ELISA NAPOLITANO E FERREIRA

Identificação de variantes de *splicing* sob influência da alta expressão do oncogene *ERBB2* em câncer de mama

São Paulo

2010

Elisa Napolitano e Ferreira

Identificação de variantes de *splicing* sob influência da alta expressão do oncogene *ERBB2* em câncer de mama

Tese apresentada ao Instituto de Biociências da Universidade de São Paulo, para a obtenção de Título de Doutor em Ciências, na Área de Ciências Biológicas (Biologia Genética).

Orientadora: Dra. Dirce Maria Carraro

Co-orientador: Sandro José de Souza

São Paulo

2010

Ferreira, Elisa Napolitano e
Abordagens para identificação de variantes de *splicing* associadas ao câncer de mama sob influência da alta expressão do oncogene *ERBB2*.
163pgs + anexos

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

1. *Splicing* alternativo 2. Bibliotecas de cDNA 3. *ERBB2*

Universidade de São Paulo. Instituto de Biociências. Departamento de Biologia Genética.

Comissão Julgadora:

_____ _____

Prof(a). Dr(a). Prof(a). Dr(a).

_____ _____

Prof(a). Dr(a). Prof(a). Dr(a).

_____

Dra. Dirce Maria Carraro
Orientador(a)

À meus pais,Ana Maria e José Carlos

pelo carinho e apoio, sempre.

À minha irmã, Mariana, pelo companheirismo e por me levar para

o mundo da Biologia.

Ao Juba, pelo amor e incentivo que me mantém firme no meu caminho.

Á meu avô, Aymoré, pelo exemplo de vida.

Á minha querida avó, Maria do Rosário,

Saudades...

A coisa mais bela que o homem pode experimentar é o mistério. É esta a emoção fundamental que está na raiz de toda ciência e arte. O homem que desconhece esse encanto, incapaz de sentir admiração e estupefação, esse já está, por assim dizer, morto e tem os olhos extintos.

Albert Eisntein

Agradecimentos

À minha orientadora, Dra. Dirce Maria Carraro, pelo grande aprendizado ao longo desses oito anos. Agradeço pela confiança, pelo incentivo, pela amizade e, principalmente, por tantas oportunidades.

Ao meu co-orientador, Dr. Sandro José de Souza, por todo apoio, pelos conselhos e pela amizade.

À Maria Cristina Rangel e ao Gustavo Molina pelas discussões e excelentes sugestões que muito contribuíram para o desenvolvimento deste trabalho.

À Mariana Maschietto por estar sempre tão disposta em ajudar. Obrigada pelo apoio e amizade durante todo o processo de elaboração dessa tese, desde o projeto até a escrita final.

Aos demais colegas do Laboratório de Genômica e Biologia Molecular, Alex, Bianca, Bruna, Carolina, Eloisa, Felipe, Giovana, Letícia, Louise, Márcia, Roberto, Tatiana e Vera pelo convívio diário, pelo apoio e pela amizade. Agradeço também aos colegas que passaram pelo laboratório e já se foram. Aprendi muito com todos vocês.

Aos colegas do Laboratório de Biotecnologia do Hospital A.C. Camargo, Helena, Renato e Eduardo pelas análises de bioinformática.

Aos colegas do Laboratório de de Biologia Computacional do Instituto Ludwig de Pesquisa sobre o Câncer pelas as análises de bioinformática. Em especial, ao Pedro Galante, pela paciência e pelos ensinamentos de bioinformática.

Aos demais colegas do Instituto Ludwig de Pesquisa sobre o Câncer pelo tempo que trabalhamos juntos e pela contínua amizade.

Ao Dr. Hugo Marques Campos e a Dra. Cynthia Osório pela ajuda com as dúvidas de patologia.

Ao Biobanco do Hospital A.C. Camargo pela disponibilização das amostras de RNA.

Ao Banco de Tumores do Hospital A.C. Camargo pela disponibilização das amostras tumorais.

Ao Centro de Pesquisas do Hospital A.C. Camargo pela excelente estrutura para o desenvolvimento da pesquisa.

Aos docentes do curso de pós-graduação do Departamento de Genética pelos ensinamentos. Em especial a Prof. Dr. Regina Célia Mingroni Netto pela supervisão no Programa de Aperfeiçoamento de Ensino.

Aos membros da banca de qualificação Dr. Luiz Eduardo Soares Netto, Gláucia Maria Machado Santalli e Dra, Mariz Vainzof, pelas críticas e sugestões.

Aos funcionários da biblioteca do Hospital A.C. Camargo pelo apoio na obtenção dos artigos.

À Coordenação de Aperfeiçoamento de Pessoal de nível Superior (CAPES) e à Fundação de Amparo à Pesquisa do Estado de São Paulo pelo pelo apoio financeiro durante o período de realização da tese.

Às queridas amigas da faculdade de Biologia por estarem sempre ao meu lado, compartilhando os conhecimentos da Biologia e as experiências de vida.

À grande amiga Elisa Meirelles Reis pela amizade inestimável. Agradeço por ajudar a enfrentar todas as dificuldades e comemorar com tanta alegria todas as minhas conquistas.

A minha irmã, Mariana, por cuidar de mim com um carinho de mãe.

Aos meus pais pelo estímulo em aprender sempre mais, por me darem a base emocional e intelectual que me permitiram chegar até aqui e querer ir mais longe. Vocês são meu maior exemplo.

Ao Juba por torcer tanto por mim, por compreender minha dedicação ao meu trabalho, pela paciência e, principalmente, por me divertir quando mais preciso. Agradeço por todo seu amor.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

O *splicing* alternativo é o processo pelo qual diversos transcritos podem ser gerados a partir de um único gene, sendo de extrema importância para diversidade do repertório transcricional e proteico. Diferentes variantes de *splicing* são expressas entre os diferentes tecidos e estágios de desenvolvimento garantindo o funcionamento normal da célula, portanto, qualquer alteração neste padrão pode resultar no aparecimento de doenças. Neste contexto, o objetivo deste trabalho foi o estabelecimento de metodologias para identificação de variantes de *splicing* em câncer de mama sob influência do oncogene *ERBB2*, o qual é um marcador de mau prognóstico altamente expresso em cerca de 30% dos tumores de mama. Foram estabelecidas duas estratégias para construção de bibliotecas de cDNA. A construção de bibliotecas de cDNA enriquecidas para *splicing* alternativo, baseada na formação e captura de moléculas de heteroduplexes em combinação com a amplificação de RNAm, foi realizada a partir de RNA total de linhagem celular de mama e a partir de um grupo de cinco amostras tumorais, todas com alta expressão de *ERBB2.* Foram identificadas 79 possíveis variantes de *splicing* alternativo em câncer de mama, das quais 18 foram selecionadas para validação por RT-PCR. Foi obtida uma taxa de vallidação de 94% e foram identificadas duas novas variantes de *splicing* alternativo. A regulação da expressão mediada por *ERBB2* de três variantes de *splicing* foi confirmada por duas metodologias distintas, eletroforese em *chip* e estratégia baseada na ligação de sondas específicas, que revelou desbalanço de expressão entre as variantes, demonstrando a influência do oncogene na regulação de variantes de *splicing*. A segunda abordagem utilizada, foi a construção de bibliotecas de cDNA para avaliação do transcriptoma total, utilizando sequênciamento de alto desempenho. Foram utilizados RNA total de duas linhagens celulares de mama que diferem apenas na expressão do gene *ERBB2*. Foram identificadas 2.865 novas variantes de *splicing*, das quais 20, que reportaram a identificação de um novo éxon, foram selecionadas para validação, com uma taxa de validação

de 90%. Seis destas variantes apresentaram aumento de expressão na linhagem com alta expressão de *ERBB2*. Além disso, foi detectado um enriquecimento de algumas categorias de variantes na linhagem celular com alta expressão de *ERBB2*, reforçando a influência do oncogene na regulação do *splicing* alternativo, podendo resultar em variantes de *splicing* associadas a este grupo de câncer de mama, que podem ser candidatas a marcadores moleculares.

Summary

Alternative splicing is a process, by which many differente transcripts can be generated by one single gene, significantly expanding the transcriptional and proteomic diversity. Different splicing variants are generated among different transcripts and developmental stages, assuring normal cell function. Therefore, alterations in the splicing process can lead to diseases outcome. In this context, the aim of this study was the establishment of methodologies for the identification of alternative splicing in breast cancer influenced by *ERBB2* oncogene, which is a poor prognostic molecular marker, highly expressed in 30% of human breast cancer. Two strategies were established for the construction of cDNA libraries. Alternative enriched splicing libraries, based on heteroduplex capture combined with mRNA amplification, were constructed from total RNA from a cell line and also from five tumor samples, all of them presenting high *ERBB2* expression. Seventy nine putative splicing variants were identified and 18 of them were selected for RT-PCR validation. A high validation level was obtained (94%) and two novel alternative splicing variants were identified. *ERBB2* mediated regulation was confirmed for three variants by two distinct methodologies, electrophoresis on a chip and probe specific ligation approach. The alteration in the expression balance of variants suggests the influence of the oncogene in the splicing pattern regulation. The second strategy was the construction of cDNA libraries for global transcriptome analysis based on deep sequencing. Total RNA from two mammary epithelial cell lines expressing different ERBB2 levels were used and 2,865 novel splicing variants were identified. Twenty novel events reporting the inclusion of novel exons were selected for RT-PCR validation with 90% validation rate. Six variants presented higher expression in the cell line with high levels of ERBB2. Moreover enrichment in *splicing* events was detected in the ERBB2 high expressing cell line, supporting the ERBB2 influence in alternative splicing regulation, possibly resulting in splicing variants associated to this subgroup of cancer that can be tested as molecular markers.

1. Introdução

A partir da obtenção da sequência completa dos 3,2 bilhões de bases do genoma humano foi constatado que apenas 3% do genoma correspondem a genes codificadores de proteínas, aproximadamente 30 mil genes, sendo esses interrompidos por longos trechos de sequências intrônicas (LANDER et al., 2001; VENTER et al., 2001). Esse número foi surpreendentemente menor do que as estimativas feitas no inicío no início do projeto genoma que giravam em torno de 100 mil genes. O número de genes humanos identificados também pode ser considerado pequeno quando comparado com organismos mais simples, os quais muitas vezes possuem um número de genes muito similar a *Homo sapiens*, como o verme *Caenoharbditis elegans* que possui em torno de 22 mil genes (*C. ELEGANS* SEQUENCING CONSORTIUM, 1998).

Portanto, a complexidade dos organismos está mais relacionada com a existência de mecanismos que expandem a capacidade codificadora do genoma, como sítios de início transcrição alternativos (CARNINCI et al., 2006), poliadenilação alternativa (XING; LI, 2009) e o *splicing* alternativo (MODREK, et al., 2001). O *splicing* alternativo foi inicialmente proposto por Gilbert em 1978, (GILBERT, 1978) logo após a identificação da existência dos íntrons e parecia ser um fenômeno raro, com estimativas de ocorrência em cerca de 5% dos genes de eucariotos (SHARP, 1994). Atualmente sabemos que esse é um fenômeno frequente que contribui para a obtenção de, em média, seis transcritos diferentes por gene (HARROW, 2006; KIM; MAGEN; AST, 2007).

O *splicing* alternativo é um importante mecanismo de modulação da função gênica, uma vez que altera a população de transcritos em diferentes tipos celulares, estágios de desenvolvimento, condições ambientais e estados patológicos. As diferentes isoformas proteicas geradas podem alterar a afinidade de ligação, a atividade enzimática, a regulação alostérica ou localização celular das proteínas, interferindo em inúmeras atividades celulares.

1.1. O processamento do RNA mensageiro

As moléculas de RNA mensageiro (RNAm) são transcritas como moléculas precursoras que precisam ser processadas antes de serem transportada ao citoplasma para a tradução das proteínas. O processamento do RNAm ocorre inicialmente pela adição de uma molécula de 7-metilguanosina a extremidade 5' dos transcritos, chamado de 5' CAP, que ocorre concomitantemente a transcrição. Em seguida, logo após o término da transcrição, ocorre a incorporação de uma longa cadeia de adeninas a extremidade 3' do transcrito, chamada de cauda poli A. A presença do 5' CAP e da cauda poli A são extremamente importantes para permitir o transporte dos transcritos do núcleo ao citoplasma, com também para protegê-los da degradação.Para completar a maturação dos RNAm é necessário a remoção de algumas regiões específicas do pré-RNAm, pelo processo de *splicing*[1].

A maioria dos genes dos organismos eucariotos multicelulares possui uma estrutura formada por éxons, regiões funcionais que contêm a sequência de códons para a síntese proteica, intercalados por regiões não codificantes, os íntrons, que necessitam ser precisamente retirados para originar os RNA mensageiros (RNAm) maduros, pelo processo conhecido como *splicing.* Em genes humanos, os íntrons são, em média, vinte e cinco vezes mais extensos que os éxons (120 bases em média), correspondendo a mais de 90% da porção genômica dos genes (AST, 2004). No entanto, a maquinaria celular responsável pelo mecanismo de *splicing*, o spliceossomo, é capaz de reconhecer com precisão as sequências correspondentes aos éxons e íntrons para promover o processamento do RNAm e, consequentemente, formar a molécula de RNA mensageiro madura.

O spliceossomo é um grande complexo celular formado por cinco subunidades ribonucleoproteicas, chamadas snRNPs (do inglês - *small nuclear ribonucleoprotein particles* - pequenas partículas nucleares ribonucleoproteicas). Cada subunidade é composta por uma pequena molécula de RNA nuclear (snRNAs, do inglês - *small nuclear* RNA) rica em

[1]A tradução da palavra *splice* pelo dicionário é: *substantivo*. emenda, junção. *verbo*. ligar, unir, emendar. FONTE: "Novo dicionário FOLHA Webster's: inglês/português português/inglês". Em relação ao fenômeno celular não existe uma tradução utilizada, sendo utilizada a palavra em inglês.

uracila associada a um conjunto de 6 a 10 proteínas. O reconhecimento dos íntrons pela maquinaria de *splicing* ocorre pelo pareamento de bases entre os RNAs constituintes do spliceossomo e as regiões específicas do pré-RNAm, tais como: o sítio doador de *splice* localizado na extremidade 5' dos íntrons representado pelos dinucleotídeos conservados GU; o sítio de aceptor de *splice* localizado na extremidade 3' dos íntrons representado pelos dinucleotídeos conservados AG; o trato de polipirimidina que é uma região rica em bases timina e citosina localizado a cerca de 15 nucleotídeos a montante do sítio aceptor de *splice*; e, por fim, o ponto de quebra ou ponto de ramificação que é normalmente uma base adenina localizado entre 30 a 50 nucleotídeos a montante sítio aceptor de *splice* (LODISH et al., 2001) (Figura 1). Uma pequena fração dos íntrons apresenta sítios de *splice* com os dinucleotídeos AT na posição 5' do íntron e os dinucleotídeos AC na posição 3' do íntron (TARN; STEITZ, 1996). Com exceção dos sítios de *splice* as demais sequências reguladoras não apresentam um nível de conservação tão alto, o que dificulta a identificação dos limites entre éxons e íntrons a partir da sequência genômica.

Figura 1: Estrutura genômica de um gene humano hipotético. Os retângulos representam os éxons e a linha representa a região intrônica, onde estão sinalizadas as sequências reguladoras como o sítio doador de *splice* (GT), o sítio aceptor de *splice* (AG), o trato de polipirimidina (Py) e o ponto de quebra (A). Adaptado de Ferreira e colaboradores (2007).

A montagem do spliceossomo ocorre de forma sequencial como mostrado na figura 2. Inicialmente, ocorre o pareamento da subunidade snRNP U1 com sítio de doador *splice*, formando o complexo E (do inglês *early* – inicial); a etapa seguinte é dependente de ATP e consiste no pareamento de bases entre snRNP U2 e o ponto de quebra do pré-RNAm, sendo conhecido como complexo A ou pré-spliceossomo. Nesta etapa ocorre uma alteração conformacional na molécula de RNA mensageiro, que aproxima o ponto de quebra ao sítio doador de *splice*. Posteriormente, ocorre a ligação do trio de snRNP U4/U5/U6 ao RNAm no ponto de quebra,

formando o spliceossomo propriamente dito ou complexo B. Ocorre, então, uma segunda mudança conformacional na qual ocorre o desligamento do snRNP U1 e U4 e a primeira reação de transesterificação, gerando o complexo C. A reação de transesterificação é a transferência da ligação fosfodiéster que ocorre entre o fosfato do sítio doador e uma hidroxila do ponto de quebra, unindo o sítio doador ao ponto de quebra e formando uma estrutura em laço (LODISH, 2001). Por fim, ocorre a segunda reação de transesterificação entre o fosfato do primeiro nucleotídeo do éxon a jusante com a hidroxila do último nucleotídeo do éxon a montante, resultando na junção dos éxons adjacentes. O íntron é rapidamente degradado e o complexo spliceossomo é desfeito (HARTMUTH et al., 2002).

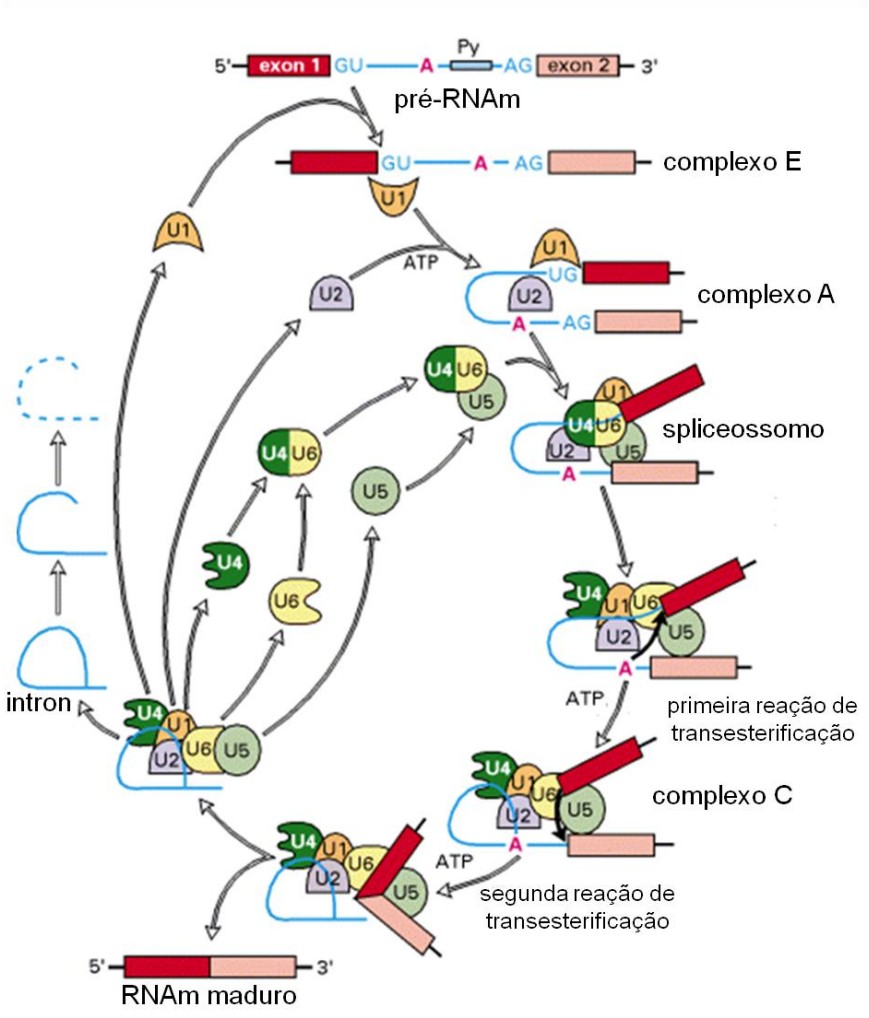


Figura 2: Montagem do *splice*ossomo e as etapas do *splicing*. GU – sítio doador de *splice*. AG – sítio aceptor de *splice*. Py – trato de polipirimidina. A – ponto de quebra. U1, U2, U4, U5 e U6 – subunidades ribonucleoproteicas (snRNP) envolvidas com o *splicing*. Adaptado de Lodish e colaboradores (2001).

O fato de a maioria dos genes humanos ser constituída por dois ou mais éxons aumenta a plasticidade do genoma, pois permite que diferentes combinações entre os éxons sejam utilizadas na formação do RNAm maduro, gerando transcritos distintos a partir de uma única molécula de RNAm imaturo. Esse processo, denominado *splicing* alternativo, aumenta significativamente a diversidade transcricional dos organismos.

Estimativas recentes sugerem que cerca de 95% dos genes humanos, com múltiplos éxons, sofrem *splicing* alternativo (PAN, 2008), gerando, em média, seis transcritos por gene (HARROW, 2006). Além disso, uma vez que cerca de 80% dos eventos ocorrem dentro da região codificante do gene (MODREK et al., 2001), o aumento do repertório transcricional é acompanhado por um aumento na diversidade proteômica, podendo gerar isoformas com alteração na função, na localização celular, na atividade enzimática e na afinidade pelo substrato.

Sob uma perspectiva evolutiva, o *splicing* alternativo parece contribuir de forma significativa para o aumento da complexidade fenotípica dos organismos eucariotos multicelulares. Em protozoários, os íntrons e, consequentemente, os genes com múltiplos éxons são extremamente raros; em fungos e leveduras estão presentes em apenas 4% dos genes, são pequenos e não apresentam eventos de *splicing* alternativo. Já nos metazoários, os íntrons são mais prevalentes e são detectados eventos de *splicing* alternativo. Portanto, o *splicing* alternativo explica, em parte, a maior complexidade de organismos superiores frente ao respectivo número de genes, proporcionando grande expansão da capacidade codificadora dos genomas (AST, 2004).

Os padrões de *splicing* alternativo podem ser classificados em quatro tipos (Figura 3). O tipo mais estudado é o uso alternativo do éxon (Figura 3A), no qual os éxons chamados éxons cassetes (ou regulados), podem estar incluídos ou excluídos da molécula de RNA mensageiro maduro (Figura 3B). O uso de sítios de *splice* alternativos, tanto sítio doador (Figura 3C) como sítio aceptor de *splice* (Figura 3D), altera o tamanho dos éxons. Por fim, como o próprio nome sugere, a retenção de íntrons gera variantes de *splicing* nas

quais um dos íntrons não é removido, sendo incorporado à molécula de RNAm madura (Figura 3E). Podem ocorrer ainda casos de éxons mutuamente exclusivos, nos quais apenas um ou o outro éxon é incluído na molécula de RNAm madura, isto é, não ocorre a presença dos dois éxons em um mesmo transcrito (STEPHAN et al., 2007).

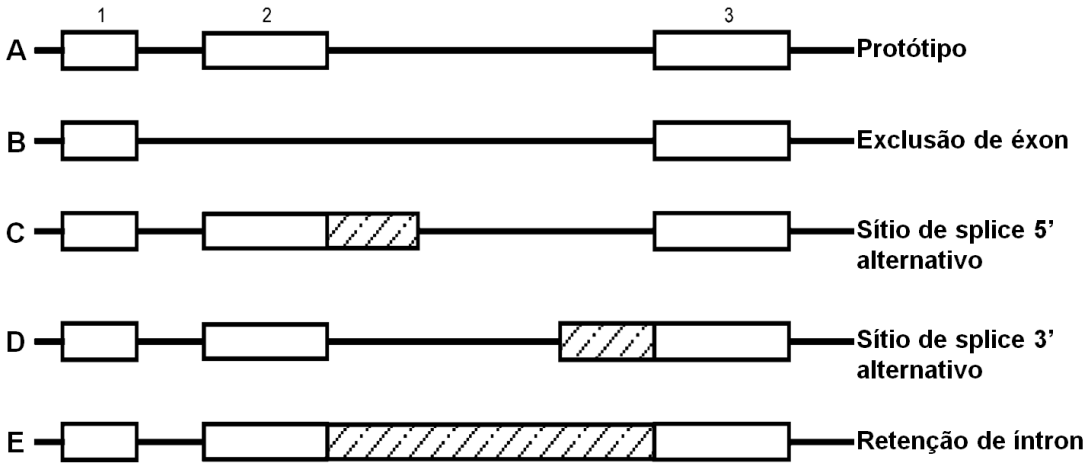


Figura 3: Padrões de *splicing* alternativo. No esquema estão representados três éxons de um gene hipotético, numerados 1, 2 e 3. A – Protótipo: transcrito modelo de referência. B – Uso alternativo de éxon: o éxon 2, éxon cassete pode ou não ser incluído no RNAm maduro. C – Uso de sítio doador de *splice* alternativo (ou sítio de *splice* 5'). D – Uso de sítio aceptor de *splice* alternativo (ou sítio de *splice* 3'). E – Retenção de íntron. Modificado de Ferreira e colaboradores (2007).

