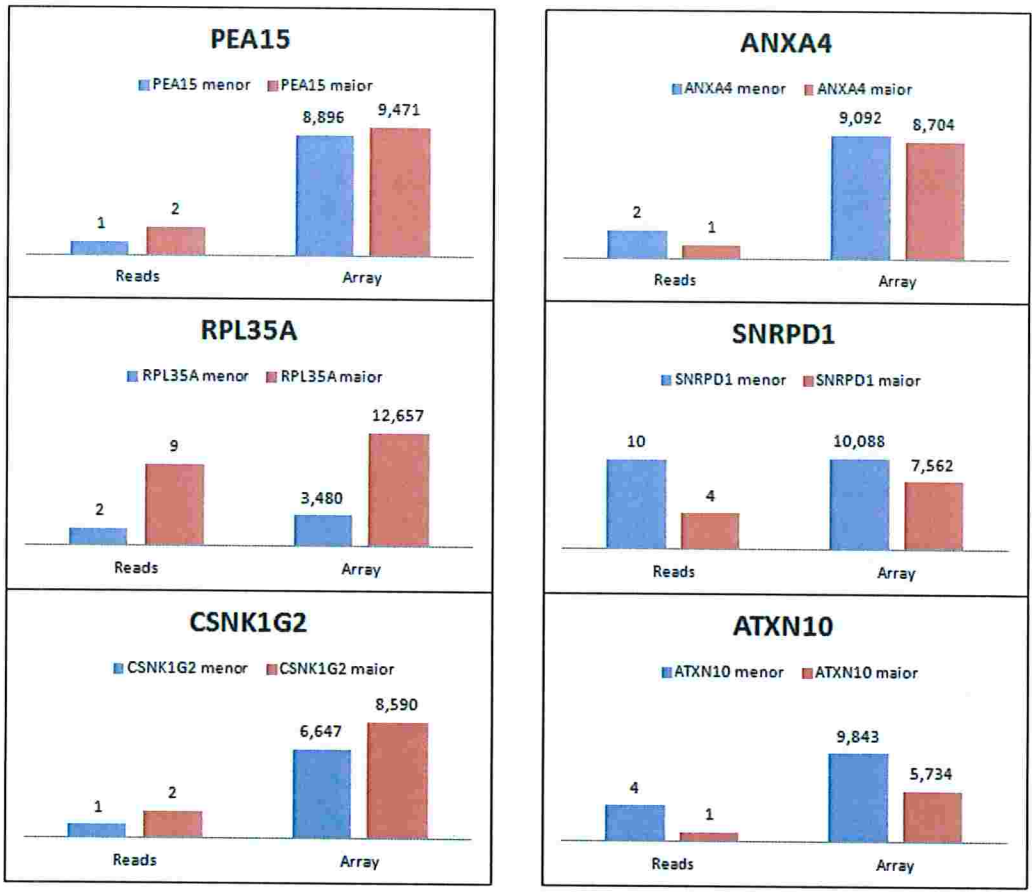


Figura 37: Principais genes nos quais se observa qualitativamente a expressão diferencial entre os variantes de poliadenilação, através do seqüenciamento (à esquerda) e do microarray (à direita). Em azul, o variante mais curto, ou menor, de cada gene. Em vermelho, o variante mais longo, ou maior, de cada gene.

4.13 - Poliadenilação alternativa em astrocitomas de diversos graus

A plataforma Codelink de Análise de Expressão (Codelink Expression Analysis ®) foi utilizada em 19 experimentos de microarray de tecidos de cérebro humanos. Foram analisados tecidos não neoplásicos: 3 amostras de córtex normal e 3 amostras de substância branca normal; bem como tecidos tumorais: 4 amostras de astrocitoma pilocítico (grau I), 6 amostras de astrocitoma de grau II e 3 amostras de glioblastoma (grau IV).

O objetivo desses estudos foi demonstrar que a regulação pós-transcricional também pode ser importante na caracterização do grau de malignidade dos tumores. Foi desenvolvido um algoritmo de bioinformática para mostrar que alguns genes com variantes de poliadenilação têm expressão similar nos tecidos não neoplásicos e podem ter o transcrito mais curto ou o mais longo diferencialmente expresso em astrocitomas.

De todas as 55 mil probes presentes na plataforma Codelink, 52.493 probes puderam ser identificadas de acordo com seu transcrito proveniente através do cruzamento dos dados entre as tabelas obtidas do Ensembl. Entre elas, 44.888 probes foram mapeadas em uma posição única no genoma humano.

Nós utilizamos as posições genômicas das probes para identificar que 37.763 dessas probes estão mapeadas na região genômica de 22.487 genes com transcritos poliadenilados. No entanto, nem todas as probes estavam na região exônica desses genes; somente 22.401 probes localizam-se na região exônica de 20.499 genes com transcritos poliadenilados. Seguindo com a análise, identificamos que somente 14.695 probes estão mapeadas no último exon de seu transcrito designado, o que representa 13.025 genes. Além disso, 4496 transcritos representados no microarray estão presentes em genes com eventos de poliadenilação alternativa, mesmo que não seja possível identificar os variantes do mesmo gene.

Foram selecionados a seguir os genes com eventos de poliadenilação alternativa nos quais existem pelo menos duas probes mapeadas em transcritos

diferentes, de acordo com sua identificação pelo Ensembl. Dessa maneira, selecionamos 157 genes com 2 ou 3 transcritos diferentes, representados por um total de 314 probes. No entanto, essa seleção não garante que as probes identifiquem os transcritos distintos de um mesmo gene, e é necessária mais uma etapa.

No site de Tabelas do UCSC (http://genome.ucsc.edu/cgi-bin/hgTables) foi montada uma *track* customizada, para que pudéssemos localizar cada probe no transcrito respectivo. Assim, foi possível visualizar de forma individual cada probe que era capaz de identificar os transcritos diferentes de um mesmo gene com poliadenilação alternativa, o que resultou em 24 genes selecionados. A maioria das probes que não são capazes de distinguir os variantes de poliadenilação representa somente o transcrito maior.

Para melhor caracterizar os transcritos selecionados foi realizada uma validação *in silico*, selecionando aqueles com maior expressão em GBM comparado ao tecido não neoplásico e aos graus mais baixos de astrocitoma.