Na figura 3 podemos notar que o éxon 1 do gene hipotético, diferentemente dos éxons 2 e 3, está presente em todos os transcritos de forma inalterada. Éxons que não sofrem alterações devido ao *splicing* alternativo são denominados éxons constitutivos, pois estão sempre presentes nas moléculas de RNA mensageiro maduro, enquanto que os éxons regulados por *splicing* alternativo são chamados éxons alternativos.

1.2. Regulação do *splicing*: os elementos *cis* e *trans*

O *splicing* alternativo é um importante mecanismo de modulação da função gênica, uma vez que altera a população de transcritos em diferentes tecidos e estágios de desenvolvimento, interferindo em inúmeras atividades celulares. A regulação do *splicing* alternativo é um mecanismo extremamente complexo que depende tanto dos elementos *cis* (sequência de bases nos éxons e íntrons) quanto dos fatores *trans* (proteínas reguladoras). A regulação do padrão de *splicing* tem implicações em diversos processos cruciais como a determinação sexual em *Drosophila melanogaster* (BLACK, 2003), apoptose (BOISE et al., 1993), audição (FETTIPLACE; FUCHS, 1999), entre outros. Assim, a geração dos transcritos variantes entre diferentes indivíduos, tecidos e células depende tanto da sequência de bases nos éxons e íntrons quanto da disponibilidade de proteínas reguladoras nas células, que controlam rigorosamente a geração das variantes, garantindo o funcionamento normal das células. Os elementos *cis* são sequências curtas de DNA presentes tanto nos éxons como nos íntrons que podem agir facilitando o processo de reconhecimento de sítios de *splice* ou inibindo a utilização de determinados sítios (LAM, 2002). Os elementos *cis* são nomeados de acordo com sua localização e ação: *exonic splicing enhancers* são ativadores de *splicing* localizados nos éxons, *exonic splicing silencers* são inibidores de *splicing* localizados nos éxons, *intronic splicing enhancers* são ativadores de *splicing* localizados nos íntrons e *intronic splicing silencers* são inibidores de *splicing* localizados nos íntrons (Black, 2003) (Figura 4).

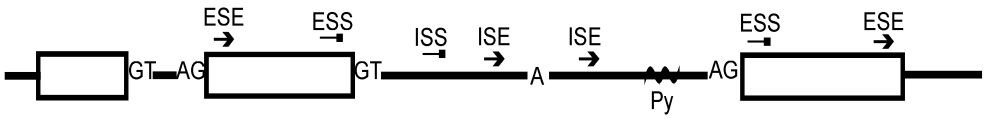


Figura 4: Os elementos *cis* de regulação do *splicing*. A regulação do *splicing* pode ser controlada por elementos ativadores ou *enhancers* (➔) de localização exônica (ESE) ou intrônica (ISE) ou elementos inibidores ou *silencers* (⊣) de localização exônica (ESS) ou intrônica (ISS). Os retângulos representam os éxons e a linha representa os íntrons. GT – sítio doador de *splice*. AG – sítio aceptor de *splice*. Py – trato de polipirimidina. A – ponto de quebra. Adaptado de Ferreira e colaboradores (2007).

A sequência de bases, o número e a localização dos elementos *cis* é variável entre os diferentes genes. Em contrapartida, os sítios doador e aceptor de *splice* estão sempre presentes em regiões muito bem definidas. No entanto, alterações na composição de bases que flanqueiam esses sítios podem também contribuir para regulação do *splicing*, interferindo com o balanço entre o *splicing* constitutivo e alternativo por influenciar a estabilidade do pareamento de bases entre o pré-RNAm e os snRNAs do spliceossomo (AST, 2004). Sítios de *splice* considerados fortes, possuem uma determinada sequência de bases nas posições adjacentes o que acarreta ligação estável com os snRNAs, sendo comumentemente reconhecido pelo spliceossomo (Figura 5A). Alterações em uma única base, por exemplo, a troca de uma adenina por uma guanina a jusante ao sítio doador de *splice*, é suficiente para diminuir a força de interação entre o íntron e o snRNA U1, aumentando a frequência de ocorrência do *splicing* alternativo (Figura 5B). Demais alterações na sequência de bases podem acarretar ausência de reconhecimento do sítio pela maquinaria de *splicing* (Figura 5C).

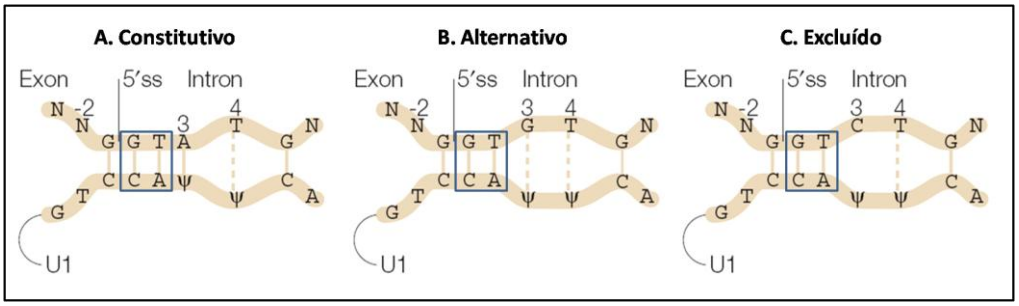


Figura 5: Pareamento das bases adjacentes ao sítio doador de *splice* no íntron e o snRNP U1. A – pareamento estável entre o íntron e o snRNP U1 acarreta sítio de *splice* constitutivo. B – alteração de uma base na posição 3 diminui a estabilidade da interação entre os RNAs resultando em um aumento da taxa de *splicing* alternativo em relação ao constitutivo. C – alteração em bases nas posições 3 e 4 interferem com a ligação entre snRNP U1, aumentando a taxa de exclusão/inclusão do éxon. O quadrado azul delimita os dinucleotídeos constituintes do sítio doador de *splice* (5'ss). Modificado de Ast (2004).

Os elementos de ação *cis* (*enhancers* e *silencers*) agem como sítios de ligação para diversas proteínas reguladoras, que são os fatores *trans* ou fatores de *splicing*. De modo geral, os *enhancers* são sítios de ligação para

proteínas da família SR (proteínas ricas em serina e argenina), as quais são fatores de *splicing* essenciais, que recrutam direta ou indiretamente componentes do spliceossomo (GRAVELEY, 2000). Por outro lado, as proteínas hnRNP são as principais ligantes das sequências silenciadoras (*silencers*) e podem interferir direta ou indiretamente na montagem dos componentes do spliceossomo, bloqueando interações durante a definição dos éxons ou bloqueando a ação das proteínas SR (BLACK, 2003; CAPUTI et al., 1999; DEL GATTO-KONCZAK et al., 1999). Além dessas mais de 300 proteínas, conhecidas como fatores de *splicing*, parecem estar envolvidas na regulação do processo de *splicing* alternativo (NILSEN, 2003; ZHOU et al., 2002), sendo responsáveis pelo reconhecimento e pela determinação dos sítios de *splice* a serem utilizados, resultando na geração das mais diversas variantes (KRAMER, 1996).

Assim, fica evidente que a regulação do *splicing* alternativo é um mecanismo extremamente complexo, dependente de uma combinação entre a sequência de bases dos genes e do nível de expressão das proteínas reguladoras. A presença de sítios fortes ou fracos determina o maior ou menor reconhecimento pelo spliceossomo, no entanto a afinidade dessa ligação pode ser alterada na presença das proteínas reguladoras associadas aos sítios ativadores ou silenciadores, alterando o balanço de expressão das variantes de *splicing* (Figura 6).

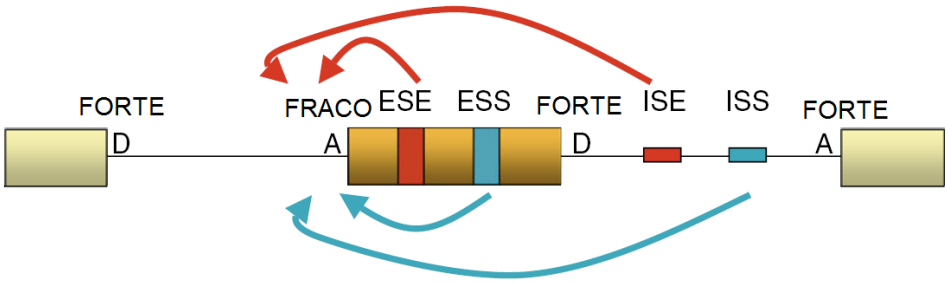


Figura 6: A regulação do *splicing* ocorre pela interação entre os fatores *trans* e os elementos em *cis*. D – sítio doador de *splice*. A – sítio aceptor de *splice*. ESE – *exonic splicing enhancer*. ESS – *exonic splicing silencer*. ISE – *intronic splicing enhancer*. ISS – *intronic splicing silencer*. As flechas indicam a influencia da ligação de fatores *trans* no elementos *cis* favorecendo ou inibindo o reconhecimento de um sítio fraco de *splice* pela maquinaria celular. Adaptado de Srebrow e Kornblihtt (2006).

Outros fenômenos transcricionais que interferem no padrão de *splicing* foram descritos mais recentemente. A diminuição na velocidade de incorporação de nucleotídeos durante a transcrição pela enzima polimerase II pode acarretar o aumento da taxa de inclusão de éxons alternativos (KORNBLIHTT, 2006). Além disso, a regulação do *splicing* alternativo pode ser um mecanismo de regulação de expressão transcricional. Um terço das variantes de *splicing* alteram a sequência aberta de leitura da proteína de forma a inserir um códon de parada prematuro, que sinaliza degradação do transcrito pela via *non-sense mediated decay* (NMD) (LEWIS et al., 2003; SCHELL; KULOZIK; HENTZE, 2002).

A busca pela obtenção de um padrão global de *splicing* que possa prever o padrão de variantes geradas para um determinado gene em um determinado tecido ou célula tem sido o objetivo de diversos grupos (BARASH et al., 2010; MATLIN; CLARK; SMITH, 2005; FU, 2004). A definição de um código de *splicing* seria de extrema importância para entender os mecanismos de regulação tecido-específica e como esses mecanismos são alterados nas doenças humanas. Dois exemplos muito bem estudados de regulação de *splicing* são a determinação de sexo em *Drosophila melanogaster* e o perfil de variantes do gene humano *CD44* que serão discutidos a seguir como forma ilustrativa da complexidade e importância dessa regulação.

1.2.1. Regulação do *splicing* alternativo e desenvolvimento embrionário: o exemplo da determinação de sexo em *Drosophila melanogaster*

Um exemplo muito bem estudado dos mecanismos de regulação de *splicing* ao longo do desenvolvimento é a determinação sexual da mosca *Drosophila melanogaster*. O gene mais importante desse sistema, chamado *sex-lethal* gene (*Sxl*) é transcrito apenas nas moscas fêmeas dessa espécie, devido a diferença na dosagem do cromossomo X (ERICKSON; QUINTERO,

2007; SALZ; ERICKSON, 2010). A proteína SXL contém dois domínios conservados do tipo RRM, sendo um importante fator de *splicing* para diversos genes (SALZ; ERICKSON, 2010). A presença da proteína SXL no início do desenvolvimento nos embriões fêmea, promove autoregulação, garantindo geração de transcritos funcionais do gene *Sxl* pela exclusão do éxon 3. Em machos, a ausência da proteína SXL inicial resulta na formação de transcritos do gene *Sxl* com a inclusão do éxon 3, codificando um códon de parada prematuro, sem a formação da proteína funcional (SALZ;ERICKSON, 2010). A proteína SXL interfere com o padrão de *splicing* de um segundo gene, o *Transformer* (*Tra*). Apenas na presença de SXL ocorre a geração de uma variante de *splicing* funcional de *Tra*. Nos indivíduos do sexo masculino, a ausência da proteína SXL promove a geração de uma variante distinta do gene *Tra*, a qual contém um códon de parada prematuro, resultando em uma proteína truncada não funcional (AMREIN; GORMAN; NÖTHIGER, 1988). A proteína Transformer, por sua vez, também age como reguladora de *splicing* dos genes *doublesex* (*dsx*) e *fruitless* (*fru*), gerando variantes de *splicing* distintas nos machos e nas fêmeas, que resultam no dimorfismo sexual de *Drosophila melanogaster* (HOSHIJIMA et al., 1991). Assim, variantes de *splicing* específicas para indivíduos machos e fêmeas serão produzidas, resultando em alterações morfológicas e comportamentais (SIWICKI; KRAVITZ, 2009). O gene *doublesex* está mais associado à determinação das caracteristicas morfológicas enquanto o gene *fruitless* parece agir principalmente na determinação das características comportamentais. Os genes *fru* e *dsx* agem como fatores de transcrição, interferindo na expressão de diversos genes.

1.2.2. Regulação do *splicing* alternativo e a expressão tecido-específica: as variantes do gene *CD44*

O gene *CD44* codifica uma glicoproteína envolvida principalmente com a adesão célula-célula e célula-matriz, além de outros processos celulares como ativação de linfócitos, angiogênese e liberação de citocinas, entre

outros (SNEATH; MANGHAM, 1998). Esse gene possui 5.725 pares de bases distribuídos em 19 éxons, dos quais os 5 primeiros e os 5 últimos são conservados enquanto os 9 éxons intermediários são alternativos. Os cinco primeiros éxons conservados codificam um domínio de ligação ao ácido hialurônico, juntamente com os éxons conservados 6 e 7. O éxon conservado 8 codifica um domínio transmembrânico, enquanto os éxons conservados 9 e 10 codificam um domínio intracelular que interage com o citoesqueleto. Por outo lado, os éxons alternativos codificam diferentes domínios extracelulares que irão promover interação com diferentes proteínas, alterando a função do gene (Figura 7) (SNEATH; MANGHAM, 1998).

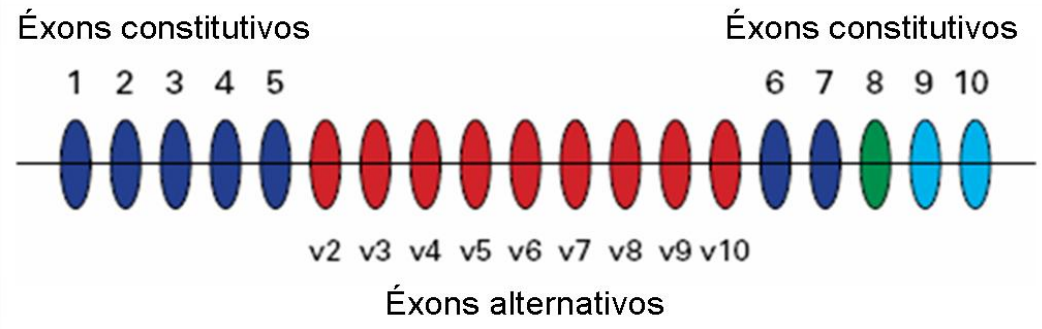


Figura 7: Estrutura do gene *CD44*. Os éxons estão representados por círculos ovalados, sendo os éxons conservados numerados de 1 a 10 e os éxos alternativos numerados de v2 a v10. Os éxons conservados coloridos em azul marinho juntamente com os éxons alternativos compõem a região extracelular da proteína. O éxon em verde (número 8) codifica o domínio transmembrânico e os éxons em azul claro (numerados 9 e 10) representam a porção intracelular da proteína. Adaptado de Sneath e Mangham (1998).

A isoforma padrão desse gene (CD44s) é formada apenas pelos éxons constitutivos e é a única variante encontrada em linfócitos. Em outros tecidos como gengiva, laringe, língua e esôfago as diversas variantes estão presentes em alto nível de expressão. Em estômago e intestino, também se observa a produção das diversas variantes, no entanto existe um padrão de expressão diferencial entre elas, onde a isoforma padrão é a mais abundante, seguida das isoformas que contenham uma combinação dos éxons v7, v8 e v9, e por fim, as variantes que contêm apenas um dos éxons alternativos (v6 ou v7 ou v8) são as menos frequentes (Tabela 1) (SNEATH; MANGHAM, 1998). Portanto, é possível inferir que não apenas alterações entre presença

e ausência de uma variante, mas também o balanço de expressão entre elas é de extrema importância para a manutenção fisiológica das células.

Tabela 1: Perfil de Expressão das variantes de *splicing* do gene *CD44* em diferentes tecidos humanos saudáveis.

Tecido	*CD44s*	*CD44v6*	*CD44v7*	*CD44v8*	*CD44v7 e v8*	*CD44v8 e v9*	*CD44v7, v8 e v9*	*Intron 9*
Pele	+++	+++	+++	+++	+++	+++	+++	++
Gengiva	+++	+++	+++	+++	+++	+++	+++	+++
Língua	+++	+++	+++	+++	+++	+++	+++	+++
Laringe	+++	+++	+++	+++	+++	+++	+++	+++
Esôfago	+++	+++	+++	+++	+++	+++	+++	+++
Brônquio	++	++	++	++	++	++	++	+
Pulmão	+++	++	++	++	++	++	++	+
Estômago	+++	+	+	+	++	++	++	-
Duodeno	+++	+	+	+	++	++	++	-
Jejuíno	+++	+	+	+	++	++	++	-
Íleo	+++	+	+	+	++	++	++	-
Cólon	+++	+/++	+/++	+/++	++	++	++	-/+
Reto	+++	+/++	+/++	+/++	++	++	++	-/+
Glândula paratireóide	+++	++	++	++	+++	+++	+++	+
Glândula submandibular	+++	++	++	++	+++	+++	+++	+
Glândula tireóide	+	+	+	+	++	+++	++	+
Pâncreas	++	++	++	++	+++	+++	+++	+
Glândula adrenal	-	-	-	-	+	+	+	-
Ovário	-	-	-	-	+	+	+	-
Miométrio	+++	-	-	-	+	+	+	-
Útero	+++	+++	+++	+++	+++	+++	+++	+++
Fígado	-	-	-	-	-	-	-	-
Ducto bilear	-	-	-	-	-	-	-	-
Córtex renal	+	-	+	+	+	+	+	-
Ureter	+	-	-	-	-	-	-	-
Bexiga	+++	++	++	+++	+++	+++	+++	++
Baço	++	++	++	++	++	++	++	-
Linfócitos	+++	-	-	-	-	-	-	-

Adaptado de Sneath e Mangham (1998).

1.3. Identificação de variantes de *splicing*: busca por novas variantes e variantes associadas ao câncer

Devido à alta complexidade transcricional decorrente do *splicing* alternativo e de sua implicação em diversas doenças humanas, diferentes estratégias foram propostas para explorar o repertório de variantes transcricionais presente nas diferentes células, tecidos e estágios de desenvolvimento. As abordagens experimentais desenvolvidas para análise do padrão de *splicing* em larga escala, podem ser divididas em: métodos baseados em RT-PCR (*reverse transcriptase polimerase chain reaction*), plataformas de microarranjos de DNA e metodologias baseadas em sequência.

Independentemente da abordagem escolhida, estratégias de avaliação do transcriptoma em larga escala são em geral altamente dependentes de ferramentas de bioinformática. Em relação à identificação de variantes de *splicing*, a bioinformática tem papel fundamental. Apesar da identificação *in silico* de variantes de *splicing* unicamente a partir da sequência genômica (identificação *ab initio*) ser altamente complexa devido à baixa conservação das sequências sinalizadoras dos éxons e íntrons, o processamento da grande quantidade de dados experimentais gerados tem sido possibilitado pelo contínuo desenvolvimento de ferramentas e estratégias computacionais (FERREIRA et al, 2007). Além disso, apesar das predições computacionais necessitarem de confirmação experimental, elas podem identificar possíveis novas variantes, bem como sugerir a presença de elementos reguladores (CARTEGNI et al., 2003; HSU et al., 2005; SCHWARTZ; HALL; AST, 2009; SMITH et al., 2006).

1.3.1. Utilização de RT-PCR na identificação de variantes de *splicing*

A abordagem de RT-PCR é baseada na propriedade da enzima transcriptase reversa de sintetizar DNA complementar (cDNA) a partir de moléculas de RNA. Assim, os diferentes transcritos gerados por uma célula em um determinado momento são convertidos em cDNA e detectados após amplificação em cadeia da enzima polimerase. Uma etapa importante nessa abordagem consiste no desenho de oligonucleotídeos utilizados como iniciadores da reação de amplificação das variantes de interesse. Assim, o perfil de expressão das variantes é analisado por RT-PCR utilizando cDNA de diversos tecidos de interesse, como tecido do tumor e tecido não neoplásico correspondente. Os resultados permitem identificar variantes específicas ou associadas aos tumores. Essa abordagem é laboriosa, sendo mais apropriada para estudos em pequena escala quando há um interesse muito grande na obtenção do perfil de expressão das variantes de um ou poucos genes.

No entanto, com o desenvolvimento de robôs que auxiliam o preparo das reações em placas e da possibilidade da eletroforese capilar, também em placa, diminuindo o intenso trabalho manual, análises baseadas em RT-PCR, em larga escala, tornam-se viáveis. A análise do produto de RT-PCR por eletroforese capilar permite avaliar isoladamente o tamanho e a quantidade de cada um dos diferentes produtos gerados em uma mesma reação. Com isso, é possível determinar o nível de expressão relativo entre as variantes de um mesmo gene e avaliar alterações no balanço de expressão entre elas. Um estudo recente utilizando essa estratégia avaliou o perfil de expressão de variantes de *splicing* de 600 genes humanos relacionados com câncer em 21 amostras de tecido não neoplásico e 26 amostras de câncer de mama. Esse estudo foi capaz de detectar um perfil de expressão de variantes de *splicing* de 41 genes capazes de separar as amostras tumorais das amostras normais (VENABLES, 2008a).

Apesar de resultados promissores, a avaliação da expressão de variantes de *splicing* por RT-PCR pode gerar resultados artefatuais em

relação ao balanço de expressão das variantes devido a uma competição dos transcritos na reação. Nesse caso, a amplificação de uma das variantes pode ser favorecida em relação à amplificação das demais variantes, devido a características como tamanho e conteúdo GC, entre outros, gerando um falso valor de diferença de expressão. Entretanto, algumas estratégias foram propostas com o intuito de reduzir a super representatividade da variante mais abundante em relação às demais, permitindo identificar variantes de *splicing* novas ou raras utilizando a metodologia de RT-PCR (VENABLES; BURN, 2006; VENABLES, 2008b; GORLOV; SAUDERS, 2002).

1.3.2. Microarranjos de DNA

A tecnologia de microarranjos de DNA é extremamente utilizada para análises de expressão gênica, pois permite avaliar o perfil de expressão de uma grande quantidade de transcritos simultaneamente, sendo, portanto, um instrumento valioso para análise de expressão de variantes de *splicing*. No entanto, para melhor avaliação do perfil de expressão de variantes de *splicing* algumas plataformas específicas são mais apropriadas, como plataformas que cobrem grandes regiões genômicas (*tiling microarrays*) (FAN et al., 2006; HU et al., 2001), plataformas de éxons (GARDINA et al., 2006) ou ainda plataformas que representem não apenas os éxons, mas também as junções éxon-éxon conhecidas (JOHNSON et al., 2003; RELÓGIO et al., 2005). Estas plataformas permitem uma análise detalhada do padrão de expressão dos diferentes éxons de um gene.

Plataformas de grandes regiões genômicas permitem a identificação de novas variantes de *splicing*, bem como o perfil de expressão dessas. No entanto, são de extrema complexidade em termos de análises bioinformáticas para a definição das porções exônicas e intrônicas, sendo mais sensíveis na detecção de eventos de uso alternativo do éxon ou retenção de íntrons. Essa abordagem busca identificar grupos de sondas localizadas em regiões genômicas próximas que apresentam expressão similar, porém diferente da

média de expressão de todas as sondas correspondentes ao gene (HU et al., 2001).

As plataformas de microarranjos de DNA que contêm sequências exônicas necessitam de um conhecimento prévio sobre os éxons de interesse. Um estudo desenvolvido pelo nosso grupo selecionou sequências candidatas a éxons mais expressos em tecidos tumorais, identificados por análises bioinformáticas (KIRSCHBAUM-SLAGER et al., 2005), e imobilizou-as em uma plataforma de microarranjos de DNA. Com o intuito de obter identificar variantes de *splicing* mais expressas em tumores de mama, estas variantes foram interrogadas por amostras de tumor e normais de mama. No total, foram confirmados três genes, *MK-STYX*, *BRRN1* e *TRIM37*, cujas variantes de *splicing* apresentaram nível de expressão elevado em amostras de tumores em relação a amostras normais. Adicionalmente, a expressão da variante de *splicing* que contêm o éxon adicional do gene *TRIM37* apresentou associação positiva com a presença de expressão dos receptores hormonais de estrógeno e progesterona, bem como com ausência de mutação no gene *p53*, avaliado por imunohistoquímica. Estas associações não foram observadas quando consideramos o nível de expressão das variantes sem esse éxon (RANGEL, 2008), corroborando com a sugestão de modulação específica da expressão individual das variantes de *splicing*.

As plataformas de éxons e junções éxon-éxon são baseadas em um cuidadoso desenho de sondas especificas que correspondem a regiões exônicas, intrônicas e a junções éxon-éxon, como mostrado na figura 8. Para comparação do padrão de expressão das variantes entre duas amostras, o cDNA oriundo de tecidos diferentes é marcado com moléculas fluorescentes distintas e hibridadas em uma mesma lâmina contendo as sondas correspondentes aos éxons, íntrons e às junções éxon-éxon. No exemplo da figura 8, está esquematizado o desenho de sondas para detecção da expressão de duas variantes de um gene, que diferem no uso alternativo do éxon 2.

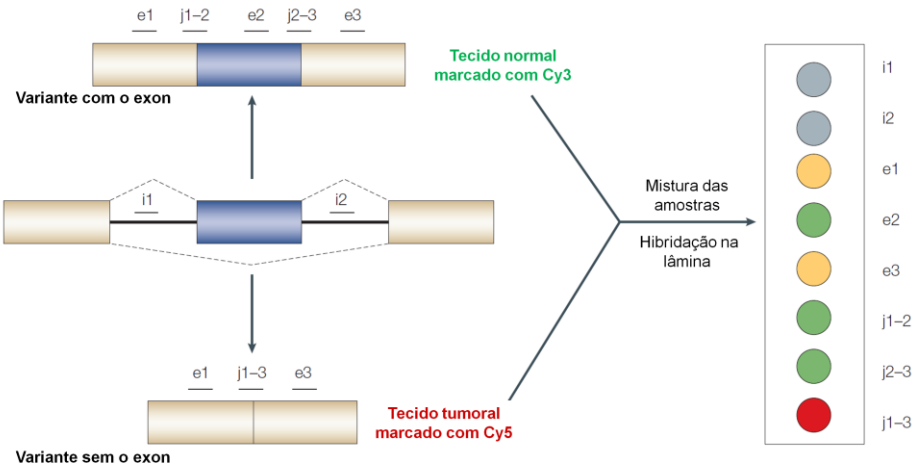


Figura 8: Desenho de sondas para análise de *splicing* por microarranjos de DNA. A figura mostra um gene hipotético que apresenta duas variantes de *splicing* com uso alternativo do éxon 2. Foram desenhadas sondas nas junções éxon-éxon entre o éxon 1 e 2 (j1-2), entre os éxons 2 e 3 (j2-3) e também na junção entre os éxons 1 e 3 (j1-3) que são distintas entre as variantes. Foram também utilizadas sondas éxon-específicas desenhadas nos éxons 1, 2 e 3 (e1, e2 e e3, respectivamente). Como controle de contaminação de RNAm não processado foram desenhadas sondas nos íntrons 1 e 2 (i1 e i2). O RNAm dos diferentes tecidos foram marcados com moléculas fluorescentes de comprimento de onda distintos, Cy3 e Cy5, misturados e hibridados na mesma lâmina. À direita da figura, o resultado da co-hibridação mostra um perfil de expressão tecido-específico entre as variantes. Adaptada de Matlin, Clark e Smith (2005).

A cor azul mostra ausência de detecção de sinal nas regiões correspondentes às sondas intrônicas nas duas amostras de cDNA. Os éxons 1 e 3 presentes nas duas variantes apresentaram nível equivalente de expressão dos respectivos cDNAs provenientes dos 2 tecidos, representado pela cor amarela, no resultado da hibridação. No entanto, o éxon 2 e as junções éxon1/éxon2 e éxon2/éxon3 apresentam maior nível de expressão no tecido normal, indicado pela cor verde no resultado da hibridação. Por outro lado, a junção éxon1/éxon3 apresentou maior expressão no tecido tumoral, representado pela cor vermelha no resultado da hibridação.

Usando uma plataforma com sondas desenhadas nas junções éxon-éxon de 10 mil genes humanos o perfil de variantes de *splicing* foi avaliado em 52 tecidos e linhagens celulares humanas, revelando eventos de *splicing* tecido-específicos bem como novos eventos (JOHNSON et al., 2003). Pan e colaboradores (2004) utilizaram uma plataforma contendo sondas não

apenas correspondentes às junções éxon-éxon mas também sondas éxon-específicas para avaliar alterações tecido-específicas no transcriptoma de 10 tecidos de camundongo. Esses dados também sugerem uma modulação diferencial do nível de expressão de variantes de *splicing* específica nos diferentes tecidos.

Apesar de abordagens de microarranjos de DNA para análise de *splicing* alternativo ter gerado grande quantidade de dados em relação ao perfil de expressão tecido-específico, esses experimentos são restritos a genes com estrutura genômica conhecida e bem definida e focados na análise de eventos de uso de éxons alternativos. Os eventos de *splicing* do tipo retenção de íntrons e uso de sítios de *splice* alternativo não foram avaliados. No mais, o uso de sondas nas junções éxon-éxon, dificulta o desenho de sondas com propriedades iguais de conteúdo GC e temperatura de anelamento, podendo influenciar na eficiência de hibridação, gerando resultados enviezados (CUPERLOVIC-CULF et al., 2006). Além disso, estas abordagens exigem análises computacionais complexas que permitam distinguir entre as alterações de expressão variante-específicas e as diferenças de expressão do gene como um todo, uma vez que os sinais de expressão resultam da soma das intensidades de hibridação de diversas variantes.

1.3.3. Metodologias baseadas em sequências

A grande quantidade de informação presente nos bancos de dados públicos tornou-se uma fonte potencial para o estudo do transcriptoma, acelerado pela publicação da sequência do genoma humano. O mapeamento de sequências expressas, ESTs (do inglês - *expressed sequence tags*) (ADAMS et al., 2001), tanto contra o genoma como contra as sequências completas de RNAm disponíveis em banco de dados foi uma das primeiras iniciativas de identificação de variantes de *splicing* em larga escala (BRETT et al., 2000; BURKE et al., 1998, MIRONOV, et al., 1999). Bibliotecas de ESTs são especialmente interessantes por serem geradas pelo sequênciamento

parcial de transcritos de diversos tecidos e condições patológicas. O alinhamento de ESTs e sequências de cDNA completes contra a sequência genômica permite a definição da estrutura gênica pela identificação dos limites éxon-íntron (KAN; ROUCHKA; GISH 2001;MODREK;LEE, 2003).

O alinhamento de sequências expressas contra o genoma humano, seguido de uma comparação entre os limites éxon-íntron de todas as sequências de um mesmo gene permitem a identificação de variantes de *splicing* alternativo. A melhoria das ferramentas bioinformáticas de alinhamento entre as sequências, que consideram a presença de sítios de *splice* conservados para a definição dos limites éxon-íntron (FLOREA et al., 1998), contribuiu para a obtenção de resultados mais precisos (GUPTA et al., 2004; KAN et al., 2005). As variantes de *splicing* identificadas por metodologias baseadas no alinhamento entre sequências são disponibilizadas em diversos bancos de dados (DRALYUK et al., 2000; HOLSTE et al., 2006; KIM et al., 2007; POSPISIL et al., 2004; STAMM, et al., 2006).

O alinhamento interespecífico entre sequências expressas e a sequência genômica é também uma ferramenta de grande valia na identificação de variantes. A comparação entre sequências expressas e genômicas humanas, de camundongo e rato permitiu não apenas a identificação de novas variantes de *splicing* alternativo como também uma análise evolutiva destes eventos (CHEN, et al., 2006; KAN et al., 2005).

A geração em grande escala de sequências (EST e RNAm) a partir de bibliotecas de cDNA tem disponibilizado grande quantidade de informação, contribuindo de forma importante para a identificação de variantes de *splicing*. Essas estratégias não necessitam de um conhecimento prévio das variantes e permitem a detecção das diferentes formas de *splicing* alternativo, uma vez que utilizam como material inicial bibliotecas de cDNA. No entanto, isoformas raras, expressas em baixo nível, são dificilmente identificadas por essas metodologias. Dessa forma, a implementação de estratégias que favorecem um enriquecimento de variantes de *splicing* na construção das bibliotecas contribuem enormemente para identificação das mesmas. A etapa de enriquecimento é baseada no fato de que duas variantes de *splicing* de um

mesmo gene formam estruturas de heteroduplexes, resultante da hibridação de regiões comuns entre as diferentes variantes. Assim, os heteroduplexes apresentam regiões de dupla-fita, correspondentes aos éxons comuns entre as variantes e alças de simples-fita, que correspondem a regiões unicamente presentes em uma das variantes. As alças de simples fita podem ser recuperadas por clivagem enzimática (FERREIRA et al., 2008), por ligação de oligonucleotídeos randômicos (WATAHIKI et al., 2004) ou proteínas de ligação a cDNA simples-fita (THILL et al., 2006).

Baseado no princípio de formação de heteroduplex, nosso grupo desenvolveu uma estratégia para o mapeamento de sítios de *splice* alternativos utilizando cDNA de uma linhagem luminal epitelial de mama. Nessa estratégia, a região de alça de simples fita foi digerida por uma enzima endonuclease específica de simples-fita (S1 nuclease), gerando fragmentos de cDNA de fita dupla que correspondessem a regiões adjacentes a sítios de *splice* alternativos. Esses fragmentos foram amplificados, clonados, sequenciados e alinhados contra a sequência do genoma humano, possibilitando o mapeamento de sítios de *splice* alternativos (Figura 9A). No entanto, devido a digestão inespecífica de regiões de cDNA dupla-fita pela enzima S1 nuclease, a implementação dessta estratégia para estudos em larga escala foi impossibilitada (FERREIRA et al, 2008).

As alças de simples-fita podem também ser utilizadas como isca para a captura específica e isolamento de moléculas de heteroduplexes a partir de uma amostra heterogênea. Com o intuito de identificar variantes de *splicing* diferencialmente reguladas, Watahiki e colaboradores (2004) utilizaram bibliotecas de cDNA previamente construídas a partir de dois tecidos distintos para a síntese de cDNA senso e antissenso. Após a hibridação foram formadas moléculas de heteroduplexes, representando variantes de *splicing*, bem como moléculas de dupla-fita inteiramente complementares. A captura e isolamento dos heteroduplexes foram realizados utilizando oligonucleotídeos de sequência randômica biotinilados e partículas magnéticas de estreptavidina. Os fragmentos de cDNA foram digeridos e ligados a adaptadores para amplificação, clonagem e sequenciamento (Figura 9B).

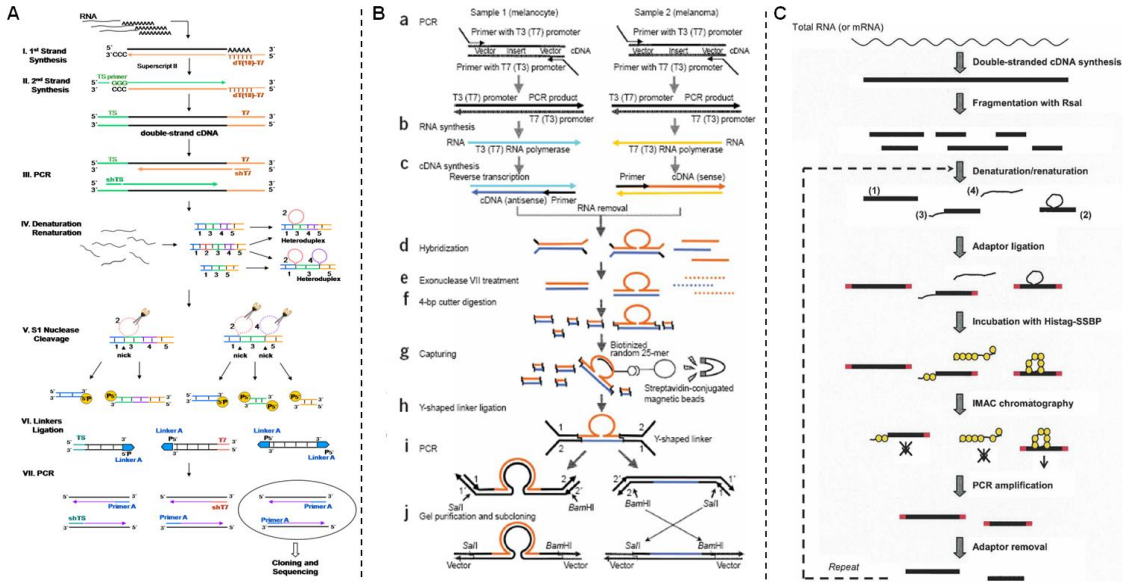


Figura 9: Metodologias de construção de bibliotecas de cDNA para análise de *splicing* alternativo, baseadas na formação de heteroduplexes. A – Metodologia proposta por Ferreira e colaboradores (2008). B – Metodologia proposta por Watahiki e colaboradores (2004). C – Metodologia proposta por Thill e colaboradoores (2006).

No total foram identificados 5.401 genes com evidências de ocorrência de *splicing* alternativo, sendo identificada uma variante nova para 436. Uma vez que essa estratégia depende da construção inicial de duas bibliotecas de cDNA parentais, trata-se de uma metodologia trabalhosa e com custo elevado.

De forma alternativa, Thill e colaboradores (2006) desenvolveram uma estratégia para construção de biblioteca de cDNA enriquecida para *splicing* alternativo a partir de RNA total de placenta. Essa estratégia é também baseada na formação de heteroduplexes, sendo a captura feita por proteínas que se ligam especificamente à região de cDNA simples-fita (Figura 9C). Essa abordagem se mostrou igualmente eficiente no enriquecimento das variantes, e, em comparação com uma biblioteca de cDNA padrão, o enriquecimento na identificação de variantes de *splicing* foi na ordem de 10 vezes.

Atualmente, com o advento de novas tecnologias de sequenciamento em grande escala, a utilização de abordagens baseadas na geração de sequências são ainda mais promissoras (BENNETT et al., 2005;

MARGULIES et al., 2005; SHENDURE et al., 2005), uma vez que os novos instrumentos são capazes de gerar grande quantidade de informação em curto período de tempo, reduzindo sobremaneira o valor de cada base gerada. A grande vantagem dessas técnicas é a dispensa da etapa de clonagem em vetores bacterianos para a construção das bibliotecas, substituída por PCR em emulsão (WILLIAMS et al., 2006) ou amplificação clonal em plataforma sólida. Em relação às tecnologias de sequenciamento, essas abordagens utilizam pirosequenciamento (454-*Roche*) (MARGULIES et al., 2005; RONAGHI; UHLÉN; NYRÉN, 1998) sequenciamento por ligação (Solid-*Applied Biosystems*) (SHENDURE et al., 2005) ou sequenciamento baseado na polimerase (GA-*Illumina*) (BENNETT et al., 2005).

As novas abordagens de sequenciamento prometem alterar a atual visão da complexidade do transcriptoma humano. Os dados gerados a partir do sequênciamento profundo do transcriptoma de diferentes tecidos humanos normais, embrionários e de linhagens celulares contribuíram com uma nova caracterização do transcriptoma humano influenciado por *splicing* alternativo (PAN et al, 2008; SULTAN et al, 2008; WANG; GERSTEIN; SNYDER, 2009). Esses trabalhos sugerem que aproximadamente 95% dos genes humanos sofrem *splicing* alternativo (PAN et al, 2008; WANG; GERSTEIN; SNYDER, 2009), e que 86% dos genes apresentam mais de uma isoforma expressa em frequencia apreciável (superior a 15%), o que sugere que as diversas variantes desempenham um papel importante para a funcionalidade das células (WANG; GERSTEIN; SNYDER, 2009). Foi observado que a maior parte dos eventos de *splicing* alternativo variam entre os diferentes tecidos, sendo as variações entre indivíduos diferentes de duas a três vezes menos comuns (WANG; GERSTEIN; SNYDER, 2009), reforçando a potencialidade da utilização de variantes de *splicing* tumor-específicas como marcadores moleculares. Além disso, esses estudos identificaram novos éxons humanos, bem como novas junções éxon-éxon e sugerem que a exclusão de éxons seja o tipo de evento mais abundante (SULTAN et al, 2008) .

Devido à grande quantidade de informação gerada a partir de tecidos tumorais, muitas dessas abordagens mencionadas foram capazes não apenas de identificar novas variantes de *splicing* como associar a presença

dessas ao aparecimento e desenvolvimento de tumores. Variantes associadas a tumores podem contribuir como marcadores diagnósticos e prognósticos, além de ser potenciais alvos terapêuticos.

1.4. Alterações no padrão de *splicing* alternativo e sua implicância no câncer

Alterações na regulação de *splicing* vêm sendo correlacionadas com 15% a 50% das doenças hereditárias humanas (KRAWCZAK; REISS; COOPER, 1992), incluindo Parkinson (HYMAN et al., 2005; JIANG et al., 2000), Alzeimer (HEINZEN et al., 2007), neurofibromas (BOTTILLO et al., 2007) e câncer (SREBROW; KORNBLIHTT, 2006). Estas alterações podem ocorrer por mutações nos elementos em *cis* como mutações nos sítios doador e aceptor de *splice* ou nas sequências ativadoras e/ou silenciadoras. Como exemplo, mais de 60% das sequências reguladoras de *splicing* do gene *BRCA1*, o qual está associado ao câncer de mama hereditário, são afetadas por mutações (PETTIGREW et al, 2005).

Usualmente as mutações pontuais detectadas no DNA genômico são classificadas em sinônimas, não-sinônimas ou *non-sense*, sem considerar as mutações em sítios de *splice*. No entanto, estimativas sugerem que 60% das mutações que causam doenças levam a defeitos no *splicing* ao invés de alterações na sequência de aminoácidos (LÓPEZ-BIGAS et al, 2005). Mutações sinônimas, isto é, sem alteração na sequência de aminoácidos, e mutações encontradas nos íntrons podem ser erroneamente consideradas como mutações neutras em câncer, uma vez que podem causar alterações importantes nos sítios reguladores de *splicing*, resultando na alteração da proteína. Há também o caso de mutações assumidas como a causa de doenças devido à alteração na sequência de aminoácido, que, após análises mais detalhadas e estudos funcionais, são reveladas como mutações que alteram o *splicing,* causando danos ainda maiores, como a perda de éxons inteiros que podem alterar substancialmente a estrutura da proteína, do que a

simples troca de um aminoácido (BLENCOWE, 2006; LÜTZEN et al., 2008). Portanto, analisar as regiões intrônicas e não codificantes podem ser extremamente importantes para caracterizar o perfil de mutações que levam a alterações no padrão de *splicing* e consequentemente nos transcritos de um gene em uma determinada doença.

A desregulação no padrão de variantes de *splicing* pode também decorrer de alterações nas proteínas reguladoras *trans* devido à ativação ou à repressão de vias de sinalização celular que interfiram com o nível de expressão desses genes, ou ainda por mutações ou variantes de *splicing* nesses genes. Alterações no nível de expressão dos fatores de *splicing* podem estar relacionadas ao aparecimento de variantes aberrantes, ou ainda causar um desbalanço entre duas ou mais variantes expressas em uma determinada célula, acarretando o aparecimento de doenças (BRINKMAN, 2004; PIND; WATSON, 2003; STIECKLER et al., 1999). Análises bioinformáticas sugerem que cerca de 80% dos fatores de *splicing* se encontram mais expressos nos tumores quando comparado ao tecido normal (KIRSCHBAUM-SLAGER et al., 2004).

Alterações no padrão de *splicing* em genes envolvidos com processos celulares importantes, como adesão, proliferação, morte celular, diferenciação, motilidade e invasão têm sido frequentemente encontradas devido à mutações na sequência de nucleotídeos dos elementos *cis* e alterações nas proteínas reguladoras, e podem contribuir para o aparecimento ou progressão tumoral (KALNINA et al., 2005) (Figura 10).

Figura 10: Alterações no padrão de *splicing* alternativo e sua implicância com o câncer. Modificado de Srebrow e Kornblihtt (2006).

Diferentes variantes de *splicing* do gene *CD44,* envolvidos com adesão celular, foram associadas à progressão (PIND; WATSON, 2003) e metástase (NAOR et al., 2002) em câncer de mama. O gene *p53*, um importante regulador do ciclo celular, possui mais de 9 variantes descritas, e a maioria resulta em uma proteína truncada não funcional (KHOURY;BOURDON, 2010). Em relação a influência da expressão de variantes de *splicing* específicas com o controle de proliferação e morte celular, podemos citar os genes *BCLX* (MERCATANTE et al., 2001) e *FAS* (IZQUIERDO et al., 2005). O gene *BCLX* possui duas variantes de *splicing* que codificam isoformas diferentes. A isoforma XS ou curta é pró-apoptótica enquanto a variante longa ou XL é anti-apoptótica (MERCATANTE et al., 2001). De forma similar, a variante de *splicing* do gene *FAS* com inclusão do éxon 6 codifica uma isoforma proteica que atua de forma pró-apoptótica enquanto a variante que não possui este éxon codifica uma isoforma que atua de forma antiapoptótica (IZQUIERDO et al., 2005).

1.4.1. Variantes de *splicing* como marcadores moleculares

Diversas variantes de *splicing* associadas a tumores humanos já foram descritas e podem contribuir para uma melhor compreensão do aparecimento

e desenvolvimento desses tumores. Variantes de *splicing* exclusiva ou preferencialmente expressas em tumores poderiam auxiliar na classificação da doença, ser indicativos da evolução ou ainda ser preditivos da resposta a determinado tratamento.