Os 24 genes selecionados estão listados na tabela abaixo, com os testes estatísticos. Esses testes foram usados para avaliar a expressão diferencial das probes, entre o variante mais curto e o mais longo. Pudemos selecionar alguns genes nos quais pelo menos uma das probes mostra uma maior expressão nas amostras de tumores mais malignos, o que demonstra que cada variante tem sua própria regulação da expressão.

Os genes com variantes de poliadenilação que possuíam as maiores diferenças de intensidade entre os tumores e o tecido normal foram: SLC30A5, PCDH1, INF2, WWP2, TAF9, GPSM1 e CTSC.

Para mostrar a relevância de nossos resultados, exploramos aqui as características de alguns dos genes mais relevantes, cujos eventos de poliadenilação alternativa são capazes de interferir com as funções celulares.

SLC30A5

Um dos genes interessantes encontrados em nossa análise com uma considerável diferença de intensidade entre as amostras foi o SLC30A5.

O SLC30 é uma das famílias de transportadores de zinco (ZnT) identificados em mamíferos. A alta expressão de proteínas ZnT (ZnT5 e Znt7) nas células epiteliais absortivas do trato gastrointestinal sugere seu importante papel na absorção e na secreção endógena de zinco. ZnT5 e ZnT7 têm a maior similaridade de aminoácidos entre os membros da família SLC30 em mamíferos (Yu, Kirschke, & Huang, 2007)

Em um estudo com câncer de próstata e glioblastoma *in vitro*, sob hipóxia, o zinco leva à diminuição dos níveis de Hypoxia Inducible Factor -1α (HIF-1α), que é responsável pelo "*angiogenic switch*" durante a progressão do tumor. Os níveis aumentados de HIF-1α em tumores levam a um crescimento mais agressivo e à quimioresistência. O zinco induz à degradação proteassômica de HIF-1α e pode ser útil como um inibidor de HIF-1α em tumores humanos por reprimir via importantes envolvidas na progressão tumoral, como as induzidas por VEGF, MDR1 e Bcl2 e potenciar as terapias anticâncer (Nardinocchi, et al., 2010).

O transcrito mais longo tem 765 amino ácidos e o mais curto tem 118 amino ácidos, o que permite somente uma passagem transmembranar. O maior transcrito apresenta 15 domínios transmembranares, segundo a análise por TMHMM (TMHMM). Como o SLC30A5 é um transportador de zinco através da membrana, a perda de domínios transmembranares claramente prejudica sua função. Pudemos observar que o transcrito mais longo é superexpresso em glioblastoma, o que demonstra que o transportador estaria ativo.

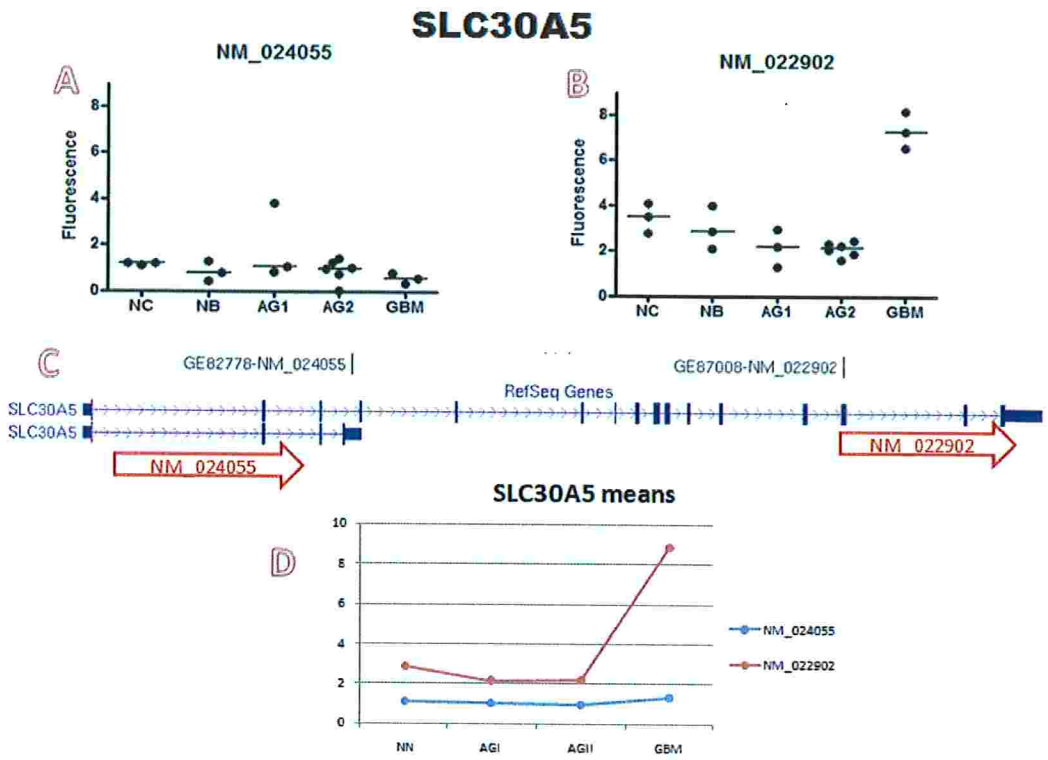


Figura 38: SLC30A5: A e B: Fluorescências das probes de cada transcrito, mostrando todas as amostras analisadas. A barra representa a mediana de cada tipo celular. C: Representação genômica do gene, com a localização das probes que distinguem cada variante. D: Gráfico das médias de fluorescência para cada variante de poliadenilação.

PCDH1 - Protocaderina 1

A Protocaderina 1 é um membro das delta-protocaderinas (PCDHs) não clusterizadas, devido à sua estrutura genômica. As PCDHs não clusterizadas são expressas predominantemente no sistema nervoso central e tem diversos padrões de expressão espaços-temporais (Kim, Chung, Sun, & Kim, 2007).

Os padrões de expressão regionais específicos das PCDHs sugerem seu papel na formação e manutenção de circuitos cerebrais. (Kim, Yasuda, Tanaka, Yamagata, & Kim, 2011).

Considera-se que alguns PCDHs não clusterizados estão envolvidos em doenças neuronais como desordens do espectro do autismo, esquizofrenia e epilepsia feminina. Além disso, alguns PCDHs são genes candidatos a serem supressores de tumor em diversos tecidos (Kim, Yasuda, Tanaka, Yamagata, & Kim, 2011).

Além disso, os genes de delta protocaderina têm ocorrência freqüente de *splicing* alternativo. PCDH1 tem sete domínios caderina extracelulares em sua forma mais longa e cada isoforma de PCDH tem seqüências citoplasmáticas diferentes, o que indica que cada isoforma pode ativar sinalizações intracelulares diferentes.

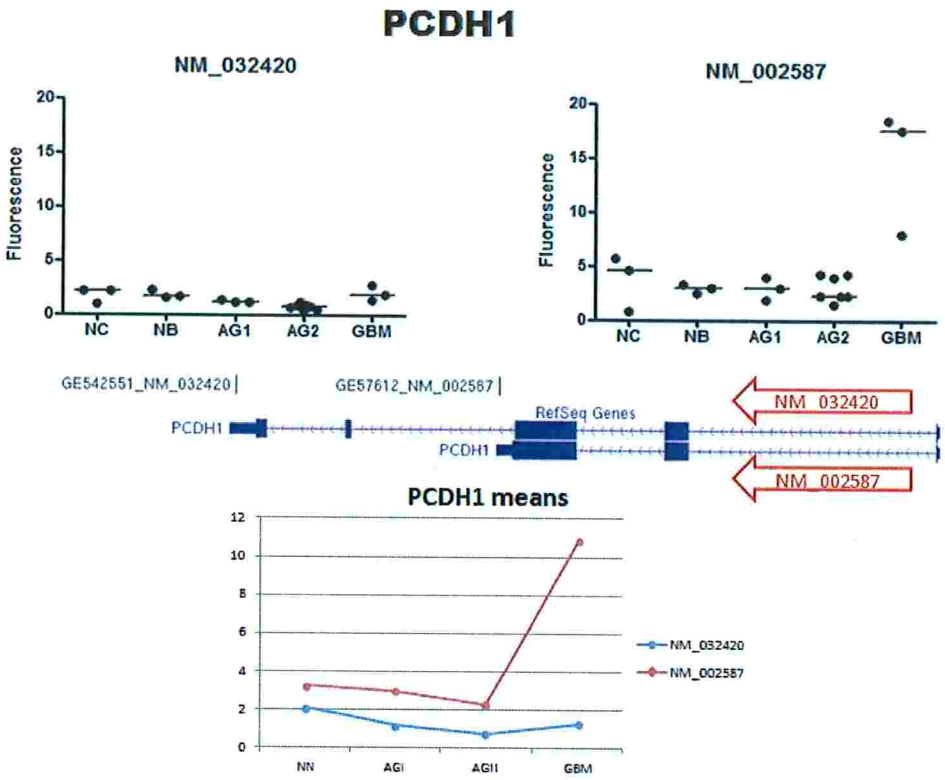


Figura 39: PCDH1: A e B: Fluorescências das probes de cada transcrito, mostrando todas as amostras analisadas. A barra representa a mediana de cada tipo celular. C: Representação genômica do gene, com a localização das probes que distinguem cada variante. D: Gráfico das médias de fluorescência para cada variante de poliadenilação.

GPSM1: G protein signaling modulator 1

As proteínas G propagam o sinal intracelular iniciado pelos receptores a ela acoplados. O GPSM1 é um ativador da sinalização da proteína G independente de receptor, é um dos diversos fatores que influenciam a atividade basal dos sistemas de sinalização de proteína G. Em humanos é conhecida também por AGS3.

A proteína contém sete repetições tetratricopeptídicas (TPR) em seu N terminal e quatro motivos regulatórios de proteína G (GPR) em seu C terminal. Múltiplos transcritos provindos de *splicing* alternativo foram encontrados codificando diferentes isoformas. Os domínios TPR são determinantes do posicionamento da proteína na célula, através de sua interação com ligantes específicos. Foi demonstrado que a perda desses domínios pela introdução de mutações provoca a redistribuição da AGS3 pelo citoplasma (Vural, et al., 2010). Ou seja, no exemplo dos variantes de poliadenilação obtidos acima, haveria uma maior expressão de AGS3 em locais onde esta se encontra ativa no citoplasma, em tumores mais agressivos, como o GBM.

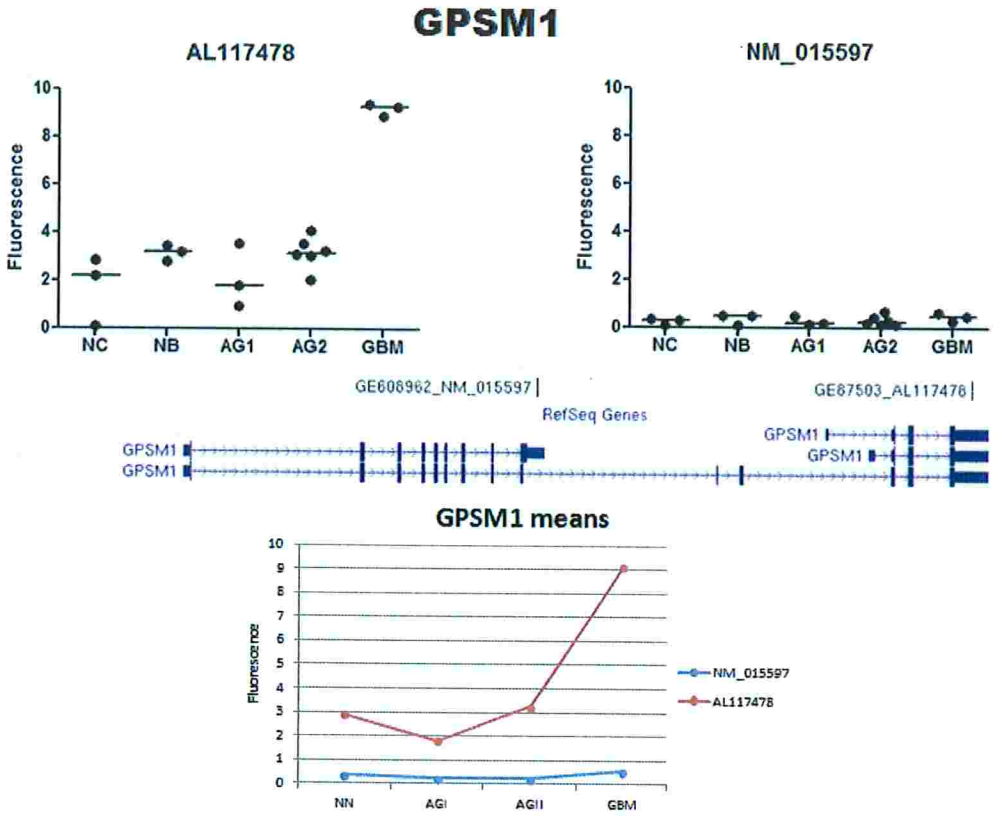


Figura 40: GPSM1: A e B: Fluorescências das probes de cada transcrito, mostrando todas as amostras analisadas. A barra representa a mediana de cada tipo celular. C: Representação genômica do gene, com a localização das probes que distinguem cada variante. D: Gráfico das médias de fluorescência para cada variante de poliadenilação.