Em relação a marcador prognóstico de invasão a variante c do gene osteopontina mostrou ser especificamente expressa em carcinoma ductal invasivo quando comparado com tecido normal de mama. Além disso, foi verificada uma associação positiva entre o nível de expressão da variante e o grau do tumor (MIRZA et al., 2008). Outro exemplo interessante em câncer de mama é a presença de alta expressão de isoformas proteicas geradas por duas variantes de *splicing* específicas do gene *VEGF* ou *vascular endotelial growth factor* associadas com menor sobrevida global nas pacientes (KONECNY et al., 2004). Este gene apresenta 9 variantes de *splicing* descritas. As variantes estudadas neste trabalho apresentam exclusão do éxon 6, que codifica a isoforma VEGF-165 e exclusão concomitante dos éxons 6 e 7, que codifica a isoforma VEGF-121, que resultam em menor interação das isoformas proteicas com a matriz celular aumentando sua permeabilidade tecidual, promovendo angiogênese e maior agressividade a uma classe de tumores de mama. Por fim, com base nesses resultados é sugerido o uso combinado de inibidores de VEGF no tratamento desses subtipos tumorais (KONECNY et al., 2004). Nesse caso, as isoformas proteicas de VEGF-165 e -121 são marcadores de agressividade e também indicadores de tratamento.

De forma similar, a variante 1 do gene *Kruppel-like* 6 (*KLF6*) está associada a mau prognóstico em câncer de próstata, pulmão e ovário e a inibição da expressão dessa variante foi acompanhada por aumento de apoptose e regressão tumoral, sendo indicada como alvo terapêutico. Esta variante é oncogênica e de ação antagônica ao transcrito *full-length* considerado um gene supressor de tumor (DIFEO; NARLA; MARTIGNETTI, 2009).

1.4.2. Variantes de *splicing* como alvo terapêutico

A utilização de variantes de *splicing* para o desenvolvimento de alvos terapêuticos é altamente promissora, uma vez que tratamentos moleculares desenvolvidos para atingir especificamente uma variante de *splicing* oncogênica, ao invés de silenciar a ação do gene como um todo, poderia resultar em um tratamento mais eficaz com menores efeitos colaterais adversos. Apesar de poucos eventos de *splicing* terem sido testados clinicamente como alvo terapêutico, é esperado um maior desenvolvimento de terapias gênicas focadas no processamento de RNA, em decorrência do aumento do conhecimento dos mecanismos funcionais dos RNAs de interferência e micro RNAs (PAJARES et al, 2007).

Uma estratégia interessante é a utilização de oligonucleotídeos sintéticos antissenso. Seu modo de ação é através da ligação do oligonucleotídeo ao RNA que pode levar à degradação do RNAm pela ação da enzima RNaseH (específica para moléculas de RNA hibridas), e impedir o reconhecimento do RNAm pelo ribossomo afetando a síntese proteica ou ainda afetar o *splicing* do pré-RNAm (LAUFER; RESTLE, 2008). No caso específico da modulação o mecanismo de *splicing* e a utilização dos oligonucleotídeos podem restaurar a função do gene pela reprogramação do *splicing*, ou inibir a expressão de uma variante pela interrupção do *splicing* (DU; GATTI, 2009). De forma detalhada, a ligação de oligonucleotídeos sintéticos a sítios de *splice* específicos na molécula de pré-RNAm pode dificultar o reconhecimento desses sítios e, assim, diminuir a geração da variante oncogênica sem interferir na geração da proteína normal (Figura 11A), como também agir na molécula de RNA mensageiro processada específica da variante oncogênica impedindo a síntese proteica ou promovendo sua degradação (Figura 11B) (PAJARES et al, 2007).

Por fim, a ação de variantes oncogênicas também pode ser bloqueada pela utilização de anticorpos monoclonais direcionados contra epítopos específicos da variante oncogênica (Figura 11C) (PAJARES et al, 2007).

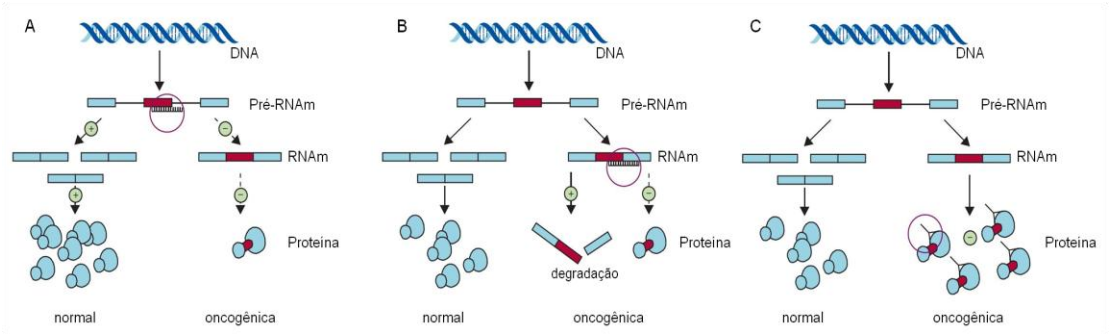


Figura 11: A utilização de variantes de *splicing* como alvo terapêutico. A – Oligonucleotídeos síntéticos podem ser usados para bloquear o reconhecimento de éxons específicos pela maquinaria de *splicing*, impedindo a produção de transcritos oncogênicos. B – Oligonucleotídeos sintéticos que reconhecem variantes de *splicing* oncogênicas podem bloquear a síntese proteica ou sinalizar para degradação do transcrito específico. C – O uso de anticorpos contra variantes oncogênicas podem bloquear especificamente a ação destas isoformas. Modificado de Pajares e colaboradores (2007).

Apesar de promissoras, as terapias baseadas na utilização de oligonucleotídeos precisam superar diversos obstáculos para possibilitar aplicação na clínica. Para que o oligonucleotídeo seja corretamente entregue à célula alvo é necessário evitar a degradação durante a circulação no sangue, conseguir atravessar a membrana celular lipoproteica e escapar das vesículas endossomais, sem causar toxicidade celular, mantendo alta eficiência e especificidade no direcionamento (DU; GATTI, 2009; LAUFER; RESTLE, 2008).

1.5. Câncer de mama: uma doença multifacetada

1.5.1. Epidemiologia, fatores de risco, prevenção e tratamento

O câncer de mama é o segundo tipo de câncer mais frequente no mundo e o primeiro entre as mulheres, sendo responsável por um quinto dos óbitos na faixa etária entre 40 e 50 anos (RADICE; REDAELLI, 2003). No Brasil, o câncer de mama já se apresenta como a neoplasia maligna mais frequente entre as mulheres. O número de casos novos estimados para o ano

de 2010 é de 49.240, sendo as maiores taxas de incidência registradas nos estados da região sul e sudeste (Figura 12). No ano de 2008 foram registradas 11.735 mortes decorrentes de câncer de mama na população brasileira feminina (fonte INCA - Instituto Nacional de Câncer).

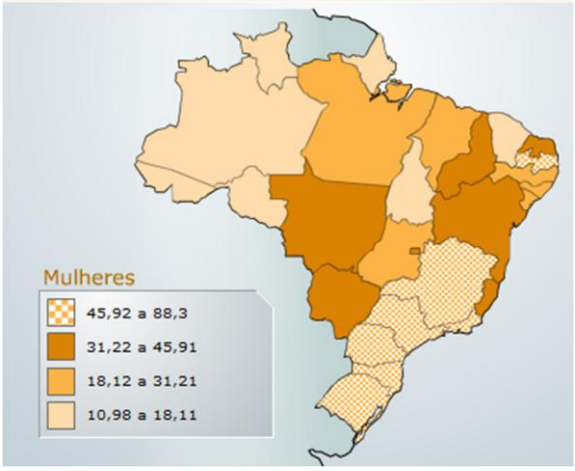


Figura 12: Taxas brutas de incidência da neoplasia maligna da mama por 100.000 mulheres estimadas para o ano 2010, segundo a Unidade da Federação. fonte INCA - Instituto Nacional de Câncer

Diversos fatores de risco que predispõem ao desenvolvimento do câncer de mama foram reportados. O fator de risco individual mais importante é a idade. A incidência desse tipo de câncer dobra a cada 10 anos em mulheres antes da menopausa (MCPHERSON; STEEL; DIXON, 2000; VOGEL, 2008). A idade da menarca e da menopausa também são considerados fatores de risco importante. Quanto mais cedo a menarca e mais tarde a menopausa, maior o risco. Estes dados sugerem associações entre o período de exposição aos hormônios femininos endógenos, estrógeno e progesterona, com o risco do câncer de mama, sendo esses importantes reguladores do desenvolvimento e proliferação do tecido mamário (ANDERSON, 2002). Dentro deste contexto, é de certa forma esperado que a nuliparidade ou idade tardia da gestação também influenciem o risco (BUTT et al., 2009), devido a um aumento de tempo de exposição aos hormônios. Adicionalmente, considerando que o tecido mamário só completa sua diferenciação durante a primeira gestação, quanto mais tarde, maiores as chances de acúmulo de mutações nas células indiferenciadas que possuem

maior capacidade proliferativa. Nesse sentido, o aumento do risco pelo uso de contraceptivos hormonais tem sido extremamente debatido e discutido. No entanto, os resultados encontrados são divergentes e sugerem o papel de outros fatores biológicos e ambientais que contribuem para o risco (HAILE et al., 2006).

Outro fator importante é a predisposição genética responsável por 5% a 10% dos casos de câncer de mama (MCPHERSON; STEEL; DIXON, 2000). Mulheres com parentes de primeiro grau que desenvolveram a doença antes dos 50 anos apresentam um risco duas vezes maior que a população em geral (MCPHERSON; STEEL; DIXON, 2000). Mutações em dois genes, *BRCA1* (*breast cancer* 1) e *BRCA2* (*breast cancer* 2), foram identificadas em alta frequência em famílias com alto risco para o câncer de mama. Estes genes participam do mecanismo de reparo do DNA e mutações que acarretem a perda da funcionalidade podem contribuir para o aparecimento do tumor. Mutações em BRCA1 e BRCA2 aumentam em até 80% a chance de desenvolver câncer de mama até os 75 anos. Algumas síndromes familiais também foram relacionadas com alto risco para o câncer de mama, como a síndrome de Li-Fraumeni (mutações no gene *p53*) e síndrome de Cowden's (mutações no gene *PTEN*).

Outros fatores ambientais parecem contribuir, porém de forma modesta para o aumento no risco do câncer de mama, como alcoolismo, tabagismo, dietas ricas em gordura e sedentarismo, entre outros.

Uma avaliação combinada desses fatores discutidos pode ser utilizada para estimar o risco individual de cada mulher desenvolver câncer de mama, baseado nos modelos de Gail (GAIL, 1989) e no modelo americano desenvolvido pelo Instituto Nacional do Câncer (NCI, do inglês, *National Cancer Institute* - http://www.cancer.gov/Bcrisktool). Esses modelos são baseados principalmente na idade, idade da menarca, idade da primeira gestação, história familial, história clínica e raça. No entanto, não levam em consideração fatores importantes como a densidade mamária, quantidade de hormônio circulante e índice de massa corpórea, que poderiam aumentar a acurácia destas ferramentas (VOGEL, 2008).

Indivíduos detectados como de alto risco para desenvolvimento do câncer de mama podem ser elegíveis para tratamentos preventivos com o uso de tamoxifeno e raloxifeno, ambos inibidores do receptor de estrógeno. Em alguns casos, uma cirurgia redutora preventiva pode ser indicada. No entanto, tratamentos preventivos são polêmicos e sua eficácia não está bem estabelecida.

De forma geral, a detecção precoce da doença é considerada um dos melhores fatores preventivos. Mulheres assintomáticas e sem história familiar devem realizar o autoexame semanalmente a partir dos 20 anos de idade e, dos 20 aos 30 anos, consultar um médico ginecologista regularmente. O exame de mamografia é recomendado a partir dos 40 anos. Esse exame aumenta significativamente a taxa de sobrevida das pacientes com câncer de mama devido à detecção precoce da doença. No entanto, 10% a 15% de todos os cânceres de mama não são detectados na mamografia. Há também exames auxiliares como a ultra-sonografia, exames citológicos (PAAF-punção aspirativa com agulha fina e citologia de descarga papilar) e histopatológicos (biópsia), que combinados aumentam ainda mais a acurácia no diagnóstico da doença.

O tratamento do câncer de mama pode ser local, cirurgia e radioterapia, ou sistêmico, quimioterapia, hormonioterapia e imunoterapia. A cirurgia é um tratamento indicado na grande maioria dos casos de câncer de mama, tanto para remoção da massa tumoral quanto para esvaziamente de gânglios linfáticos atingidos pela doença. O tratamento sistêmico é frequentemente utilizado em combinação com a cirurgia. A quimioterapia neoadjuvante tem com o objetivo diminuir o tamanho do tumor antes de sua remoção enquanto a terapia adjuvante tem o intuito de prevenir recorrência da doença ou o aparecimento de metástases à distância. O tipo de tratamento mais indicado é dependente de diversos fatores clinicohistopatológicos que serão discutidos a seguir.

1.5.2. Classificação histopatológica

A mama feminina normal adulta é constituída por tecido epitelial organizado em lóbulos, estruturas produtoras de leite, e ductos, que são pequenos canais que ligam os lóbulos ao mamilo, além de tecido adiposo, conjuntivo, vasos sanguíneos e vasos linfáticos. O epitélio mamário é formado por uma camada dupla de células sendo a mais interna o epitélio luminal envolto por células mioepiteliais de localização adjacente à membrana basal (SCHNITT 2009).

Por acometer as estruturas epiteliais o câncer de mama é classificado em carcinoma ductal ou carcinoma lobular, de acordo com a região onde ocorre, sendo mais frequente a ocorrência nos ductos do que nos lóbulos. Os tumores de crescimento restrito, que ficam confinados às estruturas epiteliais sem romper a membrana basal, são chamados carcinoma *in situ*. Os carcinomas *in situ* podem ser considerados uma lesão pré-maligna que pode ou não progredir para lesão invasiva (REIS-FILHO; LAKHANI, 2003; SCHNITT, 2009). O carcinoma ductal invasivo ou infiltrante é o tipo mais comum de câncer de mama, responsável por 80% dos casos. Apesar de seu surgimento ocorrer internamente ao ducto, as células tumorais se espalham para o tecido conjuntivo e adiposo adjacente. A identificação de fatores moleculares preditivos da progressão do carcinoma *in situ* para o invasivo é o desafio de vários estudos que têm contribuído para essa questão (CASTRO et al., 2008; HWANG et al., 2004; MA et al., 2003).

Os carcinomas lobulares invasivos são detectados em 10% das pacientes com câncer de mama. Além dos tipos hitológicos lobular e ductal outros tipos são reconhecidos pela Organização Mundial de Saúde, como o carcinoma mucinoso, carcinoma medular, carcinoma papilífero, carcinoma tubular, carcinoma apócrino, carcinoma adenocístico, carcinoma secretor, acarcinoma apócrino e carcinoma metaplásico.

A classificação clínica da doença é baseada no sistema TNM de acordo com análises de exame físico e diagnóstico por imagem. O fator T, em câncer de mama, refere-se ao tamanho do tumor e pode ser classificado em T1, T2, T3 e T4 se o tumor primário é menor ou igual a 2 cm, entre 2 e 5 cm,

mais que 5 cm ou se apresenta extensão direta à parede torácica ou à pele, respectivamente. O segundo aspecto avaliado é a presença de linfonodos acometidos (N), sendo o tumor classificado como N0 na ausência de metástases linfonodais e N1, N2 ou N3 na presença de células tumorais em linfonodos axilares homolaterais móveis, fixos, intraclaviculares, internos e/ou de aparência clínica (detectado por exame clínico ou por estudos de imagem). Por fim, a presença de metástases à distância é avaliada e sua presença ou ausência é classificada por M1 ou M0, respectivamente. Na impossibilidade de avaliação de algum dos três aspectos, é aplicada a letra “x”para designação (Tx, Nx ou Mx).

A classificação patológica é semelhante à classificação clínica, seguindo critérios pTNM, e requer o exame do carcinoma primário sem tumor macroscópico nas margens de ressecção. As categorias pT correspondem às categorias T, no qual apenas o tamanho do componente invasor é considerado. A classificação patológica requer a ressecção e o exame dos linfonodos axilares inferiores, pelo menos, e em geral inclui 6 ou mais linfonodos, sendo pN1 quando há acometimento de 1 a 3 linfonodos, pN2 de 4 a 9 linfonodos e pN3 mais de 10 linfonodos. As categorias pM correspondem às categorias M.

De acordo com a classificação TNM os tumores são agrupados em diferentes graus de estadiamento, de acordo com a tabela 2.

Tabela 2: Graus de estadiamento em câncer de mama, de acordo com a classificação de TNM. Tis – tumor *in situ*.

Estadio	T	N	M
Estadio 0	Tis	N0	M0
Estadio I	T1	N0	M0
Estadio IIA	T0	N1	M0
	T1	N1	M0
	T2	N0	M0
Estadio IIB	T2	N1	M0
	T3	N0	M0
Estadio IIIA	T0	N2	M0
	T1	N2	M0
	T2	N2	M0
	T3	N1, N2	M0
Estadio IIIB	T4	N0, N1, N2	M0
Estadio IIIC	Qualquer T	N3	M0
Estadio IV	Qualquer T	Qualquer N	M1

Outra medida de estadiamento tumoral para o câncer de mama é dada pela classificação de Scarff-Bloom-Richardson, mais conhecida como grau SBR. Essa classificação é baseada no grau de diferenciação das células tumorais, que leva em conta a capacidade do tumor formar estruturas tubulares, glandulares e papilares; na forma do núcleo celular; e no índice mitótico. Cada um destes três componentes são avaliados individualmente e pontuados de 1 a 3, e posteriormente somados, gerando valores de 3 a 9. Esta pontuação classifica o grau de diferenciação em bem diferenciado (pontuação de 3, 4 e 5), moderadamente diferenciado (pontuação de 6 a 7) e fracamente diferenciado (pontuação de 8 a 9), e quanto mais indiferenciado maior a agressividade do tumor.

1.5.3. Marcadores moleculares: nova classificação do câncer de mama baseada no perfil molecular

Além da classificação histopatológica clássica dos tumores de mama, importantes marcadores moleculares são utilizados na rotina clínica como forma de auxiliar prognóstico da doença e direcionar a conduta terapêutica (DUFFY, 2005; MOLINA et al., 2005). A detecção da expressão dos receptores hormonais de estrógeno (RE) e progesterona (RP) e a amplificação e ativação do oncogene *ERBB2* constituem os três marcadores mais usados em câncer de mama e recomendados pela Associação Americana de Oncologia Clínica (ASCO – HARRIS et al., 2007). Os receptores hormonais são membros de uma grande família de receptores nucleares, que agem como fatores de transcrição ativando a expressão de diversos genes e promovendo proliferação celular e inibição da apoptose (FRASOR et al., 2003). O receptor de estrógeno promove a ativação do gene codificador do receptor de progesterona. Portanto, a detecção do receptor de estrógeno concomitante à expressão do receptor de progesterona sinaliza que o RE está funcional. Cerca de dois terços dos cânceres de mama são positivos para receptores hormonais, e entre 60% a 80% dos tumores apresentam expressão do RE e entre 44% a 61% apresentam expressão de RP (EISENBERG; KOIFMAN, 2001).

Os receptores hormonais são marcadores moleculares indicativos de bom prognóstico. Tumores receptor hormonal positivos apresentam comportamento mais indolente, de crescimento tumoral mais lento e sua presença é sugestiva para hormonioterapia. O tamoxifeno é um inibidor seletivo do receptor de estrógeno, que interage com o receptor alterando sua estrutura conformacional e inibindo parcialmente sua ação. A ação do receptor de estrógeno depende da interação com outras proteínas que atuam como cofatores, sendo os principais AF-1 e AF-2. O tamoxifeno altera a capacidade de interação de RE inibindo sua ação via AF-2, sem interferir com a ativação gênica via AF-1 (TZUKERMAN et al., 1994; McDONNELL et al., 1995). A complexidade da inibição seletiva de tamoxifeno possibilita o

desenvolvimento de mecanismos de resistência ao tratamento (RING; DOWSETT, 2004).

O receptor *ERBB2* (HER-2 / neu) é um proto-ocogene que se encontra altamente expresso em 25 a 30% dos tumores de mama (SLAMON et al., 1989). A amplificação desse gene é considerado um fator prognóstico adverso (SLAMON et al., 1987), no entanto é o primeiro oncogene utilizado como alvo para terapia do câncer de mama (BASELGA et al., 1996; COBLEIGH, et al., 1999; GUSTERSON et al., 1992; SLAMON et al., 2001; SLAMON;PELGRAM, 2001). A classificação imunohistoquímica da expressão de *ERBB2* feita por imunohistoquímica classifica os tumores como 0 – ausência de marcação ou marcação em menos de 10% das células tumorais; 1+ - marcação fraca de membrana em mais de 10% das células tumorais e marcação apenas parcial da membrana; 2+ - marcação de membrana fraca à moderada em mais de 10% das células tumorais e marcação completa da membrana; 3+ - marcação forte e completa da membrana em mais de 30% das células tumorais. Além disso, nos casos de marcação intermediária (2+) a confirmação da amplificação gênica pode ser realizada pelo método de FISH (do inglês, *fluorescence in situ hibridization*), que avalia a amplificação genômica de *ERBB2*.

Esse gene é integrante da família de receptores do tipo tirosina-quinase, que contém 4 membros: *ERBB1* ou *EGFR*, *ERBB2*, *ERBB3* e *ERBB4*. Esses receptores transmembrânicos são ativados na presença de um ligante que proporciona uma dimerização entre dois receptores e a ativação do domínio quinase intracelular que promove uma fosforilação cruzada nos resíduos de tirosina, ativando uma cascata de forforilação e ativação de diversas vias de sinalização (HYNES; MACDONALD, 2009). Nenhum ligante específico foi descrito para o receptor ERBB2, no entanto sua ativação ocorre por heterodimerização com os demais membros da família. Em células tumorais, a alta expressão de ERBB2 leva à formação de homodímeros mesmo na ausência de um ligante, resultando em uma ativação constitutiva (DOUGALL et al., 1994). A formação de diferentes homo e heterodímeros resulta na ativação de diferentes vias de sinalização intracelular (Figura 13), sendo principalmente ativadas as vias MAPK,

PIK3/Akt e mTOR, devido a formação de hetrodímeros entre ERBB2 e ERBB3 (HYNES; MACDONALD, 2009).