5 - Discussão

Esta dissertação foi elaborada de forma a estudar a poliadenilação alternativa nos genes humanos da maneira que nos foi mais ampla e abrangente possível. Diversos métodos de estudo foram testados de forma a obter os resultados mais confiáveis. Os critérios escolhidos para a execução deste trabalho foram sempre no sentido de minimizar a ocorrência de falso-positivos.

Tivemos o cuidado de selecionar os transcritos poliadenilados de forma que os que passavam em cada etapa correspondessem à parte mais confiável de dada subconjunto. Sabemos que, dessa forma, houve uma diminuição do número de casos estudados, ou seja, nossos resultados subestimam o número total da população de nosso objeto de estudo.

Quando lidamos com a poliadenilação, os critérios de seleção utilizados devem levar em conta que podemos observar apenas um trecho pequeno da cauda poli(A), devido às regras de depósito nos bancos de dados públicos. Além disso, deve-se considerar que as seqüências podem ter baixa qualidade em suas extremidades 3' e 5' (Zhang, Hu, Recce, & Tian, 2005) e pode haver a ocorrência de internal priming (Lee, Park, & Tian, 2008).

Diversas publicações recentes examinaram os padrões da poliadenilação usando a informação genômica disponível atualmente e algoritmos de bioinformática, mostrando que mais da metade dos genes de mamíferos é sujeito à poliadenilação alternativa e que muitos sinais alternativos são evolucionariamente conservados (Lutz, 2008).

Num estudo sobre a conservação dos sítios entre humanos e camundongos, observou-se 27.654 sítios poli(A) em 14.574 genes humanos, o que totaliza aproximadamente 1,8 sítios por gene (Ara, Lopez, Ritchie, Benech, & Gautheret, 2006).

Diversos grupos estudaram a poliadenilação alternativa nos genes humanos, com resultados diferentes de acordo com o critério adotado. Trabalhos mais antigos estimaram que 29% dos genes sofrem poliadenilação alternativa (Beaudoing & Gautheret, 2001). Iseli et al. (2002) estudaram os genes anotados em um longo *contig*, e verificaram que metade dos genes possuía múltiplos sítios

de poliadenilação, com diferenças de terminação entre 300nt e 15kb, independente do número de exons desses genes (Iseli, et al., 2002).

Um banco de dados de poliadenilação (PolyA_DB) encontrou 7524 genes humanos com sítios poli(A) alternativos, equivalente a 54% dos genes (Zhang, Hu, Recce, & Tian, 2005) (Tian, Hu, Zhang, & Lutz, 2005).

Neste trabalho foram analisadas 81.664 seqüências, incluindo mRNAs e RefSeqs, e foram agrupados em 32.136 clusters, sendo que somente metade deles possui mais de 2 transcritos e somente 15 mil possuem nomes de genes segundo o MGC. Definimos que 7070 genes apresentam eventos de poliadenilação alternativa, com variantes de ao menos 25 nucleotídeos.

Podemos observar que obtivemos porcentagem semelhante de genes com sítios poli(A) alternativos (47%), considerando que os estudos anteriores também utilizaram somente genes que possuíam nomes padronizados. Notamos também que não há padronização de quantos nucleotídeos definem um variante de poliadenilação, e que nosso critério foi baseado na região onde é encontrado o sinal de poliadenilação.

Estudos anteriores, utilizando bancos de dados de ESTs, observaram um número muito menor, de 29% dos genes (Beaudoing, Freier, Wyatt, Claverie, & Gautheret, 2000). Essa diferença pode ser atribuída ao número de seqüências analisadas em cada caso. À época do estudo mais antigo, o total de seqüências usadas era pouco mais de 157 mil, enquanto que no estudo de 2005 foram analisadas 397 mil seqüências.

Assim é possível observar que a diferença quantitativa dos genes com poliadenilação alternativa se deve ao número de seqüências analisadas.

Uma característica interessante entre os diversos estudos do sinal de poliadenilação é que a porcentagem de cada sinal é mantida, independente do número de seqüências utilizadas.

Tian et al.(2005) demonstraram que em sítios poli(A) constitutivos e conservados entre humanos e camundongos, 70% possuem o hexâmero AATAAA, 15% ATTAAA e 12% dos sítios possuem outros 10 tipos de hexâmeros. Além disso, foi observado que as regiões *downstream* são mais ricas em uridinas quanto mais conservados forem os sítios entre humanos e camundongos, um

indicativo de elemento *downstream* (DSE) forte (Ara, Lopez, Ritchie, Benech, & Gautheret, 2006).

No entanto as porcentagens gerais para humanos são diferentes quando não se considera a comparação entre humanos e camundongos. O sinal AATAAA ocorre em 53,2% dos transcritos estudados em (Tian, Hu, Zhang, & Lutz, 2005) e em 58,2% dos transcritos em (Beaudoing, Freier, Wyatt, Claverie, & Gautheret, 2000). Neste trabalho obtivemos 46% do uso do hexâmero AATAAA no total de transcritos, após os filtros do alinhamento contra o genoma.

O segundo hexâmero canônico ATTAAA ocorre em 17% dos casos em (Tian, Hu, Zhang, & Lutz, 2005) e em 14,9% dos mRNAs em (Beaudoing, Freier, Wyatt, Claverie, & Gautheret, 2000). Neste trabalho obtivemos 15%.

Assim, nossos resultados reproduziram os dados da literatura.

Como a ocorrência relativa dos sinais é mantida, podemos especular que isso se deva a alguma pressão seletiva, para que a poliadenilação ocorra sempre em determinada proporção, ou para que alguns dos sinais de poliadenilação não se percam.

Ao analisar os sinais simples, duplos e triplos encontrados no conjunto de RefSeqs e mRNAs. Observamos que as proporções dos sinais *downstream* são as que mais se assemelham às dos transcritos com apenas um sinal. Podemos cogitar que o sinal mais *downstream* é o sinal funcional, mesmo em transcritos com vários sinais.