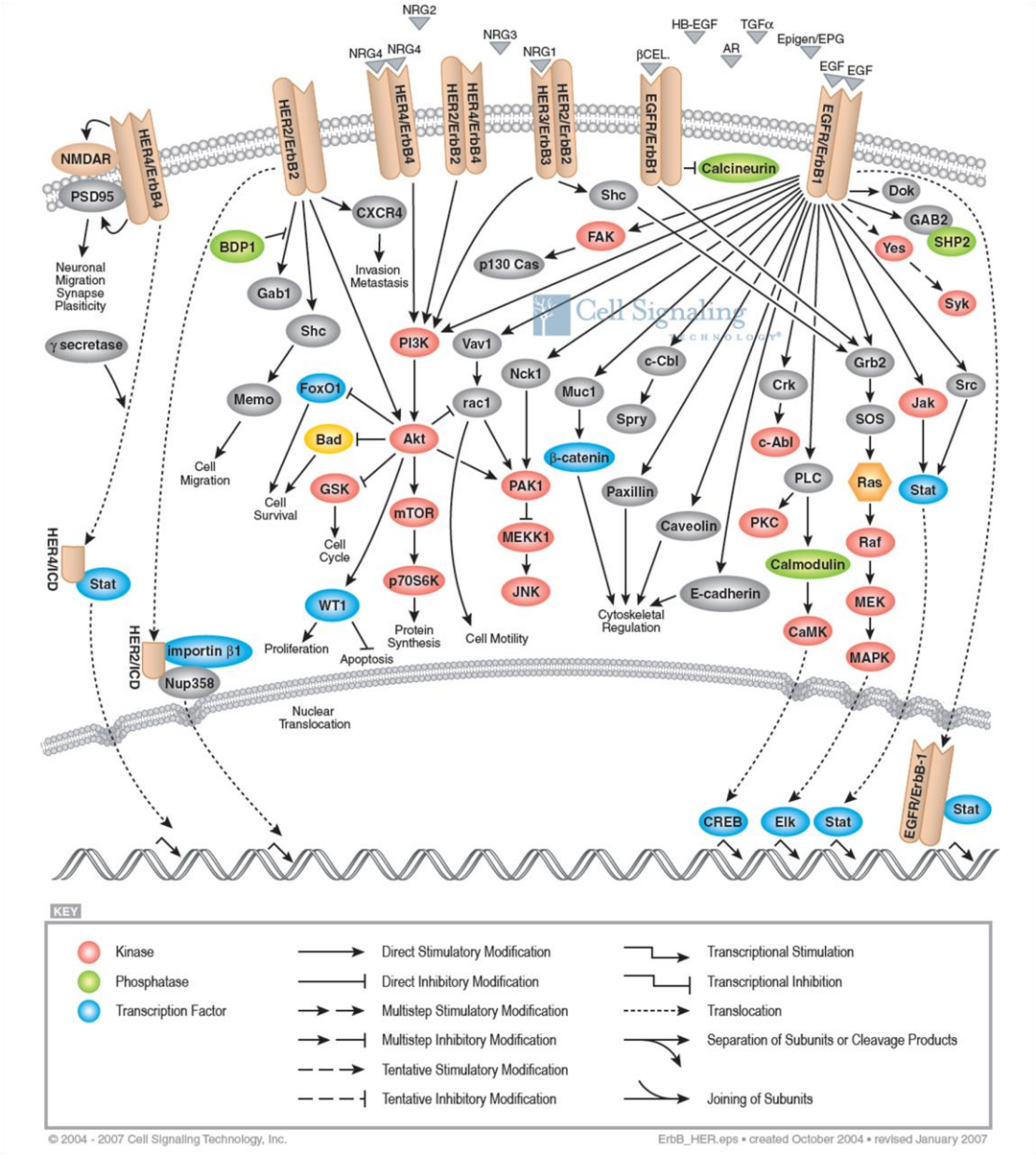


Figura 13: Vias de sinalização celular ativadas pelos receptores tirosina quinase da família ERBB. A formação de homo ou heterodímeros entre os diferentes membros da família (ERBB1, ERBB2, ERBB3 e ERBB4) ativa de diferentes proteínas intracelulares. O receptor ERBB2 ativa principalmente as vias de sinalização PI3K/AKT e a via das enzimas MAPK. Retirado de *Cell Signaling Technology*, Inc.

Devido ao fato do oncogene *ERBB2* ter papel importante de desencadeamento do processo tumoral em uma elevada frequência de casos (DI FIORE et al., 1987), foram desenvolvidos anticorpos e pequenas moléculas inibidoras de domínios tirosina-quinase para inativação da ação

desse oncogene. O anticorpo monoclonal traztuzumab (também conhecido como Herceptin) tem demostrado aumento significativo na taxa de sobrevida livre de doença em pacientes. Esse anticorpo específico reconhece o domínio extracelular do receptor, impede sua dimerização e assim inibe sua atividade de quinase (HYNES; LANE, 2005). Por outro lado, lapatinib é uma pequena molécula que reconhece a porção intracelular do receptor e também inibe sua fosforilação (HYNES; LANE, 2005).

No entanto, dois terços dos tumores de mama metastáticos classificados como ERBB2 positivos não respondem à imunoterapia (VOGEL et al, 2002). Os mecanismos de resitência ao traztuzumab ainda não foram completamente esclarecidos, no entanto alguns estudos sugerem que alterações em outros genes da via PI3K/Akt, como mutações em *PTEN*, resultariam em mecanismos alternativos de ativação da via de sinalização conferindo resistência ao tratamento (BERNS et al, 2007; HYNES; MACDONALD, 2009).

Devido à implicação do receptor ERBB2 no desenvolvimento de câncer de mama foi estabelecido um sistema modelo para investigar sua função no processo tumorigênico nesse tecido, transfectando uma linhagem luminal epitelial de mama, HB4a, com o cDNA do gene *ERBB2*, originando as linhagens HB4a C3.6 e HB4a C5.2 (HARRIS et al., 1999). A linhagem HB4a foi estabelecida a partir de células epiteliais do lúmen de mama imortalizadas (STAMPS et al., 1994) e, portanto, apresenta características típicas de uma célula luminal, como a alta expressão de de citoqueratinas 18 e 19. As células desta linhagem são cubóides, bem organizadas e apresentam inibição de crescimento por contato (HARRIS et al., 1999). A linhagem HB4a C5.2 foi originada a partir de células da linhagem HB4a transfectadas com quatro cópias de uma construção do cDNA *full-lenght* do gene *ERBB2* sob controle do promotor do vírus *long terminal repeat* (MMTVLTR) e sinais de poliadenilação SV40. Essa linhagem expressa aproximadamente 106 receptores ERBB2 por célula, a qual é uma quantidade similar a linhagem tumoral de mama SKBR3 ou ainda superior à linhagem tumoral de mama BT474. Como resultado, as células apresentam características morfológicas distintas da linhagem parental, sendo as células da linhagem C5.2 finas e

alongadas com perda de inibição de crescimento por contato (HARRIS et al., 1999).

Ensaios bioquímicos mostraram que os altos níveis de expressão de *ERBB2* foram suficientes para causar transformação *in vitro*, mas não *in viv*o (HARRIS et al., 1999). Esse modelo vem sendo utilizado por muitos trabalhos que procuram investigar a função do gene *ERBB2* no processo de tumorigênese bem como caracterizar as proteínas envolvidas nas vias de ativação de ERBB2 (DOS SANTOS et al., 2006; MACKAY et al., 2003; TIMMS et al., 2002).

Recentemente a utilização de outros marcadores moleculares com potencial prognóstico e preditivo de resposta a tratamento tem sido proposta. Em especial, dois testes baseados em avaliação do perfil molecular dos tumores estão disponíveis, e sua aplicabilidade tem sido extensamente discutida. O teste *Oncotype Dx*, disponível desde janeiro de 2005, é um ensaio de 21 genes que prediz a eficácia da hormonioterapia para pacientes com tumor em estágios iniciais (estádio I e II) que sejam RE positivo e não apresentem acometimento linfonodal. Esse teste, baseado em reações de RT-PCR, classifica os tumores com baixo, intermediário ou alto escore de recorrência, no qual pacientes com baixo escore de recorrência apresentam maior benefício do tratamento com tamoxifeno sozinho (PAIK et al, 2004). Assim, pacientes com baixo escore de recorrência não são indicados para tratamento quimioterápico adjuvante, melhorando sua qualidade de vida. No outro extremo, pacientes com alto escore de recorrência parecem ter grande benefício do tratamento quimioterápico (PAIK et al, 2004). Recomendações para pacientes com escore intermediário são incertas (SPARANO; PAIK, 2008).

O segundo teste disponível, *Mamaprint*, é baseado em uma plataforma de microarranjos de DNA contendo 70 genes e apresenta uma assinatura prognóstica do risco de desenvolvimento de metástase à distância em pacientes com câncer de mama que não apresentem comprometimento linfonodal (VAN DE VIJVER et al., 2002; VAN'T VEER et al., 2002). Esse teste também tem o intuito de melhorar a seleção dos pacientes que serão beneficiadas pelo tratamento quimioterápico sistêmico adjuvante. Apesar de

promissores, ainda é discutível a real eficácia de sua utilização, principalmente devido à falta de direcionamento no tratamento dos pacientes com risco intermediário.

Em contrapartida, devido à alta heterogeneidade encontrada entre os cânceres de mama, a utilização de um perfil molecular para melhor classificação é uma alternativa interessante em relação à classificação anatomo-patológica, uma vez que tumores com o mesmo grau, estadio e tipo histológico podem apresentar diferentes evoluções e respostas à terapia (ELSTON et al., 1991). A nova classificação proposta é baseada no perfil de expressão gênica capaz de separar os tumores de mama em cinco grandes grupos: Luminal A, Luminal B, Her2 positivos, basal e *normal-breast like* (PEROU et al., 2000; SORLIE et al., 2001).

O subtipo Luminal A é um subtipo de bom prognóstico caracterizado principalmente pela expressão do receptor de estrógeno e ausência de expressão de ERBB2 e expressão de citoqueratinas características de células epiteliais luminais como as citoqueratinas 7, 8, 18 e 19. O subtipo Luminal B também apresenta expressão do RE, no entanto, diferentemente do subtipo Luminal A, apresenta expressão de *ERBB2*, com menor expressão das citoqueratinas 7, 8, 18 e 19, e pior prognóstico. O subtipo Her2 positivo, como o próprio nome indica, tem os níveis mais elevados de expressão de *ERBB2* e ausência de expressão do RE, sendo um subtipo de mau prognóstico, no entanto apresenta resposta positiva ao tratamento com anticorpo monoclonal traztuzumab. O subtipo basal apresenta características de células basais ou mioepiteliais como a expressão das citoqueratinas 5, 6, 14 e 17, além da ausência de expressão dos receptores hormonais e ERBB2. É um subtipo de mau prognóstico que não responde aos tratamentos hormonais e imunoterápicos disponíveis. O subtipo *normal-breast like* apresenta um perfil de expressão gênica similar ao tecido adiposo e outras células não epiteliais, é um subtipo de crescimento lento e pouco agressivo, indicativo de bom prognóstico, apesar de também não responder ao tratamento hormonal e imunoterápico devido à ausência de expressão dos receptores hormonais e do ERBB2. Os dois últimos subtipos apresentados são frequentemente agrupados em uma classe chamada de tumores triplo-

negativos, devido à ausência de expressão de RE, RP e ERBB2. Apesar dessa classificação molecular dos tumores ter contribuído com maior valor preditivo da evolução e da resposta a tratamento em relação à classificação clássica histopatológica, tumores com mesma classificação histológica e molecular podem ter prognóstico e evolução da doença muito distintos.

1.6 Câncer de mama, *ERBB2* e *splicing* alternativo: considerações finais

O câncer de mama é uma doença extremamente heterogênea, sendo influenciada por diversos fatores ambientais e genéticos. Apesar da combinação da classificação histológica com o uso de marcadores moleculares, tumores agrupados em uma mesma classe podem evoluir de formas distintas. Além disso, apesar da inclusão de novos tratamentos mais específicos que a quimioterapia sistêmica, muitos tumores não apresentam a resposta esperada e evoluem de forma agressiva. Portanto a identificação de novos marcadores para o câncer de mama poderia auxiliar no desenvolvimento de marcadores prognósticos e de predição de resposta terapêutica mais acurados, e servir como novos alvos terapêuticos para tratamentos mais específicos e até individualizados (CAREY, 2010). Nesse contexto, a identificação de variantes de *splicing* associadas a tumores humanos podem ter papel importante no sentido de contribuir para uma melhor compreensão da biologia desses tumores, e ainda ser utilizadas como marcadores moleculares de diagnóstico e prognóstico mais acurados e sensíveis, bem como ser utilizadas como alvo terapêutico mais específico com menor prejuízo para as células sadias.

O receptor tirosina quinase ERBB2 é um importante desencadeador do processo tumorigênico (DI FIORE et al., 1987) em aproximadamente 30% dos tumores de mama, no qual é detectado alta expressão deste gene (SLAMON et al., 1989). O oncogene ERBB2 pode alterar a regulação do *splicing* alternativo pela fosforilação de fatores de *splicing* (MUKHERJI et al.,

2006) pela ativação das vias de sinalização PI3K/Akt e MAPK (MATTER et al., 2002; WHITE et al., 2010), levando a produção de transcritos oncogênicos.

Assim, a busca de variantes de *splicing* associadas ao câncer de mama mediado pela amplificação de *ERBB2* pode contribuir para identificação de novos marcadores moleculares que possam auxiliar a melhor classificação desse subtipo, com a identificação dos tumores com pior prognóstico e indicação da resposta do tratamento terapêutico.

2. Objetivos

2.1 Objetivo Geral

O estabelecimento de metodologias para identificação de variantes de *splicing* em câncer de mama sob influência da alta expressão do oncogene *ERBB2*.

2.2 Objetivos específicos

2.2.1 Biblioteca de cDNA enriquecida para *splicing* alternativo

a) Estabelecimento da metodologia para construção de bibliotecas de cDNA enriquecidas para *splicing* alternativo a partir RNA amplificado.

b) Identificação de variantes de *splicing* em amostras tumorais de mama com alta expressão de ERBB2.

c) Validação das variantes de *splicing* por RT-PCR seguido de sequenciamento.

d) Avaliação do padrão de expressão de variantes de *splicing* em duas amostras de mama com diferentes níveis de expressão do oncogene *ERBB2*.

2.2.2 Biblioteca de cDNA para análise de transcriptoma completo

a) Identificação de novas variantes de *splicing* em câncer de mama pela exploração de dados gerados pelo sequênciamento em larga escala de duas linhagens celulares de mama que apresentam ou não alta expressão de *ERBB2*.

b) Validação das novas variantes de *splicing* identificadas por RT_PCR seguida de sequênciamento.

c) Avaliação da regulação das variantes de *splicing* pela expressão diferencial de *ERBB2*.

3. Material e Métodos

3.1 Cultura de células

As linhagens celulares HB4a e C5.2 foram gentilmente cedidas pelo Dr. Michael O´Hare (Instituto Ludwig de Pesquisa sobre o Câncer – Nova Iorque). Essas linhagens foram cultivadas em meio RPMI com 10% de soro fetal bovino (SFB) e 1% de L-Glutamina na presença de 5 µg/ml de insulina e 5 µg/ml de hidrocortisona e mantidas em estufa úmida contendo 50 ml/l CO_2 a 37°C. O meio de cultura foi trocado a cada dois dias e após um período de 10 dias em cultura as células foram lisadas para extração de RNA.

3.2 Extração de RNA

A extração de RNA foi feita através da técnica de Sedimentação em Cloreto de Césio (Glisin et al., 1974). As células foram homogeneizadas em 9 ml de solução de lise (4 M de isotiocianato de guanidina / 25 mM citrato – pH 7.0 / 0.1 M β-mercaptoetanol), e o lisado foi aplicado em 4 ml de uma solução de gradiente de cloreto de césio (5.7 M CsCl e 25 mM de NaAc) e centrifugado a 29.000 rpm em ultracentrífuga Beckman utilizando o rotor SW40Ti durante 17 horas a 20°C. Após a centrifugação, o RNA precipitado no fundo do tubo foi solubilizado em 50 a 200 µl de água biodestilada tratada com dietilpirocarbonato DEPC, para inibição de Rnases. A qualidade e quantidade do DNA obtido foram avaliadas através de leitura em espectofotômetro (O.D 260 e 280 nm) e visualização em gel de agarose 1.0%.

3.3 Amostras tumorais

As amostras de RNA utilizadas neste estudo foram amostras disponíveis no banco de biorepositórios do Hospital A.C. Camargo provenientes de tecido congelado disponíceis no banco de tumores do departamento de anatomia patológica da mesma instituição e que contenham consentimento pré-informado e estejam em acordo com o comitê de ética do Hospiral A. C. Camargo (#952/07). Foram utilizadas 5 amostras dissecadas manualmente de carcinoma ductal invasivo de mama positivas para o marcador ERBB2 nas análises imunohistoquímicas, com sinal de positividade 3+, que consiste em marcação de membrana forte e completa em mais de 30% das células tumorais (Tabela 3). O RNA total destas amostras foram extraídos pelo método de Trizol (Invitrogen), segundo as especificações do fabricante.

Tabela 3: Características clínicas das amostras de carcinoma ductal invasivo. Idade – idade da paciente no momento do diagnóstico. TNM – classificação clínica baseada no sistema TNM. LN – Comprometimento de linfonodo sentinela. Grau – Classificação segundo o sistema de de grau segundo Scarff-Bloom-Richardson, ou grau SBR. Marcadores moleculares –RE, receptor de estrógeno; RP, receptor de progesterona; p53, proteína p53 e ERBB2, proteína ERBB2.

Amostra	Idade	Estadio	TNM	LN	Grau de SBR	Marcadores Moleculares
9T	55 anos	IIa	T2N0M0	Negativo	Grau I	RE +/ RP +/ p53 -/ ERBB2+ (3+)
20T	87 anos	IIb	T2N0M0	Negativo	Grau II	RE +/ RP -/ p53 - / ERBB2+ (3+)
22T	56 anos	IIb	T2N1M0	Positivo	Grau III	RE +/ RP -/ p53 - / ERBB2+ (3+)
28T	42 anos	IIIa	T2N2M0	Positivo	Grau II	RE +/ RP -/ p53 - / ERBB2+ (3+)
36T	45 anos	I	T1N0M0	Negativo	Grau III	RE +/ RP -/ p53 - / ERBB2+ (3+)

3.4 Tratamento com DNAse

As amostras de RNA de linhagens e de amostras tumorais foram tratadas com 1 unidade da enzima DNAseI (DNA-*free* kit, Ambion), por 30 min a 37^0C, segundo as especificações do fabricante. Para checar a eficiência do tratamento e ausência de DNA genômico nas amostras foi realizada uma reação de PCR com iniciadores localizados nos íntrons do gene *MLH1* (iniciador *foward*: TGGTGTCTCTAGTTCTGG; iniciador *reverse*: CATTGTTGTAGTAGCTCTGC) com a enzima *Platinum Taq DNA Polimerase* (Invitrogen), segundo as recomendações do fabricante. A ausência de amplificação de um produto de 254pb confirma a eficácia do tratamento. Como controle positivo da reação de PCR foi utilizado DNA genômico da linhagem HB4a.

3.5. Construção de bibliotecas de cDNA enriquecida para *splicing* alternativo

3.5.1. Síntese de cDNA a partir da amplificação de RNAm

A metodologia utilizada para amplificação do RNAm seguida de síntese de cDNA dupla-fita foi baseada na metodologia proposta por Saraiva e colaboradores (2006). Para a síntese da primeira fita de cDNA a partir de RNA total o RNA foi incubado com 0,75 µg de oligonucleotídeo oligodT-T7 [5'AAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGCGCT(24)3'], que contém o sítio da enzima *T7 RNA polimerase*, em volume final de 6 µl por 10 minutos a 70^0C. Foram adicionados 1X de tampão *First Strand Buffer*; 0,01 M DTT; 1 mM dNTP; 1,5 µg de oligonucleotídeo TS (5'AAGCAGTGGTAACAACGCAGAGTACGCGGG3'); 40 unidades da enzima RNasin® Ribonuclease Inhibitor (Promega) e 400 unidades da enzima *Superscript*II (Life Technologies) em volume final de 20 µl. A reação foi

incubada a 42^{0}C por duas horas. Em seguida, a síntese da segunda fita de cDNA foi realizada com reagentes do Advantage® cDNA PCR Kit (Clontech). Foram adicionados ao cDNA simples-fita 1X *Advantage PCR Buffer*; 1 mM dNTP; 1,4 unidades da enzima ribonuclease H (RNAse H, Invitrogen); 1X *Advantage Polimerase Mix*, em volume final de 100 µl. A reação foi incubada a 37^{0}C por 10 minutos; 94^{0}C por 3 minutos; 65^{0}C por 5 minutos e 75^{0}C por 30 minutos. Para inibir a ação das enzimas foram adicionados 50 mM NaOH e 0,1 mM EDTA e a solução incubada a 65^{0}C por 10 minutos. Por fim o cDNA dupla-fita (dscDNA) foi purificado por fenol:clorofórmio:álcool isoamílico (item 3.7.3) e em colunas *Microcon YM-100 Centrifugal Filter Unit* (Millipore – item 3.7.1). O dscDNA purificado utilizado para reação de trancrição *in vitro* com *RiboMAX™ Large Scale RNA Production Systems* (Promega). Na reação foram adicionados 1X tampão, 7,5 µM de cada rNTP (rATP, rCTP,rGTP e rUTP) e 2,5 µl *Enzyme T7 Mix* (Invitrogen) em volume final de 25,0µl. A reação foi incubada a 37°C por 6 horas. O RNA amplificado foi purificado com utilização do reagente TRIzol® Reagent (Sigma – Aldrich Corporation), segundo as recomendações do fabricante. Para síntese de primeira fita do cDNA, o RNA amplificado (6 µl) foi incubado com 1 µg de oligonucleotídeo TS (5'AAGCAGTGGTAACAACGCAGAGTACGCGGG3') por 10 minutos a 70^{0}C. Foram adicionados 1X de tampão; 0,01 M DTT; 1 mM dNTP; 40 unidades de RNasin® Ribonuclease Inhibitor (Promega); 400 unidades da enzima *Superscript*II (Life Technologies) e 0,5 µg de oligonucleotídeo dT(24) em volume final de 20 µl. A reação foi incubada a 42^{0}C por duas horas. A síntese da segunda fita do cDNA foi realizada com reagentes do Advantage® cDNA PCR Kit (Clontech), segundo protocolo descrito no acima. O dscDNA foi purificado por fenol:clorofórmio:álcool isoamílico (item 3.7.3) e em coluna *Microcon YM-100 Centrifugal Filter Unit* (Millipore – item 3.7.1).

3.5.2. Desnaturação e renaturação

Para desnaturação as moléculas de cDNA foram aquecidas a 96°C por 20 minutos. Posteriormente foram renaturadas lentamente por incubação a

42°C por 24 horas na presença de 0,2% SDS; 0,5 M NaCl; 0,05 M Tris-HCl pH 7,5 e 30% formamida em volume final de 30 µl para favorecer a formação de moléculas de heteroduplexes. Ao final da reação a amostra foi purificada pelo GFX PCR and Gel Band Purification Kit (item 3.7.1)

3.5.3. Clivagem com a enzima exonuclease VII

A enzima Exonuclease VII (USB) foi utilizada para degradação especifica de moléculas de cDNA simples-fita com extremidades livres. Cada 1 µg de amostra foi incubado com 70 mM Tris-HCI, pH 8.0; 8 mM EDTA, pH 8.0; 10 mM 2-mercaptoethanol; 50 µg/ml BSA e 0,2 unidades da enzima Exonuclease VII em volume final de 50 µl. A reação foi incubada a 37°C por 30 minutos e posteriormente a 95°C por 10 minutos para inativação da enzima. A purificação foi feita pelo método de fenol:clorofórmio:álcool isoamílico (item 3.7.3).

3.5.4. Digestão com a enzima de restrição *Dpn*II

A enzima *Dpn*II (*New England Biolabs*) reconhece o seguinte sítio de restrição:

Por ser constituído por 4 nucleotídeos é esperado que este sítio seja encontrado a cada 256pb na molécula de DNA.

Na reação foram utilizadas 10 unidades da enzima na presença de 1X tampão em volume final de 15 µl. A reação foi incubada a 37ºC por 3 horas, sendo que após 1 hora foi adicionado mais 5 unidades da enzima. A amostra foi purificada pelo método de fenol:clorofórmio:álcool isoamílico (item 3.7.3). Como controle positivo da reação de clivagem foi utilizado um fragmento de 459pb que contém um único sítio de reconhecimento da enzima, gerando dois fragmenteos menores de 127 e 332pb.

3.5.5. Recuperação das estruturas de heteroduplex por purificação biotina-estreptavidina

3.5.5.1. Ligação ao oligonucleotídeo 25-mer randômico biotinilado

Para recuperação das estruturas de heteroduplex foi utilizado um oligonucleotídeo formado por 25 nucleotídeos combinados randomicamente o qual possui uma molécula de biotina na extremidade 5'. Cem picomoles deste oligo foram incubados com a amostra de cDNA na presença de 6X SSC e 0,1% SDS a 65 ºC por 16 horas.

3.5.5.2. Preparo das partículas magnéticas

Para cada amostra foram utilizadas duas alíquotas de 1 mg de partículas magnéticas com estreptavidina (Roche). Cada alíquota foi lavada com 300 µl de tampão de ligação TEN100 (10 mM Tris-HCl; 1mM EDTA; 100 mM NaCl, pH 7.5) três vezes. As partículas magnéticas foram matidas no tampão de ligação TEN100 até o momento de uso.

3.5.5.3. Purificação biotina-estreptavidina

A solução de 100 µl de cDNA e oligonucleotídeo biotinilado foi adicionada a uma alíquota de partículas magnéticas com estreptavidina (1 mg) e incubada por 30 minutos a temperatura ambiente em rotação, permitindo a ligação entre a biotina e a estreptavidina. Em seguida, a amostra foi colocada no separador de partículas magnéticas e a solução líquida foi removida e adicionada à segunda alíquota de partículas magnéticas (1 mg) para aumentar a porcentagem de recuperação das estruturas de heteroduplex. Esta segunda alíquota foi igualmente incubada por 30 minutos a temperatura ambiente em rotação, e, em seguida, a amostra foi colocada no separador de partículas magnéticas e a solução líquida descartada. A

primeira e a segunda alíquota de partículas foram misturadas em 200 µl de tampão de lavagem TEN1000 (10 mM Tris-HCl; 1 mM EDTA; 1 M NaCl, pH 7,5). A solução final foi colocada no separador de partículas magnéticas e a solução líquida descartada. Foram realizados mais dois ciclos de lavagem com o tampão TEN1000. Para eluição foi adicionada uma solução desnaturante de 6 M Guanidina-HCl para desfazer a ligação entre biotina e estreptavidina. Após 40 minutos de incubação em rotação a solução foi colocada no separador de partículas magnéticas e a solução líquida foi removida e reservada. Após a eluição a amostra foi purificada pelo método de fenol:clorofórmio:álcool Isoamílico (item 3.7.3).

3.5.6. Ligação aos adaptadores

Os adaptadores XDPN são oligonucleotídeos sintetizados comercialmente desenhados especificamente para conter região de complementariedade do sítio GATC coesivo gerado após clivagem com a enzima *Dpn*II. A sequência e a estrutura dos oligonucleotídeos estão descritas abaixo:

XDPN12: 5'GATCTCTCGAGT3'

XDPN14: 5'CTGATCACTCGAGA3'

5' CTGATCACTCGAGA**GATC**.....GATCTCTCGAGT 3'
3' TGAGCTCTCTAG**.....CTAG**AGAGCTCACTAGTC 5'

Fragmento de cDNA
clivado por *Dpn*II

Inicialmente foram adicionados ao dscDNA 1X tampão, 400 pmol do oligonucleotídeo XDPN12 e 400 pmol do oligonucleotídeo XDPN14. Esta solução foi incubada a 55ºC por 1 minuto. Em seguida houve uma diminuição de 2ºC de temperatura a cada 2 minutos, de 54°C até 28ºC. A temperatura foi diminuída de 28°C a 14ºC por redução de 2ºC de temperatura a cada 4 minutos. Estas condições favorecem um anelamento mais acurado dos oligonucleotídeos as sequências complementares. Só então foram adicionadas 2000 unidades da enzima T4 DNA Ligase (Invitrogen) e a reação

foi incubada a 14ºC por 16 horas. Por fim, o produto foi purificado em colunas *Microcon YM-100 Centrifugal Filter Unit* (Millipore – item 3.7.1).

3.5.7. Reação em cadeia da polimerase

O volume total da reação de RT-PCR foi de 20 µl, contendo 1X tampão; 0,1 mM de dNTP; 1,5 mM de $MgCl_2$; 200 pmoles de oligonucleotídeo XDPN18 (5'CTGATCACTCGAGAGATC3'); 2 unidades da enzima de *GoTaq® DNA Polymerase* (Promega) e 10 µl do cDNA purificado (10% do volume total). O oligonucleotídeo XDPN18 possui complementariedade com a sequência do oligonucleotídeo XDPN14 e o sítio GATC, conforme esquema abaixo. O programa da reação seguiu as seguintes etapas: 4 minutos a 95ºC, seguidos de 40 ciclos de 45 segundos a 95ºC, 1 minuto a 58ºC e 4 minutos a 72ºC, seguido de 6 minutos a 72ºC.

XDPN18: 5'CTGATCACTCGAGAGATC3'

5' CTGATCACTCGAGA**GATC**..... GATCTCTCGAGT 3'
3' ← CTAGAGAGCTCACTAGTC 5'

5' CTGATCACTCGAGAGATC → 3'
3' TGAGCTCTCTAG..... **CTAG**AGAGCTCACTAGTC 5'

3.5.8. Clonagem

3.5.8.1. Ligação ao vetor

A reação de ligação ao vetor de clonagem foi feita com o InsT/Aclone PCR Product Cloning Kit (Fermentas) em volume final de 10 µl, contendo 1X tampão; 0,055 µg do plasmídeo pTZ57R/T; 1 µl PEG 4000; 1,75 unidades T4 DNA ligase e 6,65µl do produto de PCR. A reação foi incubada a 22°C por 16

horas. Por fim, a ligação foi dialisada por 20 minutos em membrana de nitrocelulose de 0,025 µm (Millipore).

3.5.8.2. Transformação

A transformação foi realizada por eletroporação (2,5 KV, 25 µFD, 200 OHMS) de bactérias *E. coli* DH10B com 3 µl da ligação. Os transformantes foram cultivados a 37°C sob agitação de 200 rpm por 40 minutos, e logo em seguida, foram plaqueadas em meio CG (Invitrogen) contendo o antibiótico ampicilina (100 mg/ml) e mantidas por 16 horas a 37ºC para crescimento de colônias individuais.

3.5.8.3. PCR de colônia

A reação de PCR de colônia foi realizada em volume final de 20 µl, contendo 1X tampão; 1,5 mM MgCl2; 0,2 mM dNTP; 2 pmoles do oligonucleotídeo M13F (5'GTAAAACGACGGCCAG3'); 2 pmoles do oligonucleotídeo M13R (5'CAGGAAACAGCTATGAC3') e 0,5 unidade da enzima *Taq Polimerase Phoneutra*. A reação foi incubada a 95°C por 4 minutos, seguidos de 35 ciclos de 95°C por 45 segundos, 60°C por 1 minuto e 72°C por 4 minutos, e por fim, 72°C por 7 minutos.

3.5.9. Sequênciamento da biblioteca

O sequênciamento dos fragmentos das bibliotecas de cDNA enriquecidas para *splicing* alternativo foi realizado pelo método Sanger no aparelho ABI Prism 3130 (Applied Biosystems). A reação foi realizada com *BigDye Terminator v3.1 cycle sequencing kit* (Applied Biosystems), segundo as especificações do fabricante.

Para facilitar a identificação das sequências foi estabelecida uma nomenclatura que contém: 3 letras iniciais que identifiquem o tecido da

biblioteca (BES – *breast*); dois números que indicam a biblioteca (01; 02; em sequência); três números que identificam a placa de sequênciamento (001; 002;...); uma letra e dois números que identificam a posição na placa de sequênciamento (A01; A02; B01...).

3.6. Construção da biblioteca de cDNA para análise de transcriptoma completo

3.6.1. Síntese de cDNA a partir de RNA PoliA$^+$

Vinte microgramas de RNA total tratados com *DNase*I foram utilizados para purificação de RNA PoliA$^+$ utilizando o *mRNA Isolation kit for total RNA* (Miltenyi Biotec), segundo as especificações do fabricante.

Para a síntese de cDNA foram utilizados 200 ng de RNA PoliA$^+$ incubados com oligodT(18) por 10 minutos a 70^{0}C. Em seguida foram adicionados na reação 1X tampão *first strand buffer*, 0,01 M DTT, 1 mM dNTP, 40 unidades de RNasin® Ribonuclease Inhibitor (Promega) e 400 unidades de *Superscript*II (Invitrogen) em volume final de 20 µl. A reação foi incubada a 42^{0}C por duas horas. Em seguida, a síntese da segunda fita de cDNA foi realizada utilizando 1X tampão *second strand buffer*, 1 mM dNTP, 10 unidades *E.coli* DNA ligase (Invitrogen), 40 unidades DNA polimeraseI (Invitrogen), 1,4 unidades de RNase H, em volume total de 100µl. A reação foi incubada a 16°C por 2 horas. Em seguida foram adicionadas 10 unidades T4 DNA polimerase e a reação foi incubada a 16°C por mais 5 minutos. A amostra foi purificada pelo método de fenol:clorofórmio:álcool isoamílico (item 3.7.3).

3.6.2. Clivagem enzimática com *Dpn*II

A reação de digestão do dscDNA com a enzima *Dpn*II foi realizada de forma similar ao descrito no item 3.5.4. Foram utilizadas 15 unidades da enzima na presença de 1X de tampão em volume final de 25 µl. A reação foi incubada a 37ºC por 3 horas. A amostra foi purificada pelo método de fenol:clorofórmio:álcool isoamílico (item 3.7.3).

3.6.3. Ligação de adaptadores em Y

Os adaptadores utilizados na biblioteca para análise de transcriptoma completo formam uma estrutura em Y, como proposto por Watahiki e colaboradores (2004). Os adaptadores foram formados por oligonucleotídeos sintetizados comercialmente que possuem uma região de 11pb de complementariedade, permitindo a hibridização entre eles, formando uma região de fita-dupla e também apresentam uma segunda região de 9pb a qual é única para cada um dos oligos. A região comum dos oligonucleotídeos apresentam 4 bases que servem como identificadoras da amostra de origem, chamadas *tag* (do inglês, etiqueta), que são distintas entre os adaptadores de amostras diferentes. O oligonucleotídeo antisenso apresenta uma extremidade 5' coesiva que contém a sequência de quatro bases complementares ao sítio de clivagem GATC, que as moléculas de cDNA apresentam após digestão com a enzima *Dpn*II. Estes oligonucleotídeos também apresentam na extremidade 5' um grupo fosfato, que possibilita ligação do adaptador à cadeia do cDNA (Figura 14).

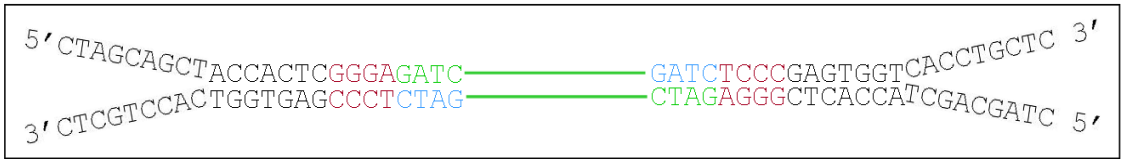


Figura 14: Adaptadores utilizados para construção das bibliotecas para análise de transcriptoma completo. Em verde está representado um fragmento de cDNA clivado pela enzima *Dpn*II. A região em vermelho

corresponde à sequência da *tag* do adaptador e em azul a sequência de quatro bases complementares ao sítio coesivo GATC.

A sequência dos oligonucleotídeos utilizados para formação do adaptador para amostra HB4a foram:

Oligo antisenso: 5' PO4-GATCTCCCGAGTGGTCACCTGCTC 3'

Oligo senso: 5' GAGCAGGTGACCACTCGGGA 3'

A região em vermelho corresponde à sequência da *tag* do adaptador e em azul a sequência de quatro bases complementares ao sítio coesivo GATC.

A sequência dos oligonucleotídeos utilizados para formação do adaptador para amostra C5.2 foram:

Oligo antisenso: 5' PO4-GATCCCCTGAGTGGTCACCTGCTC 3'

Oligo senso: 5' GAGCAGGTGACCACTCAGGG 3'

A região em vermelho corresponde à sequência da *tag* do adaptador e em azul a sequência de quatro bases complementares ao sítio coesivo GATC.

Para formação dos adaptadores 100 pmoles de cada oligonucleotídeo foram misturados, aquecidos a 98ºC por 2 minutos e resfriados lentamente a temperatura ambiente em 1X tampão *Nuclear Extraction Buffer* 2 (New England Biolabs)

3.6.4. Reação em cadeia da polimerase

O volume total da reação de PCR foi de 20 µl, contendo 1X tampão; 0,2 mM de dNTP; 2 mM de $MgCl_2$; 10 pmoles de oligonucleotídeo Primer PCR-16 (5'GAGCAGGTGACCACTC3'); 10 pmoles de oligonucleotídeo Primer PCR-10 (5'CTAGCAGCT3'); 2 unidades da enzima *Platinum Taq DNA Polymerase High-Fidelity* (Invitrogen) e 5% da amostra de cDNA purificada. O programa da reação seguiu as seguintes etapas: 5 minutos a 95ºC, seguidos de 40 ciclos de 40 segundos a 95ºC, 40 segundos a 62ºC e 2 minutos a 68ºC, seguido de 10 minutos a 68ºC.

3.6.5. Validação das bibliotecas por sequênciamento *Sanger*

Previamente ao sequênciamento em larga escala, 100 clones de cada biblioteca foram seqüenciados no equipamento ABI3130 (Applied Biosystems) para validação das bibliotecas. As etapas de clonagem e sequenciamento seguiram o protocolo apresentado acima (itens 5.6.8 e 5.6.9). Para clonagem no vetor T/A foi necessário a adição de resíduos de adenina aos produtos de RT-PCR, que foi realizada em 10 µl de reação contendo 1X tampão, 1,5 mM $MgCl_2$, 0,2 mM de dATP e 2,5 unidades da enzima *Platinum Taq Polimerase*. Esta reação foi incubada a 95ºC por 2 minutos para ativação da enzima e a 72ºC por 30 minutos.

3.6.6. Sequênciamento em larga escala

Após a validação qunatidades equimolares de cDNA das bibliotecas das duas linhagens celulares foram misturadas e um total de 2 µg foram enviados para sequenciamento em larga escala utilizando a plataforma *Genome Sequencer FLX System 454 Roche-Life Sciences*, segundo as especificações do fabricante.

3.7. Métodos de purificação utilizados

3.7.1. Purificação dos fragmentos de cDNA em colunas

Três colunas diferentes de purificação foram utilizadas durante as etapas de construção das bibliotecas de cDNA. Para purificação de fragmentos de RT-PCR, fragmentos isolados de gel de agarose 1% e dos produtos de desnaturação e renaturação foi utilizado o *GFX PCR and Gel Band Purification Kit* (Amersham Biosystems), seguindo as especificações do fabricante. Para purificação de fragmentos de PCR na etapa de validação foi

utilizado o kit *QIAquick PCR Purification* (Qiagen), segundo as recomendações do fabricante. A purificação do dscDNA e do produto ligado aos adaptadores foi realizada em colunas *Microcon YM-100 Centrifugal Filter Unit* (Millipore), seguindo as especificações do fabricante.

3.7.2. Extração orgânica de gel de agarose *low point melting*

A purificação de DNA de agarose *Low Point Melting* foi feita segundo protocolo descrito em Sambrook e Russel (2001).

3.7.3. Purificação pelo método fenol: clorofórmio: álcool isoamílico

A solução de phenol:clorofórmio:álcool isoamílico (25:25:1) pH 8.0 foi adicionada a solução a ser purificada em volume equivalente (1:1). A solução foi centrifugada a 12000 rpm por 10 minutos. O sobrenadante foi transferido a um tubo limpo contendo o volume equivalente de clorofórmio e novamente centrifugado a 12000 rpm por 10 minutos. O sobrenadante foi novamente transferido a um tubo limpo contendo a metade do volume equivalente de acetato de amônio (CH_3COONH_4) e ainda a quantidade de 2 vezes e meia do volume equivalente de isopropanol. A solução foi incubada por 30 minutos a temperatura ambiente seguida de precipitação centrifugada a 12000 rpm a 4ºC por 20 minutos para precipitação. O precipitado foi lavado 3 vezes com etanol 70%.

3.8. Análises bioinformáticas

3.8.1. Análise das bibliotecas de cDNA enriquecidas para *splicing* alternativo

As sequências das bibliotecas de cDNA enriquecidas para *splicing* alternativo foram analisadas por ferramentas de bioinformática com auxílio do Laboratório de Biologia Computacional do Instituto Ludwig de Pesquisa sobre o Câncer, coordenado pelo Dr. Sandro José de Souza. Inicialmente as sequências de vetor foram identificadas e trimadas, e as sequências com menos de 80% das bases referentes ao vetor foram analisadas. Em seguida, a qualidade das sequências foi analisada em janelas de 20 pb, sendo que apenas as janelas com no mínimo 15 pb com valor de Phred ≥ 20 (EWING et al., 1998; EWING; GREEN, 1998) foram consideradas e as demais regiões foram descartadas. As sequências foram então agrupadas por similaridade com auxílio do programa CAP3, gerando sequências consensuais (consenso) entre grupos de sequências contíguas (contigs) ou sequências únicas (singlets), isto é, sem similaridade com nenhuma outra sequência.

As sequências consenso foram alinhadas contra a sequência do genoma humano (NCBI #36.1) disponível pela ferramenta BLAT (do inglês, *Blast-like alignment tool*) (KENT, 2002) utilizando um critério de alinhamento de ≥ 93% de identidade e ≥ 55% de cobertura. Apenas as sequências com alinhamento único no genoma foram selecionadas e novamente alinhadas utilizando a ferramenta Sim4 (FLOREA et al., 1998), que leva em consideração a presença de sítios de *splice* conservados nos íntrons aumentando a confiabilidade do alinhamento. A etapa seguinte foi a clusterização ou agrupamento, dos consensos com sequências de mRNAs dos bancos de dados Genbank (244.284 sequências) (BENSON et al., 2008), RefSeq (26.040 sequências) (PRUITT et al, 2007) e dbEST (8.133.299 sequências) (BOGUSKI; LOWE; TOLSTOSHEV, 1993) disponíveis pela Universidade de Santa Cruz, Califórnia (UCSC – setembro de 2007). Por fim, foi realizada a busca de variantes de *splicing* através de comparação par a

par das coordenadas genômicas de bordas entre éxons e íntrons de todas as sequências de um mesmo cluster para detecção de limites éxon-éxon com alinhamento diferente. O alinhamento de todas as sequências de cada *cluster* contra o genoma humano foi checado visualmente com auxílio da ferramenta *Genome Express Browser System* (*GEBrowser*) (STEIN et al., 2002).

3.8.2. Análise das bibliotecas de cDNA para análise de transcriptoma completo

As sequências geradas pelo sequenciamento das bibliotecas para análise de transcriptoma completo foram analisadas pelo Laboratório de Biotecnologia do Hospital A.C. Camargo. Inicialmente foi verificada a presença de adaptadores, sendo que sequências sem adaptadores ou com adaptadores internos foram descartadas. Em seguida, sequências com alta similaridade (E-value ≤ $1x10^{-20}$, identidade ≥ 85% e cobertura ≥ 85%) a RNA ribossomal ou sequências mitocondriais foram filtradas utilizando a ferramenta Mega BLAST (ZHANG et al., 2000). Após o filtro, as sequências foram alinhadas contra o genoma humano (release hg18, Março 2006) com auxílio da ferramenta BLAT (KENT, 2002) e os melhores alinhamentos foram selecionados pelo programa de filtragem pslReps disponibilizado pela Universidade de Santa Cruz, Califórnia (UCSC) seguindo os parâmetros de cobertura mínima de 70%, alinhamento mínimo de 96% e nearTop = 0.005. Sequências com alinhamento significativo em múltiplas regiões genômicas foram descartadas. Para anotação gênica, as sequências foram mapeadas contra transcritos presentes no banco de dados KnownGene (HSU et al., 2006) disponível pela Universidade de Santa Cruz, Califórnia (UCSC). O alinhamento das sequências com qualquer nucleotídeo de uma unidade de transcrição, definida como a sequência genômica completa entre o primeiro e a último nucleotídeo do transcrito, foi considerado válido. Para identificação de variantes de *splicing,* as coordenadas de alinhamento genômico dos limites éxon-íntron das sequências foram comparadas com as coordenadas genômicas dos limites éxon-ntron dos transcritos correspondentes presentes

no banco de dados KnownGene. Todas as variantes foram checadas quanto a presença de sítios de *splice* confiáveis pela identificação de sítios de *splice* conservados (5'GT – AG3') nos intervalos de alinhamento das sequências contra o genoma humano maiores que 30 nucleotídeos. As novas variantes foram classificadas em três classes principais: uso alternativo de éxons (incluindo tanto inclusão quanto exclusão de um éxon), retenção de íntrons e sítio de *splice* doador e aceptor alternativos.

Para comparação entre o número de eventos identificados para cada amostra, foram utilizados critérios mais estringentes para correta identificação das sequências em relação à amostra de origem. Para isso, além da identificação das regiões correspondentes aos adaptadores, foram identificadas as regiões correspondentes as *tags* e ao sítio de clivagem enzimático. Os eventos identificados neste subgrupo de sequências foram normalizados pelo total de eventos identificados para cada amostra. O número de eventos identificados para cada categoria de eventos de *splicing* para cada amostra foram comparadas pelo teste estatístico do qui-quadrado.

3.9. Validação por RT-PCR

Para validação por RT-PCR foram desenhados iniciadores baseados nas sequências obtidas, sendo complementares as regiões mais extremas das ASSETs originadas pelas bibliotecas enriquecidas para *splicing* alternativo, e complementares ao novo éxon e a um éxon adjacente, no caso das sequências provenientes das bibliotecas para análise do transcriptoma completo. A sequências dos iniciadores utilizados estão descritas na tabela 4. O volume total da reação de PCR foi de 20 µl, contendo 1X tampão; 0,2 mM de dNTP; 2 mM de $MgCl_2$; 10 pmoles do iniciador *foward*; 10 pmoles do iniciador *reverse*; 2 unidades de *PlatinumTaq DNA Polimerase* (Invitrogen) e cDNA da linhagem C5.2 ou HB4a ou cDNA das amostras tumorais tratado com DNAse (item 3.4). A reação seguiu as seguintes etapas: 5 minutos a 95ºC, 40 ciclos de 40 segundos a 95ºC, 40 segundos na temperatura ideal de

anelamento dos iniciadores e 2 minutos a 72ºC, seguido de 10 minutos a 72ºC. Os produtos da reação foram visualizados em gel de acrilamida 8% corados com nitrato de prata.

Tabela 4: Sequência dos iniciadores utilizados na validação das variantes de *splicing* por RT-PCR. O gene está representado pelo símbolo oficial e a sequência do iniciador *foward* e iniciador *reverse* estão apresentadas no sentido 5'a 3'.