No grupo dos variantes mais curtos, 20% dos transcritos não possuíam nenhum dos 10 sinais procurados, enquanto que no grupo dos variantes mais longos, 11% dos transcritos não possuíam sinal.

Podemos especular que caso a poliadenilação não seja feita na posição do sítio mais curto devido à inexistência ou à existência de sinal mais fraco, a presença de maior porcentagem de sinal no variante mais longo garantiria a poliadenilação daquele gene na posição do sítio mais longo.

A descoberta de que a região 3'UTR contém elementos de regulação cruciais (para a expressão dos genes) oferece uma nova dimensão ao estudo dos cânceres, doenças auto-imunes e inflamatórias. Diversos campos de pesquisa foram abertos para o estudo das seqüências em *cis* (ARE) e dos fatores em *trans*, as proteínas ligantes de AU e os miRNAs (Nguyen-Chi & Morello, 2008).

O estudo das mutações não deve se limitar à região codificante dos genes, mas também às regiões 3'UTR. Tal análise deve levar em conta a possível existência de diversos sítios de poliadenilação num mesmo mRNA, e que uma região 3' não traduzida, maior ou menor, pode ser mais ou menos rica em ARE e em sítios de fixação de pequenos RNAs não codificantes.

Podemos observar também que com o exemplo das doenças causadas pela perda do sinal de poliadenilação, como as talassemias, há pouca flexibilidade de seqüência em relação ao hexâmero sinal (Danckwardt, Hentze, & Kulozik, 2008). Uma única mudança de base é capaz de impedir a ligação do sinal com o CPSF e assim impedir a clivagem e poliadenilação naquele sítio.

Ao avaliar os bancos de dados de SNPs, pudemos constatar que muitos SNPs podem ser encontrados na mesma posição genômica dos sinais de poliadenilação. No entanto, ao verificar, na seqüência dos transcritos de referência utilizados em nossas análises, RefSeqs, mRNAs e ESTs, não foi possível encontrar um SNP que modificasse o sinal nas posições indicadas pelo banco de dados. Isso era esperado, pois são seqüências de referência, que apresentam o nucleotídeo principal do SNP e não o alternativo.

O mesmo não ocorreu ao analisar as seqüências das células de cultura de HCC1954 e HCC1954BL, com as quais pudemos comprovar o método, observando a mudança de base nas posições indicadas pelo banco de dados. No entanto, não pudemos observar um SNP que destruísse completamente o sítio de poliadenilação, pois somente observamos SNPs em sinais muito próximos ou lado a lado (em *tandem*).

É possível observar que em diversos casos de SNP no sinal, a mudança de base produz um dos sinais canônicos AATAAA ou ATTAAA, que aparecem em uma proporção semelhante à observada entre os sinais afetados por SNP. Podemos especular que há uma pressão seletiva que mantém um sinal funcional mesmo na presença de um SNP naquela posição, ou mesmo, que o SNP somente aparece onde há múltiplos sinais de poliadenilação lado a lado, para que não se perca a sua funcionalidade.

No caso dos variantes de poliadenilação, comparamos a incidência de SNPs em sinais da extremidade de variantes curtos e longos, além do total de SNPs que colocalizam com sinais de poliadenilação. Assim, podemos observar que a incidência de SNPs em sinais de poliadenilação mantém a freqüência esperada

de sua presença no transcriptoma. No entanto, os SNPs afetam em maior proporção os sinais canônicos dos variantes curtos que dos variantes longos. Isso pode ser indício de alguma pressão seletiva, para que, se algum sinal tenha que ser modificado, que este seja o do variante mais curto, para que a poliadenilação continue sendo feita no sítio onde se encontra o variante mais longo e que o transcrito não seja perdido.

O mesmo raciocínio pode ser feito para os genes com sobreposição 3´-3´. Os genes em senso mantêm a proporção de sinais dos genes com poliadenilação alternativa, enquanto que os genes sobrepostos, em antissenso, mantêm proporção semelhante ao total de genes poliadenilados. A partir dessa observação, podemos imaginar que a clivagem e poliadenilação sejam realizadas no gene senso de maneira normal e mais freqüente, que possui maior ocorrência de sinais canônicos. A regulação pelo antissenso seria apenas realizada quando fatores externos promovessem a expressão concomitante do gene em antissenso, que possui uma proporção maior de sinais mais fracos.

RNA-Seq versus microarrays

Metade dos genes humanos tem sítios alternativos de poliadenilação e a escolha do sítio depende do tecido onde o gene é expresso (Zhang, Lee, & Tian, 2005).

As linhagens celulares linfoblastóides (HCC1954BL) são geradas pela transformação dos linfócitos B periféricos pelo vírus Epstein-Barr (EBV). Essas células são de fácil manutenção e tem uma taxa de mutação somática de 0,3%.

O RNA dessas células tem sido usado para avaliar a mutação, e já foi usado no estudo de um sítio de splicing do gene ALS2 e do gene ABCA4. No entanto, os variantes de splicing devem ser avaliados com cuidado, pois o mecanismo varia de acordo com o tipo celular.

Por outro lado, o transcriptoma da linhagem HCC1954 foi obtido a partir de células de tumor de mama, da mesma paciente de onde se obteve a linhagem linfoblastóide. No entanto, o transcriptoma dos dois tipos de culturas celulares não é comparável, devido ao fato de terem sido imortalizados por métodos diferentes. As células HCC1954 foram imortalizadas por serem provenientes de câncer de

mama, enquanto as células HCC1954BL foram imortalizadas em cultura primária com o vírus Epstein-Barr (EBV).

O transcriptoma das linhagens celulares foi obtido por RNA-Seq, e após o alinhamento com o genoma humano, obtivemos uma boa representação da posição dos reads junto aos genes com variantes de poliadenilação.

Em comparação com os dados de microarray da plataforma Codelink, pudemos constatar uma grande diferença de representatividade entre a localização das probes e dos reads nos variantes de poliadenilação.

Embora a maior parte das probes da plataforma Codelink estejam localizadas na região 3' dos genes conhecidos, há uma subrepresentação dos variantes de poliadenilação, dado que o microarray não foi originalmente projetado para este fim.