Gene	Primer Foward	Primer Reverse
COL7A1	GGTGCTCCTGGTGTTG	GATCACAGCTGAGTCTC
GSTP1	GCCAGGAAAGGAGAG	GTTCGGCAAATTATCCAG
ITGB5	GGCACTGTCTCTGTGG	GTTGTCAGGTTTCCCAGAG
FLNA	GGGATGCCAGTCGTGTTC	CAAAGCGGATGCAGTAGGTG
RBM10	CGATGCACTACAGTGACCC	CCCAGTGGCAGTGTCTG
AOF2	GAGTTCACTGGCAGCC	GTCGTACTGCCAACATG
FN1	GGGAGTGGTGGTTAC	CAGGAGGAAATAGCCC
TRIP6	GGCTGTTACAAGTGCG	GGGAACTGGAACTGAG
EIF4A3	CAAAGGGAGAGATGTC	GTGGGAGCCAAGATC
SFRS9	GGTCCTGCCAGCTGC	CCCTTCGCCTTCGTGC
ATP1A1	GTGTACTACGCAGAC	CTGGATGAGGAGCTG
CDK5RAP2	GCCTCTCTTGTTGCTTC	CAGCATTTGCAGCAGAG
PTPLA	CAATCACCTCTCCATGAAG	TCCGCTATTCCTTCTACAC
ALDH3A2	CCAATTCTTCCAATAGTGC	GAACAGGATCATTCTTCAG
RPS2	CGATGGAGAGCTTGGC	CAAGGATGAGGTTTTGAAG
KRT18	CATGGCAGACATCCGG	GGACTGGACTGTACG
CLCT	GTTGGAAGCAGCACTGG	GAGACAGCACCATCAGC
PSMC2	GATCTGAGCTTGTACAG	CATTGTTCTCTGCACTTC
AET1G	CAGGTGTGTGCGAACAG	CTTCAGCTCAATCCCAATC
APEX1	GTTTGTCATTCCCTTGATG	CTCCTGCTGCCTCTTTGTC
BC039445	GTCGACCTCGCAACAG	CCTATGAAATAGTCTCGGC
CLTC	CCTAGAAACTGCATGGAG	CTTTTCCTTTATTGCATCAAC
CSRP2BP	GATTCATCCTGTTTTGCTTCTG	CCTTTGGCTTCATGGTTCC
CUEDC2	GGATTACAGGCATGAACC	GATGAGAGCTGCACCG
FLJ00150	CTCAGAAAGGGATAGTAGC	GTAGCCCAGGACAACCATG
FTJ3	GTCTGCTGCACTCATATCC	CCTGGCATCAAGCAATCC
KIAA1033	CATTCGTACATAGCCCATAGC	CAGTAAACAAAGTCTCTGTCC
KIF2A	GCAACAGCAAGAACTTAGAG	CTTCTAGGAAATAATACCACC
NR_002599	GCGAAGAGCCGTTAGTC	GAAAACAGAATTCAAGCTACTG
NR2C1	CAAGTGCTGTCACAATCTG	GTGGCAATAGAATCGGTAC
PAWR	CCACCTAGAACAGTTTCAG	CATTCTCTTCACCCTCCAAC
PDE6D	CCATGTGCCAAGTGAGTG	CCTAACTCCACAAATACCTG
PP2R2A	CTTTCAAGTTATACCCTTCTGG	GTATAGTGGAGAAGCCTGG
PRCC	CACCTAGTAGCTGAGAACAG	GTTGGCTGCTCACCTTTC
RPLP1	CATGGCCTCTGTCTCC	GGCAATTACACCGAAAGAG
RPS19	GTTCATCTTTCAGTCCTCAG	GCTTGCTCCCTACGATG
RWDD1	GAAAGCCAAGTTTGATG	CTCTTTTTCTGTGAATTTC
ZNF567	GCTCAGAAGACTCTATATATGG	CTGGGTAAGTGAAGACAC

3.10. Eletroforese em *chip*

As amostras de RT-PCR foram analisadas por eletroforese em *chip* sem a necessidade de nenhuma purificação prévia, utilizando o equipamento Lab*Chip*® GX (Caliper Lifesciences), segundo as recomendações do fabricante. A expressão do gene *GAPDH* também foi analisada nas duas linhagens e o valor obtido para este gene foi utilizado como fator de normalização. A diferença de expressão foi calculada pela razão dos valores de concentração normalizados obtidos para C5.2 e HB4a. Os transcritos com razão de expressão ≥|2| foram considerados como diferencialmente expressos.

3.11. Validação por sonda-específica

Para validação foram desenhos dois pares de sonda para cada gene de interesse, sendo cada par específico para uma das variantes de *splicing*. As sondas esquerda e direita foram posicionadas de forma adjacente em relação ao transcrito, sendo que o limite entre elas corresponde a uma junção de éxons específica da variante alvo. A sonda esquerda contém em sua extremidade 5’ uma sequência de reconhecimento para o iniciador *foward* (GGGTAGGCTAAGGGTAGGA), uma sequência de composta por 38 nucleotídeos (CCGTTGCCAGTCTGCTCAGACCTCCCTCGCGCCATCAG) de preenchimento (*stuffer sequence*) seguida de uma região complementar a sequência alvo. A sonda direita foi fosforilada em sua extremidade 5’, sendo esta região complementar ao alvo seguida de uma sequência de reconhecimento para o iniciador *reverse* na região mais a 3’ (TCTAGATTGGATCTTGCTGGCAC). Além disso, foram desenhados oligonucleotídeos para a síntese de cDNA específica para cada gene avaliado, sendo este oligo posicionado na região 3’ do gene para permitir a síntese de todas variantes de interesse. A sequência das sondas e dos oligos

usados na síntese de cDNA estão descritas na tabela 5. Para síntese de cDNA foi utilizado 1 ug RNA total das linhagens HB4a e C5.2 tratado com DNAse, 0,2 ug oligonucleotídeo específico, 0,01 M DTT, 1 mM dNTP, 40 unidades de RNasin® Ribonuclease Inhibitor (Promega) e 400 unidades de *Superscript*II (Invitrogen) em volume final de 20 µl. Inicialmente, 250 ng do cDNA gene específico foi aquecido a 98°C por 2 minutos e colocado no gelo. Em seguida, foi adicionado ao cDNA as sondas específicas e 1X tampão de hibridação MLPA (MRC-Holland) e incubados a 60°C por 16 horas. Para ligação foram adicionados 1X tampão ligase A, 1X tampão ligase B e 1ul Ligase 65 (MRC-Holland) e incubados a 54ºC por 15 minutos e 98ºC por 5 minutos. Como controle negativo, foram realizadas hibridações e ligações na ausência de cDNA molde para todos os pares de sonda. As sondas unidas foram amplificadas com a enzima *PlatinumTaq DNA Polimerase* (Invitrogen), segundo as recomendações do fabricante. O produto de RT-PCR foi analisado em gel de acrilamida 8% corado com nitrato de prata.

Tabela 5: Sequência dos oligonucleotídeos para validação baseada na ligação de sondas específicas. O gene está representado pelo símbolo oficial e a sequência dos oligonucleotídeos estão apresentadas no sentido 5'a 3'. As sondas utilizadas para análise da variante ASSET estão identificadas com a letra A e as sondas utilizadas para análise da variante heteroduplex estão identificadas com a letra H.

Gene	Oligonucleotídeo/Sonda	Sequência
SFRS9	Oligo síntese de cDNA	GGTCCTGCCAGCTGC
	Sonda esquerda-A	GGGTAGGCTAAGGGTAGGA GAGGATGCTATTTATGGAAGAAATG
	Sonda esquerda-H	GGGTAGGCTAAGGGTAGGA GATTTCCGAGTTCTTGTTTCAG
	Sonda direita-A/H	GACTTCCTCCGTCAGGC TCTAGATTGGATCTTGCTGGCAC
FLNA	Oligo síntese de cDNA	CAAAGCGGATGCAGTAGGTG
	Sonda esquerda-A/H	GGGTAGGCTAAGGGTAGGA CCGACCAGCACGTGCCTG
	Sonda direita-A	AAATTAGCATCCAGGATATGACAG TCTAGATTGGATCTTGCTGGCAC
	Sonda direitaH	GCAGCCCCTTCTCTGTG TCTAGATTGGATCTTGCTGGCAC
TRIP6	Oligo síntese de cDNA	GGGAACTGGAACTG
	Sonda esquerda-A	GGGTAGGCTAAGGGTAGGA TTGCTTCTTTTTCAACAG
	Sonda esquerda-H	GGGTAGGCTAAGGGTAGGA CATTGGCTGTTACAAGTGCGAG
	Sonda direita-A/H	GAGTGTGGGCTGCTGCTC TCTAGATTGGATCTTGCTGGCAC

3.12. Análise da sequência aberta de leitura e domínios proteicos das variantes de *splicing*

A putativa sequência aberta de leitura das novas variantes de *splicing* identificadas foi predita com a ferramenta ORFfinder (*open reading frame finder*) disponível pelo *National Center for Biotechnology Information* (NCBI).

Os domínios proteicos da variante conhecida e das novas variantes foram preditos com auxílio da ferramenta InterProScan (QUEVILLON et al., 2005) que prediz a ocorrência de domínios funcionais, repetições e sítios importantes presentes no banco de dados InterPro (HUNTER et al., 2009).

3.13. Anotação funcional das variantes de *splicing*

As categorias funcionais dos genes correspondentes as variantes de *splicing* identificadas foram anotadas de acordo com os processos biológicos das categorias de ontologia gênica e de acordo com as vias de sinalização do banco de dados Kegg. Para identificar as categorias enriquecidas em genes que reportam variantes de *splicing* foi utilizada a ferramenta BinGO (MAERE; HEYMANS; KUIPER, 2005).

4. Resultados

Neste trabalho foram utilizadas duas abordagens distintas para a identificação de variantes de *splicing* em câncer de mama sob a influência da expressão do gene *ERBB2*. A primeira estratégia foi o estabelecimento da metodologia para construção de bibliotecas de cDNA enriquecidas para *splicing* alternativo utilizando fontes de RNA de amostras que apresentam expressão aumentada desse oncogene. Os dados referentes a esta parte do trabalho foram organizados em um manuscrito, aceito para publicação na revista BMC Genomics (Anexo A).

A segunda estratégia utilizada para identificar transcritos variantes de *splicing* em câncer de mama foi a exploração de dados gerados pelo sequenciamento global do transcriptoma de duas linhagens celulares de mama, uma com expressão basal e outra com expressão aumentada do gene *ERBB2*. Cabe esclarecer que esse projeto de sequenciamento global foi um estudo desenvolvido pelo nosso grupo com a colaboração de outros grupos de nossa instituição e também do *M. D. Anderson Cancer Center* e que teve como objetivo avaliar os aspectos quantitativos e qualitativos do transcriptoma influenciado pela alta expressão do gene *ERBB2*. Nessa tese foram apresentados somente os dados referentes à variação transcricional gerada por *splicing* alternativo e moduladas pela expressão aumentada do gene *ERBB2*. Os dados completos gerados pelo projeto foram organizados em um manuscrito, e submetidos para o jornal *PLoS Genetics* (Anexo B).

4.1 Bibliotecas de cDNA enriquecidas para *splicing* alternativo

4.1.1 Estabelecimento da metodologia de construção de bibliotecas de cDNA enriquecidas para *splicing* alternativo

Foi utilizado cDNA dupla-fita convertido de RNA amplificado para construção da biblioteca de cDNA enriquecida para *splicing* alternativo baseada no protocolo proposto por Watahiki e colaboradores (2004). Cabe ressaltar que o uso de RNA amplificado além de possibilitar a utilização de pequena quantidade de RNA total inicial, tornando nosso protocolo mais simples, rápido e de menor custo, não necessita da utilização de bibliotecas de cDNA parentais, como proposto por Watahiki e colaboradores (2004). Um esquema geral da metodologia estabelecida está apresentado na figura 15.

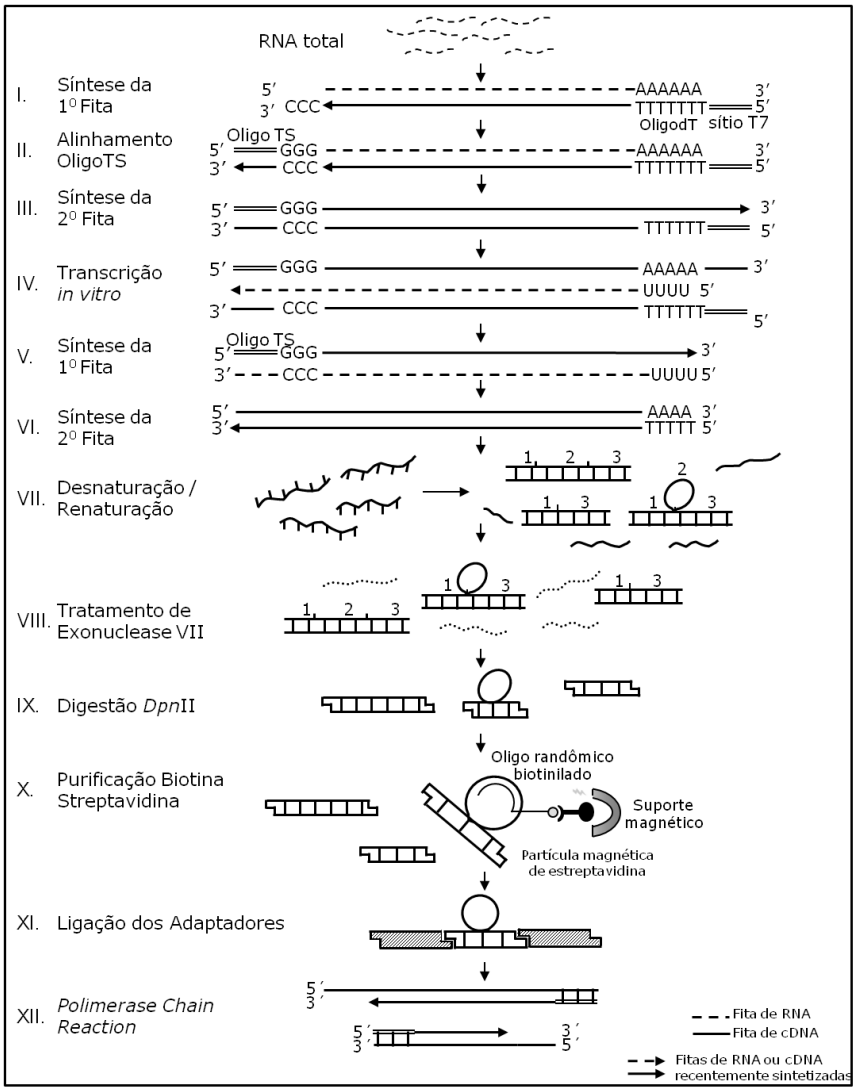


Figura 15: Esquema geral da construção da biblioteca de cDNA enriquecida para *splicing* alternativo. I – OligodT com sítio de reconhecimento da enzima T7 RNA polimerase foi usado para síntese da primeria fita de cDNA com a enzima *Superscript*II que adiciona resíduos de citosina na extremidade 5' do cDNA. II – A região rica em citosina serviu de âncora para anelamento do oligo TS, permitindo continuidade da síntese da primeira fita. III – Síntese da segunda fita de cDNA utilizando o oligo TS. IV – Transcrição do RNA antisenso pela enzima T7 RNA polimerase. V – Síntese da primeira fita do cDNA usando o oligo TS. VI – Síntese da segunda fita do cDNA usando o oligo dT. Linhas contínuas representam fitas de cDNA e linhas tracejadas representam fitas de RNA. Os traços duplos representam as regiões dos oligonucleotídeos oligodT-T7 e oligo TS. As flechas representam as fitas recém-sintetizadas.VII – Denaturação e renaturação e a formação de heteroduplex. Os números 1, 2 e 3 representam éxons de um gene hipotético. VIII – Tratamento com exonuclease VII. As linhas pontilhadas representam fitas de cDNA degradadas. IX – Digestão com *Dpn*II formando extremidades coesivas no cDNA. X – Anelamento do oligo randômico biotinilado a região de alça de simples-fita do heteroduplex e filtragem pelas partículas magnéticas de estreptavidina. XI – Ligação de adaptadores. XII – Amplificação dos fragmentos pela reação de RT-PCR.

De forma resumida, 18 µg de dscDNA foram aquecidos para desnaturação da dupla-fita e novamente hibridizados em condições favoráveis, permitindo a formação de estruturas de heteroduplexes entre duas variantes de *splicing* de um mesmo gene (Figura 15 - VII). As moléculas de heteroduplexes apresentam uma região de dupla-fita formada pelo anelamente entre regiões complementares das variantes e uma região de alça de simples-fita correspondente a uma região única de uma das variantes. Em seguida, os fragmentos de simples-fita de cDNA que possuem extremidades livres foram degradados pela enzima Exonuclease VII (Figura 15 - VIII). Posteriormente, a amostra foi clivada com a enzima de restrição *Dpn*II que reconhece, no cDNA de fita-dupla, um sítio de quatro nucleotídeos GATC (Figura 15 - IX), com extremidades coesivas de sequência conhecida, que permitiu a utilização de adaptadores específicos. Foi utilizado como controle independente da clivagem um fragmento de 459pb que possui um único sítio de restrição, o qual é clivado em dois fragmentos menores de 127pb e 332pb (Figura 16).

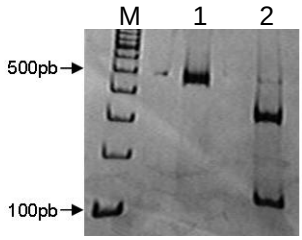


Figura 16: Clivagem do fragmento controle com a enzima de restrição *DpnII*. M – marcador 100pb. 1 – Fragmento antes da clivagem (459pb). 2 – Fragmentos gerados após clivagem (127pb e 332pb). Gel de acrilamida 8%.

Para enriquecimento das moléculas de heteroduplexes, a amostra foi incubada com oligonucleotídeo de 25 nucleotídeos randômicos, contendo na extremidade 5’ uma molécula de biotina. O fato de este oligonucleotídeo possuir sequência randômica permite ligação por complementariedade a sequências variadas de fita simples presentes nas alças das estruturas de heteroduplexes. Através da ligação entre as moléculas de biotina com partículas magnéticas de estreptavidina (Figura 15 - X) as moléculas de cDNA com estrutura de heteroduplexes foram isoladas das demais moléculas presentes na amostra. A recuperação das moléculas de heteroduplex foi feita

na presença de solução desnaturante. Foi feita então a ligação de adaptadores específicos (XDPN12 e XDPN14) nas extremidades livres coesivas dos heteroduplexes de dscDNA (Figura 15 – XI). Em seguida foi realizada amplificação por reação em cadeia da polimerase utilizando o oligonucleotídeo XDPN18 (item 3.5.7).

Uma vez que diversos hetroduplexes correspondendo a variantes de diferentes genes estão presentes na amostra é esperado que moléculas de tamanhos distintos sejam amplificadas, formando um *smear* no gel de agarose. Como controle positivo da reação foi utilizado um fragmento previamente clivado com a enzima *Dpn*II e ligado aos mesmos adaptadores que a amostra de dscDNA (Figura 17).

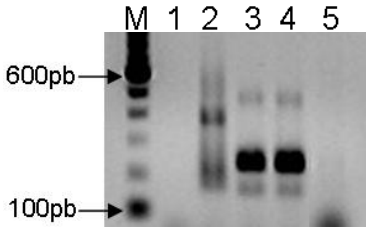


Figura 17: Reação de RT-PCR da amostra C5.2. M – marcador 100pb; 1 – 25 ciclos de amplificação; 2 – 40 ciclos de amplificação; 3 – controle positivo, 25 ciclos de amplificação; 4 – controle positivo, 40 ciclos de amplificação; 5 – controle negativo, sem molde. Gel de agarose 1%.

A reação de 25 ciclos de RT-PCR não produziu amplificação da amostra que pudesse ser visualizada no gel. Apenas na condição de 40 ciclos foi observado amplificação dos fragmentos. Dessa forma, o restante desta reação foi aplicado em gel de agarose 1% para seleção de fragmentos de tamanho entre 250pb e 600pb. Na tentativa de diminuir a redundância da biblioteca, uma vez que a abordagem não é quantitativa, a região de 500pb foi retirada. O produto foi purificado e clonado, sendo esta biblioteca nomeada BES01.

As colônias de bactérias foram utilizadas como molde da reação de PCR de colônia (Figura 18). A amplificação apresentou fragmentos de tamanhos variados os quais foram seqüenciados.

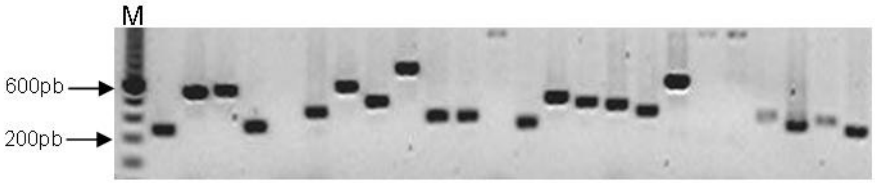


Figura 18: Reação de PCR de colônia da biblioteca BES01. M – marcador 100pb

Inicialmente foram seqüenciados aproximadamente 200 clones para análise de qualidade e validação da biblioteca. As sequências foram analisadas quanto a presença das regiões correspondentes aos adaptadores e ao sítio de clivagem enzimático, e alinhadas contra a sequência genômica e bancos de dados de sequências expressas (mRNAs e ESTs), com auxílio da ferramenta BLAT. A observação de variantes de *splicing* conhecidas e de evidências de novas variantes de *splicing* alternativo neste grupo de 200 sequências foi fundamental para classificar a biblioteca como validada. Assim, a biblioteca BES01 foi submetida a sequênciamento de aproximadamente 1000 clones.

4.1.2 Biblioteca enriquecida para variantes de *splicing* a partir de amostras tumorais de mama

Com o estabelecimento da metodologia de construção de bibliotecas de cDNA enriquecidas para *splicing* alternativo, prosseguimos com a construção de uma biblioteca a partir de amostras tumorais de mama, com características semelhantes à linhagem C5.2, no que se refere ao nível de expressão do gene *ERBB2*. Foram selecionadas cinco amostras de carcinoma ductal invasivo (CDI) que apresentam aumento da expressão da proteína ERBB2, conforme classificação por imunohistoquímica. Quantidades equivalentes de RNA total das cinco amostras foram misturadas e utilizadas como uma amostra única contendo 5 µg de RNA total para a construção da biblioteca de cDNA enriquecida para *splicing* alternativo.

As etapas para confecção da biblioteca enriquecida para *splicing* alternativo foram as mesmas discutidas anteriormente e estão representadas

na figura 15. A amostra enriquecida foi amplificada por 40 ciclos de reação de RT-PCR (Figura 19).

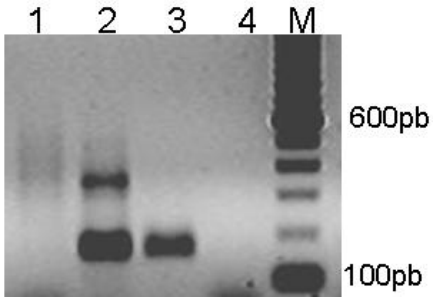


Figura 19: Reação de RT-PCR do grupo de amostras de CDI. M – marcador 100pb; 1 – 40 ciclos de amplificação; 2 e 3 – controles positivos; 4 – controle negativo. Gel de agarose 1%.

O volume restante da reação de RT-PCR foi aplicado em gel de agarose 1% e purificado. Os fragmentos foram clonados e amplificados em reação de PCR de colônia (Figura 20). Esta biblioteca foi denominada BES02.

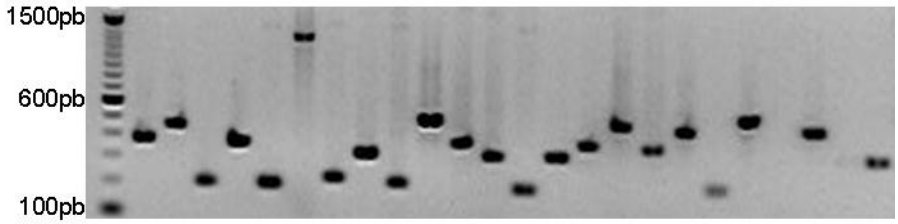


Figura 20: Reação de PCR de colônia da biblioteca BES02. M – marcador 100pb. Gel de agarose 1%.

De forma semelhante, a validação da biblioteca foi obtida pelo sequenciamento inicial de aproximadamente 200 clones. Após validação, aproximadamente 1000 clones foram sequenciados.