As diferenças entre o número de genes com variantes de poliadenilação entre o RNA-Seq e o microarray utilizados se deve principalmente ao número de genes inicial. No microarray tínhamos a representação de 22 mil genes sendo que a posição genômica das probes era conhecida. No entanto, somente um número ínfimo poderia representar os variantes de poliadenilação de forma fidedigna. No RNA-Seq, realizamos o alinhamento dos reads e dessa forma pudemos obter um número muito superior de reads em genes com eventos de poliadenilação alternativa. Sendo assim, obtivemos aproximadamente 553 genes com variantes de poliadenilação no microarray, e por volta de 2584 genes com bom alinhamento e mesmo nome segundo o MGC, no RNA-Seq.

A expressão gênica é freqüentemente quantificada pelos níveis de mRNA. No entanto, há uma incerteza sobre quão fidedignos os níveis de mRNA se relacionam com os níveis das proteínas correspondentes. Muitos autores sugerem apenas uma leve correlação.

Ao constatar que existe expressão diferencial entre variantes de poliadenilação de alguns genes, abre-se a possibilidade para que as técnicas de diagnóstico levem em conta esses transcritos na detecção do câncer.

Essa pesquisa pode abrir portas para descobertas futuras que levem ao diagnóstico precoce dos astrocitomas, ao avaliar a expressão dos variantes de poliadenilação de determinados genes.

6 - Bibliografia

Adereth, Y., Dammai, V., Kose, N., Li, R., & Hsu, T. (2005). RNA-dependent integrin α3 protein localization regulated by the Muscleblind-like protein MLP1. *Nature Cell Biology , 7(12)*, 1240-1247.

Anant, S., Houchen, C. W., Pawar, V., & Ramalingam, S. (2010). Role of RNA-binding proteins in colorectal carcinogenesis. *Curr Colorectal Cancer , 6* (2), 68-73.

Andreassi, C., & Riccio, A. (2009). To localize or not to localize: mRNA fate is in 3'UTR ends. *Cell* .

Annovazzi, L., Mellai, M., Caldera, V., Valente, G., & Schiffer, D. (2011). SOX2 Expression and Amplification in Gliomas and Glioma Cell Lines. *Cancer Genomics and Proteomics , 8* (3), 139-147.

Ara, T., Lopez, F., Ritchie, W., Benech, P., & Gautheret, D. (2006). Conservation of alternative polyadenylation patterns in mammalian genes. *BMC Genomics , 7* (189).

Aulak, K., Mishra, R., Zhou, L., Hyatt, S., de Jonge, W., Lamers, W., et al. (1999). Post-transcriptional regulation of the arginine transporter Cat-1 by amino acid availability. *Journal of Biological Chemistry , 274* (43), 30424-32.

Bakheet, T., Williams, B. R., & Khabar, K. S. (2006). ARED 3.0: the large and diverse AU-rich transcriptome. *Nucleic Acids Research , 34* (Database Issue), D111-D114.

Bashirullah, A., Cooperstock, R., & Lipshitz, H. (2001). Spatial and temporal control of RNA stability. *PNAS , 98* (13), 7027.

Beaudoing, E., & Gautheret, D. (2001). Identification of alternative polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Research , 11*, 1520-1526.

Beaudoing, E., Freier, S., Wyatt, J., Claverie, J., & Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. *Genome Research* , 10(7):1001-10.

Bevilacqua, A., Ceriani, M. C., Capaccioli, S., & Nicolin, A. (2003). Post-transcriptional regulation of gene expression by degradation of messenger RNAs. *Journal of Cellular Physiology , 195*, 356-372.

Boguski MS, L. T. (1993). dbEST--database for "expressed sequence tags". *Nat Genet* , 4(4):332-3.

Brown, K. M., & Gilmartin, G. M. (2003). A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im. *Molecular Cell , 12* (6), 1467-1476.

Cañadillas, J. M., & Varani, G. (2003). Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *The EMBO Journal , 22*, 2821-2830.

Caput, D., Beutler, B., Hartog, K., Thayer, R., Brown-Shimer, S., & Cerami, A. (1986). Identification of a Common Nucleotide Sequence in the 3'-Untranslated Region of mRNA Molecules Specifying Inflammatory Mediators. *PNAS , 83* (6), 1670-1674.

Chatterjee, S., & Pal, J. K. (2009). Role of 5'- and 3'- untranslated regions of mRNAs in human diseases. *Biol. Cell , 101*, 251-262.

Colgan, D., & Manley, J. (1997). Mechanism and regulation of mRNA polyadenylation. *Genes & Development , 11* (21), 2755-66.

Dahary, D., Elroy-Stein, O., & Sorek, R. (2005). Naturally occurring antisense: Transcriptional leakage or real overlap? *Genome Research , 15*, 364-368.
Danckwardt, S., Hentze, M., & Kulozik, A. (2008). 3'end mRNA processing: molecular mechanisms and implications for health and disease. *The EMBO Journal , 27*, 482-498.
de Vries, H., Rüegsegger, U., Hübner, W., Friedlein, A., Langen, H., & Keller, W. (2000). Human pre-mRNA cleavage factor IIm contains homologs of yeast proteins and bridges two other cleavage factors. *The EMBO Journal , 19*, 5895-5904.
Denome, R., & Cole, C. (1988). Patterns of polyadenylation site selection in gene constructs containing multiple polyadenylation signals. *Mol Cell Biol , 8(11)*, 4829-4839.
Edwalds-Gilbert, G., Veraldi, K., & Milcarek, K. (1997). Alternative poly(A) site selection in complex transcription units: Means to an end? *Nucleic Acids Research , 25*, 2547-2561.
Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research* , 967-974.
Fuke, H., & Ohno, M. (2008). Role of poly(A) tail as an identity element for mRNA nuclear export. *Nucleic Acids Research , 36* (3), 1037-1049.
Galante, P. A., Vidal, D. O., de Souza, J. E., Camargo, A. A., & de Souza, S. J. (2007). Sense-antisense pairs in mammals: functional and evolutionary considerations. *Genome Biology , 8*.
Gieselmann, V., Polten, A., Kreysing, J., & von Figura, K. (1989). Arylsulfatase A pseudodeficiency: loss of a polyadenylylation signal and N-glycosylation site. *Proc Natl Acad Sci USA , 86* (23), 9436-40.
Gilmartin, G. M. (2005). Eukaryotic mRNA 3' processing: a common means to different ends. *Genes & Development , 19*, 2517-2521.
Hall-Pogar, T., Zhang, H., Tian, B., & Lutz, C. S. (2005). Alternative polyadenylation of cyclooxygenase-2. *Nucleic Acids Research , 33* (8), 2565-2579.
Harvey, J., Carey, W., & Morris, C. (1998). Importance of the glycosylation and polyadenylation variants in metachromatic leukodystrophy pseudodeficiency phenotype. *Human molecular genetics , 7*, 1215-1219.
Higgs, D., Goodbourn, S., Lamb, J., Clegg, J., Weatherall, D., & Proudfoot, N. (1983). Alpha-thalassaemia caused by a polyadenylation signal mutation. *Nature , 306* (5941), 398-400.
Holmberg, J., He, X., Peredo, I., Orrego, A., Hesselager, G., Ericsson, C., et al. (2011). Activation of neural and pluripotent stem cell signatures correlates with increased malignancy in human glioma. *PLoS One , 6* (3), e18454.
Hu, J., Lutz, C. S., Wilusz, J., & Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA , 11*, 1485-1493.
Huang, Y., & Carmichael, G. G. (1996). Role of polyadenylation in nucleocytoplasmic transport of mRNA. *Molecular and Cellular Biology , 16* (4), 1534-1542.
Iseli, C., Stevenson, B., de Souza, S., Samaia, H., Camargo, A., Buetow, K., et al. (2002). Long-range heterogeneity at the 3' ends of human mRNAs. *Genome Research , 12*, 1068-1074.
Jambhekar, A., & Derisi, J. (2007). Cis-acting determinants of asymmetric, cytoplasmic RNA transport. *RNA , 13*, 625-642.

Kent WJ, S. C. (2002). The human genome browser at UCSC. *Genome Research* , 996-1006.
Kent, W. J. (2002). BLAT - The BLAST-like Alignment Tool. *Genome Research , 12*, 656-664.
Khabar, K. S. (2005). The AU-Rich transcriptome: More than interferons and cytokines and its role in disease. *Journal of interferon & cytokine research , 25*, 1-10.
Kim, S., Chung, H., Sun, W., & Kim, H. (2007). Spatiotemporal expression pattern of non-clustered protocadherin family members in the developing rat brain. *Neuroscience , 147* (4), 996-1021.
Kim, S., Yasuda, S., Tanaka, H., Yamagata, K., & Kim, H. (2011). Non-clustered protocadherin. *Cell Adh Migr , 5* (2), 97-105.
Kislauskis, E., Zhu, X., & Singer, R. (1997). beta-Actin messenger RNA localization and protein synthesis augment cell motility. *Journal of Cell Biology , 136* (6), 1263-70.
Kozak, M. (2004). How strong is the case for regulation of the initiation step of translation by elements at the 3' end of eukaryotic mRNAs? *Gene* , 41-54.
Kress, T., Yoon, Y., & Mowry, K. (2004). Nuclear RNP complex assembly initiates cytoplasmic RNA localization. *J Cell Biology , 165*, 203-211.
Kumar, M., & Carmichael, G. (1998). Antisense RNA: Function and Fate of Duplex RNA in Cells of Higher Eukaryotes. *Microbiology and molecular biology reviews* , 1415-1434.
Lavorgna, G., Dahary, D., Lehner, B., Sorek, R., Sanderson, C., & Casari, G. (2004). In search of antisense. *Trends in Biochemical Science , 29* (2), 88-94.
Lee, J. Y., Park, J. Y., & Tian, B. (2008). Identification of mRNA Polyadenylation Sites in Genomes Using cDNA Sequences, Expressed Sequence Tags, and Trace. *Methods in Molecular Biology* , 23-37.
Lin, H., Huang, L., Su, H., & Jeng, S. (2009). Effects of the multiple polyadenylation signal AAUAAA on mRNA 3'-end formation and gene expression. *Planta , 230(4)*, 699-712.
Lipman, D. J. (1997). Making (anti)sense of non-coding sequence. *Nucleic Acids Research , 25* (18), 3580-3583.
Losekoot, M., Fodde, R., Harteveld, C., van Heeren, H., Giordano, P., Went, L., et al. (1991). Homozygous beta+ thalassaemia owing to a mutation in the cleavage-polyadenylation sequence of the human beta globin gene. *Journal of Medical Genetics , 28* (4), 252-255.
Louis, D., Ohgaki, H., Wiestler, O., Cavenee, W., Burger, P., & al, e. (2007). WHO Classification of Tumours of the Central Nervous System. *Acta Neuropathology , 114*, 97-109.
Lutz, C. S. (2008). Alternative Polyadenylation: A twist on mRNA 3' end formation. *ACS Chemical Biology , 3* (10), 609-617.
Mandel C.R., B. Y. (2007). Protein factors in pre-mRNA 3'-end processing. *Cell. Mol. Life Sci.*
Mangus, D., Evans, M., & Jacobson, A. (2003). Poly(A) binding proteins: Multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biology , 4* (7), 223.
Martin, K. C., & Ephrussi, A. (2009). mRNA Localization: Gene expression in the spatial dimension. *Cell , 136*, 719-730.
Martincic, K., Campbell, R., Edwalds-Gilbert, G., Souan, L., Lotze, M., & Milcarek, C. (1998). Increase in the 64-kDa subunit of the polyadenylation/cleavage

stimulatory factor during the G0 to S phase transition. *Proc. Natl. Acad. Sci USA , 95*, 11095-11100.
Missier, P., Embury, S., & Stapenhurst, R. (2008). Exploiting Provenance to Make Sense of automated decisions in scientific workflows. *IPAW2008, LNCS 5272* , 174-185.
Munroe, D., & Jacobson, A. (1990). mRNA poly(A) tail, a 3' enhancer of translational initiation. *Molecular Cell Biology , 10* (7), 3441-3455.
Nagaraj, S. H., Gasser, R. B., & Ranganathan, S. (2006). A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics , 8* (1), 6-21.
Nardinocchi, L., Pantisano, V., Puca, R., Porru, M., Aiello, A., Grasselli, A., et al. (2010). Zinc downregulates HIF-1α and inhibits its activity in tumor cells *in vitro* and in vivo. *PLoS One , 5* (12), e15048.
Natalizio, B., Muniz, L., Arkin, G., Wilusz, J., & Lutz, C. (2002). Upstream elements present in the 3'-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals. *Journal of Biological Chemistry , 277* (45), 42733-40.
Neugebauer, K. (2002). On the importance of being co-transcriptional. *Journal of cell science* , 115(Pt 20):3865-71.
Nguyen-Chi, M., & Morello, D. (2008). Contribution de la régulation post-transcriptionnelle à l'émergence de maladies. *Medecine/Sciences* , 290-296.
Orkin, S., Cheng, T., Antonarakis, S., & Kazazian, H. J. (1985). Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. *EMBO Journal , 4* (2), 453-456.
Pandit, S., Wang, D., & Xiang-Dong, F. (2008). Functional integration of transcriptional and RNA processing machineries. *Current Opinion in Cell Biology , 20*, 260-265.
Prévôt, D., Darlix, J.-L., & Ohlmann, T. (2003). Conducting the initiation of protein synthesis: the role of eIF4G. *Biology of the cell , 95*, 141-156.
Rüegsegger U, B. K. (1996). Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. *J Biol Chem.* , 271(11):6107-13.
Rund, D., Dowling, C., Najjar, K., Rachmilewitz, E., Kazazian, H. J., & Oppenheim, A. (1992). Two mutations in the beta-globin polyadenylylation signal reveal extended transcripts and new RNA polyadenylylation sites. *Proc Natl Acad Aci USA , 89* (10), 4324-4328.
Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., & Burge, C. B. (2008). Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science , 320*.
Scorilas, A. (2002). Polyadenylate Polymerase (PAP) and 3' End pre-mRNA Processing: Function, Assays and Association with Disease. *Critical Reviews in ClinicalLaboratory Sciences , 39* (3), 193-224.
Sheets MD, O. S. (1990). Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation *in vitro. Nucleic Acids Res.* , 18(19):5799-805.
Sherry, S., Ward, M., Kholodov, M., Baker, J., Smigielski, E., & Sirotkin, K. (2001). dbSNP:the NCBI database of genetic variation. *Nucleic Acids Research* , 308-311.
Sie, L., Loong, S., & Tan, E. (2009). Utility of Lymphoblastoid Cell Lines. *Journal of Neuroscience Research , 87*, 1953-1959.
Sun, G., Wang, Y., Sun, L., Luo, H., Liu, N., Fu, Z., et al. (2011). Clinical significance of Hiwi gene expression in gliomas. *Brain research , 1373*, 183-188.

Sun, M., Hurst, L., Carmichael, G., & Chen, J. (2005). Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts. *Nucleic Acids Research , 33* (17), 5533-5543.
Tabaska, J. E., & Zhang, M. Q. (1999). Detection of polyadenylation signals in human DNA sequences. *Gene , 231*, 77-86.
Tan, W.-C. (2004). Research Problems in Data Provenance. *IEEE Data Engineering Bulletin , 27* (4), 45-52.
Tarun Jr, S., Wells, S., Deardoff, J., & Sachs, A. (1997). Translation initiation factor eIF4G mediates *in vitro* poly(A) tail- dependent translation. *Proc. Natl. Acad. Sci. , 94*, 9046–9051.
Tian, B., Hu, J., Zhang, H., & Lutz, C. S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research , 33*, 201-212.
TMHMM. (s.d.). Fonte: www.cbs.dtu.dk/services/TMHMM/
van Hoof, A., & Parker, R. (2002). Messenger RNA Degradation: Beginning at the end. *Current Biology , 12*, R285-R287.
Vanhée-Brossollet, C., & Vaquero, C. (1998). Do natural antisense transcripts make sense in eukaryotes? *Gene , 211* (1), 1-9.
Velculescu VE, Z. L. (1995). Serial analysis of gene expression. *Science* , 270(5235):484-7.
Venkataraman, K., Brown, K., & Gilmartin, G. (2005). Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes & Development , 19*, 1315-1327.
Vural, A., Oner, S., An, N., Simon, V., Ma, D., Blumer, J. B., et al. (2010). Distribution of Activator of G-Protein Signaling 3 within the Aggresomal Pathway: Role of Specific Residues in the Tetratricopeptide Repeat Domain and Differential Regulation by the AGS3 Binding Partners Giα and Mammalian Inscuteable. *Molecular and Cellular Biology , 30* (6), 1528-1540.
Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature , 456* (7221), 470-476.
Wells, S. E., Hillner, P. E., Vale, R. D., & Sachs, A. B. (1998). Circularization of mRNA by eukaryotic translation initiation factors. *Molecular Cell , 2* (1), 135-140.
Werner, A., & Sayer, J. A. (2009). Naturally occurring antisense RNA: function and mechanisms of action. *Current Opinion in Nephrology and Hypertension , 18* (Renal pathophysiology), 343-349.
Yelin, R., Dahary, D., Sorek, R., Levanon, E., Goldstein, O., Shoshan, A., et al. (2003). Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnology , 21* (4), 379-386.
Young, L. E., & Dixon, D. A. (2010). Posttranscriptional regulation of cyclooxygenase 2 expression in colorectal cancer. *Curr Colorectal cancer , 6* (2), 60-67.
Yu, Y., Kirschke, C., & Huang, L. (2007). Immunohistochemical analysis of ZnT1, 4, 5, 6, and 7 in the mouse gastrointestinal tract. *J Histochem Cytochem , 55* (3), 223-234.
Zarudnaya, M., Kolomiets, I., Potyahaylo, A., & Hovorun, D. (2003). Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. *Nucleic Acids Research , 31* (5), 1375-1386.

Zhang, H., Hu, J., Recce, M., & Tian, B. (2005). PolyA-DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Research , 33* (Database issue).
Zhang, H., Lee, J. Y., & Tian, B. (2005). Biased alternative polyadenylation in human tissues. *Genome Biology , 6* (12).
Zhao J, H. L. (1999). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev.* , 63(2):405-45.
Zhao, Q., Caballero, O., Levy, S., Stevenson, B., Iseli, C., de Souza, S., et al. (2009). Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc Natl Acad Sci USA , 106* (6), pp. 1886-91.
Zhao, W., & Manley, J. (1998). Deregulation of Poly(A) Polymerase Interferes with Cell Growth. *Molecular and cellular biology , 18* (9), 5010-5020.
Zhu, Y., States, J., Wang, Y., & Hein, D. (2011). Functional effects of genetic polymorphisms in the N-acetyltransferase 1 coding and 3' untranslated regions. *Birth Defects Res A Clin Mol Teratol , 91* (2), 77-84.
Zlotorynski, E., & Agami, R. (2008). A PASport to cellular proliferation. *Cell , 134*, 208-210.

Esta tese foi confeccionada de acordo com as Diretrizes para apresentação de dissertações e teses da USP da Associação Brasileira de Normas Técnicas (ABNT) e com a Reforma Ortográfica da Língua Portuguesa de 1971.