4.1.3 Análise das sequências das bibliotecas BES01 e BES02

Todas as sequências geradas das bibliotecas foram inicialmente analisadas quanto a qualidade e presença de sequência correspondente ao vetor de clonagem. No total foram geradas 2.048 sequências de boa qualidade das bibliotecas de cDNA enriquecidas para *splicing* alternativo,

sendo 946 sequências provenientes da biblioteca da linhagem celular C5.2 (BES01) e 1.102 sequências oriundas da biblioteca de amostras tumorais (BES02) (Tabela 6). Em seguida, as sequências foram agrupadas por similaridade, gerando sequências consensuais (consenso) entre grupos de sequências contíguas (*contigs*) ou sequências únicas (*singlets*), isto é, sem similaridade com nenhuma outra sequência gerada na biblioteca. A redundância média considerando as duas bibliotecas foi de 85,5% (Tabela 6) e o tamanho médio das sequências consenso obtidas foi de 256 nucleotídeos.

Tabela 6: Análise das sequências das bibliotecas BES01 e BES02. Número de sequências geradas no total, número de sequências consensuais, *contigs* e *singlets*, e a redundância calculada para cada biblioteca.

Biblioteca	Número de sequências	Número de *Contigs*	Número de *Singlets*	Número de consensos	Redundância
BES01	946	96	53	149	84,25%
BES02	1102	74	72	146	86,75%
Total	2048	167	125	295	-

Após a clusterização, as sequências consenso foram analisadas manualmente. Seis consensos, dos 289, foram descartados de análises posteriores por apresentarem mais de um adaptador na região inicial da sequência, o que pode ser um artefato gerado durante a construção da biblioteca, resultando em 289 consensos submetidos para as análises posteriores. As sequências consenso foram denominadas ASSETs, *alternative spliced sequence-enriched tag,* em acordo com publicações anteriores (WATAHIKI, et al., 2004; THILL et al., 2006).

As 289 ASSETs foram alinhadas contra a sequência genômica humana pela ferramenta BLAT (Universidade de Santa Cruz, Califórnia - UCSC) e pela ferramenta Sim4 (FLOREA et al., 1998), que promove um alinhamento mais confiável das bordas dos transcritos por levar em consideração os sítios de *splice* constitutivos, sendo especialmente importante em análises de busca por variantes de *splicing* (KIM; SHIN; LEE, 2004). Após esta análise, 214 ASSETs foram selecionadas (151 *contigs* e 63 *singlets*). Este alinhamento permite definir a estrutura das sequências consenso, sendo que 93 são formadas por múltiplos éxons e 121 são

sequências *intronless*, isto é, sequências formadas por um único éxon que alinham continuamente na sequência genômica. A etapa seguinte da análise bioinformática foi a clusterização, ou agrupamento, das sequências consenso com sequências de RNAm (banco de dados do Genbank e RefSeq) e ESTs (banco de dados dbEST), resultando em 164 *clusters* ou grupos. Destes, 83 apresentaram apenas sequências da biblioteca BES01, 65 apresentam apenas sequências da biblioteca BES02 e 16 apresentam sequências de ambas bibliotecas, indicando uma sobreposição de aproximadamente 10% das sequências geradas para cada biblioteca. Dos 164 *clusters* 142 contêm ao menos uma sequência de RNAm com anotação (sequência do banco de dados RefSeq). As análises bioinformáticsa estão resumidas na figura 21. As 79 ASSETs, união das ASSETs com múltiplos éxons das duas bibliotecas, foram utilizadas para busca por eventos de *splicing* alternativo.

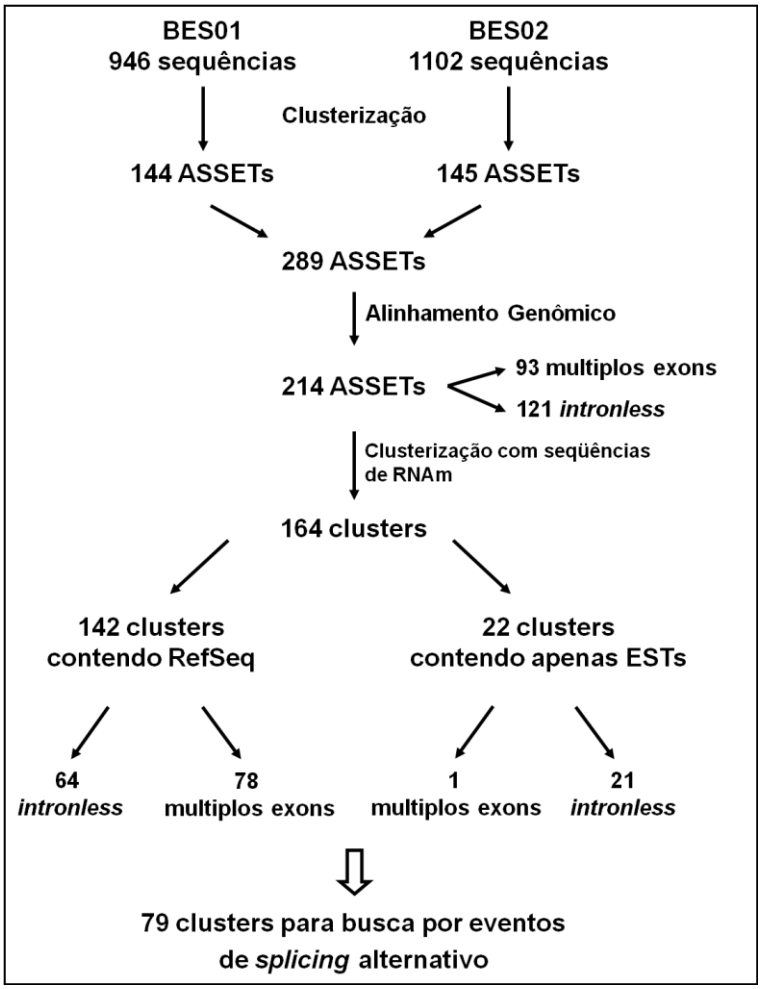


Figura 21: Fluxograma das análises bioinformáticas das bibliotecas enriquecidas para variante de *splicing*.

O agrupamento das ASSETs com sequências de RNAm dos bancos de dados permite além da identificação de variantes de *splicing*, analisar a posição relativa das ASSETs ao longo dos transcritos RefSeq. Esta análise de cobertura por transcrito é muito informativa a respeito da eficiência da estratégia de amplificação do RNAm para construção das bibliotecas de cDNA. A distribuição da localização das ASSETs em relação aos transcritos RefSeq de cada cluster resultou em uma curva de distribuição normal, indicando que não houve representação preferencial de nenhuma porção dos transcritos (Figura 22). Além disso as porções 5' e 3' dos transcritos foram representadas de forma similar, indicando ausência de viés para porção 3'. Estes resultados sugerem que a metodologia de amplificação de RNAm baseada no oligonucleotídeo TS e transcrição *in vitro*, adaptada nesse estudo, foi apropriada para geração de transcritos completos e identificação de eventos de *splicing* em toda extensão do transcrito.

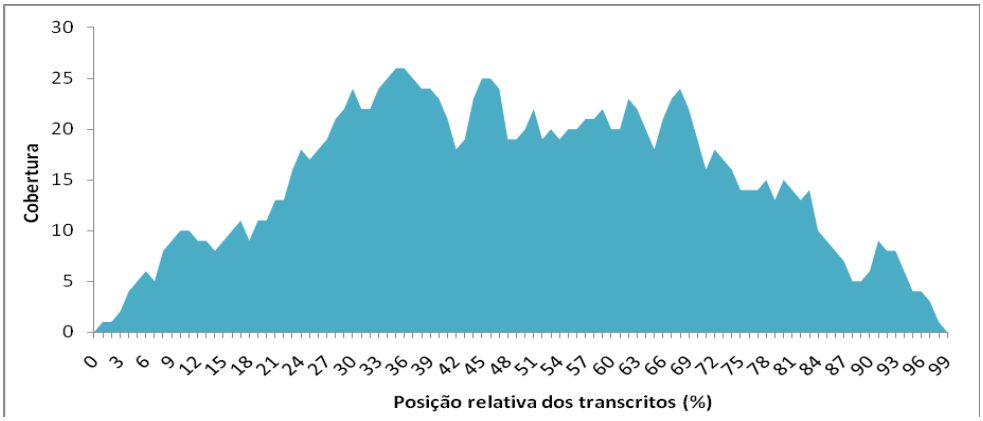


Figura 22: Distribuição relativa das ASSETs em relação aos transcritos RefSeq. A posição relativa dos transcritos é dada em forma proporcional, sendo que 1% corresponde a extremidade 5' e 100% corresponde a extremidade 3'.

4.1.4 Identificação de variantes de *splicing* alternativo utilizando bibliotecas de cDNA enriquecidas para *splicing* alternativo

Uma vez que durante a etapa de clusterização das sequências de cada biblioteca não foram encontrados pares de sequências representativas de variantes de *splicing*, a identificação da suposta outra variante de *splicing*,

envolvida na formação do heteroduplex, foi realizada através da comparação das ASSETs com sequências de cDNA dos bancos de dados.

Os 79 *clusters* contendo ASSETs com múltiplos éxons foram analisados para busca dos pares de variantes de *splicing* através de comparações entre as coordenadas genômicas de cada ASSET e as coordenadas genômicas de todas as sequências referência (RNAm e EST dos bancos de dados públicos) do mesmo *cluster,* para detecção de bordas entre éxons e íntrons com coordenadas de alinhamento diferentes (Figura 23). Para esta análise, as sequências comuns entre as duas bibliotecas (10 sequências representado 5 genes) foram agrupadas e a sequência consenso resultante foi analisada. Assim, entre os 79 ASSETs analisadas, 75 representam genes distintos e foram utilizadas para busca por *splicing* alternativo.

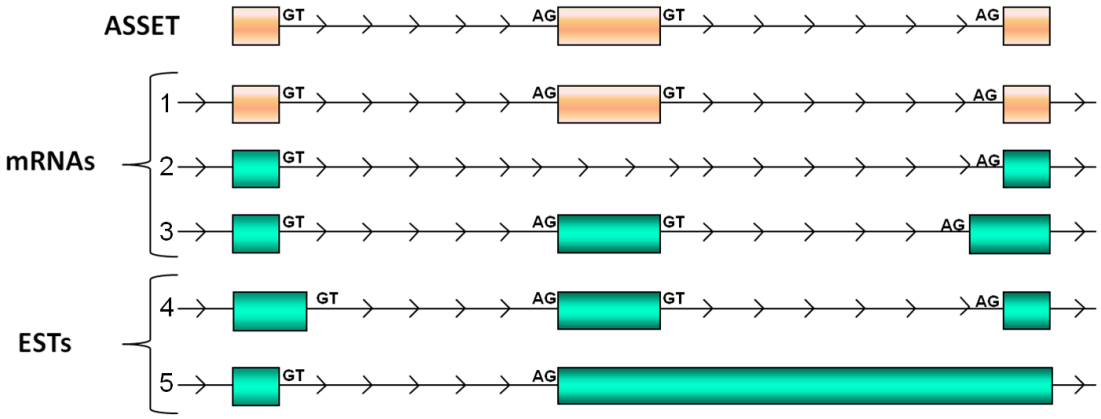


Figura 23: Estratégia para identificação de variantes de *splicing*. Através da comparação das coordenadas genômicas dos limites éxon/íntron entre as ASSETs e as sequências de cDNA dos bancos de dados foram identificados os pares de variantes de *splicing*. Os retângulos representam os éxons e as linhas representam os íntrons. As flechas nos íntrons representam o sentido da transcrição do gene em relação ao genoma. Em laranja claro estão representado os éxons da ASSET e de uma sequência de RNAm que reporta o mesmo alinhamento que a sequência ASSET (1). Em verde estão representados os éxons de sequências consideradas variantes de *splicing* alternativo em relação a ASSET. A variante 2 reporta um evento do tipo uso alternativo do éxon, com a exclusão de um éxon. A variante 3 reporta no último íntron um sítio aceptor de *splice* alternativo, enquanto a variante 4 reporta no primeiro íntron o uso de um sítio doador de *splice* alternativo. A variante 5 reporta a retenção do último íntron.

Todas as 75 ASSETs representam transcritos conhecidos, uma vez que ao menos uma sequência correspondente com as mesmas coordenadas genômicas foi encontrada nos bancos de dados. Entre as 75 ASSETs, 39 (52%) apresentaram transcritos alternativos nos bancos de dados que possam ter participado da formação do heteroduplex; e 36 ASSETs não apresentaram nenhum transcrito alternativo sendo apenas anotado o gene correspondente. Para as 39 ASSETs que apresentaram um transcrito alternativo no banco de dados foi também anotado o número de variantes distintas no banco de dados que possa ter participado da formação do heteroduplex e o número e tipo de eventos de *splicing* reportados (Tabela 7). Das 39 ASSETs para as quais foram identificadas transcritos variantes nos bancos de dados, 22 foram identificadas na biblioteca BES01, 12 foram identificadas na biblioteca BES02 e 5 ASSETs são comuns entre as duas bibliotecas (Tabela 7).

Tabela 7: Caracterização do número e tipo de evento de *splicing* alternativo identificado para 39 ASSETs, representadas pelo símbolo dos genes correspondentes.

Biblioteca	Gene	Retenção de Íntron	Éxon Alternativo	Sítio de *splice* 3'	Sítio de *splice* 5'	Total de eventos
BES01	*FLNA*		1	1	2	4
	NAP1L1		3	2	2	7
	CAMK2G		1			1
	GNPTAB		1	1		2
	RANBP1		1			1
	RPL6	2		1	1	4
	PPIB	1		1	2	4
	GAPDH	1		1	2	4
	CTSH	2				2
	RPL28	1		2	1	4
	CD320	1		1	1	3
	SETD2			1		1
	UQCRC1			2	2	4
	STK25			1		1
	ATP5A1			1		1
	MAN1B1			1		1
	ATP1A1			1	1	2
	RPS4X				2	2
	GNAS				1	1
	C6orf108				1	1
	ELF3				1	1
	SFRS9				1	1
BES02	*ALDH3A2*		5	1	1	7
	INTS9		1			1
	CCNB1		1	1	2	4
	RPS5	1	1	1	1	4
	RPS2	1		1		2
	FN1	1			1	2
	AOF2			2		2
	ST13			1		1
	CREBB3			1	2	3
	SEC61G			1		1
	DNAJC10				1	1
	MY01C				1	1
BES01 e BES02	*KRT18*	1			2	3
	DDX46		1		1	2
	GSTP1			1	1	2
	CLCT			1	1	2
	PSMC2				1	1
TOTAL	**39**	**12**	**16**	**28**	**35**	**91**

Na maioria dos casos mais de um transcrito variante foi identificado para um mesmo gene, sendo identificados no total 79 variantes de *splicing* para as 39 ASSETs, reportando 91 eventos de *splicing* distintos. Foi identificado em média 2,3 eventos de *splicing* alternativo para cada ASSET.

O tipo de evento mais encontrado foi o uso de sítios de *splice* alternativos, sendo que 35 eventos reportaram o uso alternativo de sítios de *splice* 5' (sítio doador) e 28 eventos reportaram o uso alternativo de sítios de *splice* 3' (sítio aceptor). Dezesseis eventos reportaram o uso alternativo de éxons e 12 eventos reportaram retenção de íntrons (Figura 24; Tabela 7).

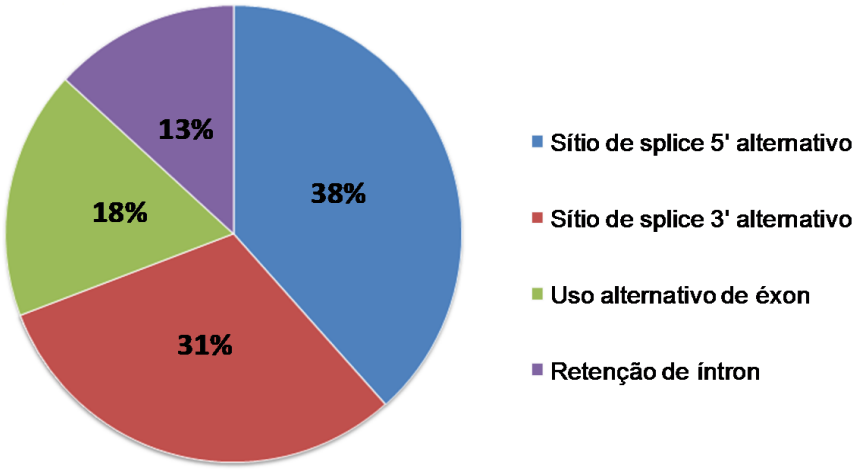


Figura 24: Identificação de eventos de *splicing* alternativo. A porcentagem de eventos de cada tipo está representada em relação ao número total de eventos identificados.

As demais 36 ASSETS para as quais nenhum evento de *splicing* alternativo foi reportado pelas sequências nos bancos públicos podem ter sido resultado de hibridações com transcritos ainda não caracterizados e representam um grupo com potencial para identificação de transcritos novos. No entanto, não pode ser descartada a possibilidade de terem resultado de hibridações das ASSETs com transcritos de uma mesma família proteica que compartilhem regiões de alta similaridade ou entre transcritos originados de pseudogenes.

4.1.5 Validação de eventos de *splicing* alternativo identificados pelas bibliotecas de cDNA enriquecidas para *splicing* alternativo: BES01 e BES02

Dezoito ASSETs foram randomicamente selecionadas para validação por RT-PCR, sendo 6 ASSETs exclusivas da biblioteca BES01, 7 ASSETs exclusivas da biblioteca BES02 e 5 ASSETs detectadas pelas duas bibliotecas. O processo de validação foi realizado em duas etapas. A primeira etapa, denominada validação da ASSET, consistiu na confirmação da expressão do transcrito correspondente à sequência ASSET utilizando a mesma fonte de RNA utilizada para construção das bibliotecas (Figura 25). Para as ASSETs identificados na biblioteca BES01 foi utilizado cDNA sintetizado a partir de RNA da linhagem C5.2 e para as ASSETs identificadas na biblioteca BES02 foi utilizado cDNA sintetizado a partir de RNA correspondente às 5 amostras tumorais de mama. Para as 5 ASSETs detectadas por ambas bibliotecas foram utilizados cDNAs sintetizados tanto a partir de RNA da linhagem C5.2 quanto a partir do grupo de amostras tumorais de mama. Os iniciadores utilizados para validação foram desenhados nas extremidades da sequência de cada ASSET (Figura 25). A segunda etapa, denominada validação do heteroduplex, consistiu na identificação de transcritos alternativos que poderiam ter participado no processo de formação do heteroduplex durante a construção da biblioteca, utilizando os mesmos oligonucleotídeos utilizados na validação da ASSET (Figura 25). Das 18 ASSETs selecionadas para validação, 12 apresentam variantes de *splicing* nos bancos de dados e 6 não apresentam variantes nos bancos de dados públicos.

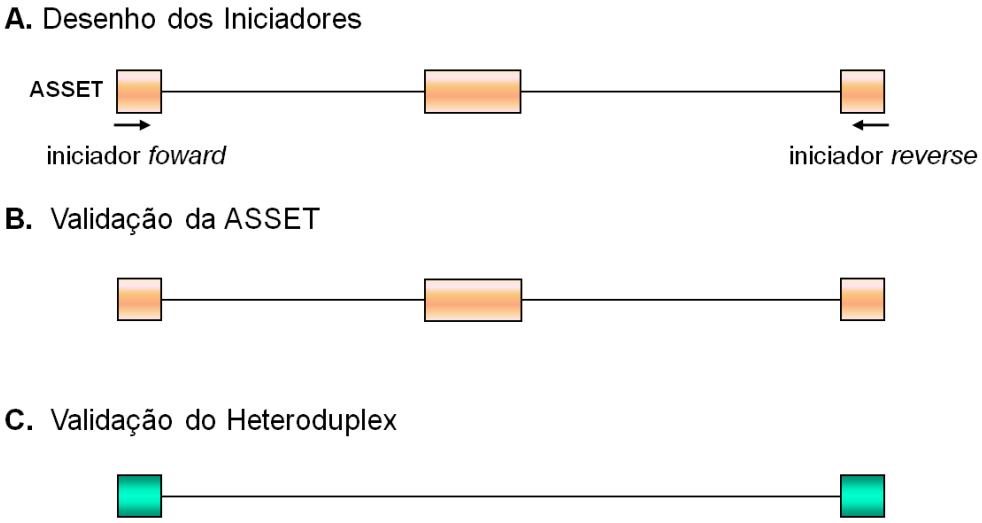


Figura 25: Validação das ASSETs. O esquema mostra um exemplo de uma ASSET. A – Desenho dos iniciadores nas extremidades da sequência ASSET. B – Etapa 1 de validação, ou validação da ASSET, que consistiu na amplificação da mesma variante identificada na biblioteca. C – Etapa 2 da validação, ou validação do heteroduplex, que consistiu na amplificação de uma segunda variante que possa ter participado da formação do heteroduplex. Os retângulos representam os éxons e as linhas representam os intros.

Em relação à etapa 1, validação das ASSETs, a taxa de validação foi de 94,4%, onde apenas uma (*CDK5RAP2* da biblioteca BES01) das 18 ASSETs não foi validada (Tabela 8). Em relação à etapa 2, validação do heteroduplex, a taxa de validação foi de 35,3%, na qual um transcrito alternativo que possa ter participado da formação do heteroduplex foi identificado para seis dos 17 genes (*SFRS9, FLNA, ALDH3A2, PTPLA, RPS2* e *TRIP6*) (Tabela 8; Figura 26). Todas as ASSETs e variantes identificadas pela reação de RT-PCR foram confirmadas através de sequênciamento.

Tabela 8: Resultado das etapas de validação para as 18 ASSETs selecionadas. As ASSETs estão representadas pelo símbolo do gene correspondente e estão agrupadas em relação as bibliotecas de origem, BES01 e BES02.

	Gene	Evento de *splicing* alternativos no banco de dados	Validação da ASSET	Validação do Heteroduplex	Validação cruzada
BES01	*SFRS9*	Sítio de *splice* alternativo 3'	sim	sim	sim
	ATP1A1	Sítio de *splice* alternativo 5'	sim	não	não
	CDK5RAP2	Sítio de *splice* alternativo 5'	não	não	não
	ITGB5	Sem evento	sim	não	sim
	FLNA	Uso alternativo de éxon	sim	sim	sim
	RBM10	Sem evento	sim	não	sim
BES02	*PTPLA*	Sem evento	sim	sim	sim
	ALDH3A2	Uso alternativo de éxon	sim	sim	sim
	RPS2	Retenção de íntron	sim	sim	sim
	FN1	Sem evento	sim	não	sim
	TRIP6	Sem evento	sim	sim	sim
	COL7A1	Sem evento	sim	não	sim
	AOF2	Sem evento	sim	não	sim
BES01 e BES02	*DDX46 / EIF4A3*	Uso alternativo de éxon e sítio de *splice* alternativo 3'	sim	não	-
	KRT18	Sítio de *splice* alternativo 5'	sim	não	-
	GSTP1	Sítios de *splice* alternativos 3' e 5'	sim	não	-
	CLCT	Sítios de *splice* alternativos 3' e 5'	sim	não	-
	PSMC2	Sítio de *splice* alternativo 5'	sim	não	-

As variantes identificadas para os genes *SFRS9* (Figura 26A) e *PTPLA* (Figura 26B) reportaram eventos de uso alternativo de sítios de *splice*, sendo o sítio de *splice* 3' para o gene *SFRS9* e 5' para o gene *PTPLA*. As variantes identificadas para o gene *FLNA* (Figura 26C) e *ALDH3A2* (Figura 26D) reportam eventos de exclusão de éxons (*exon skipping*). Já as variantes identificadas para os genes *RPS2* (Figura 26E) e *TRIP6* (Figura 26F) reportam eventos de retenção de íntron.

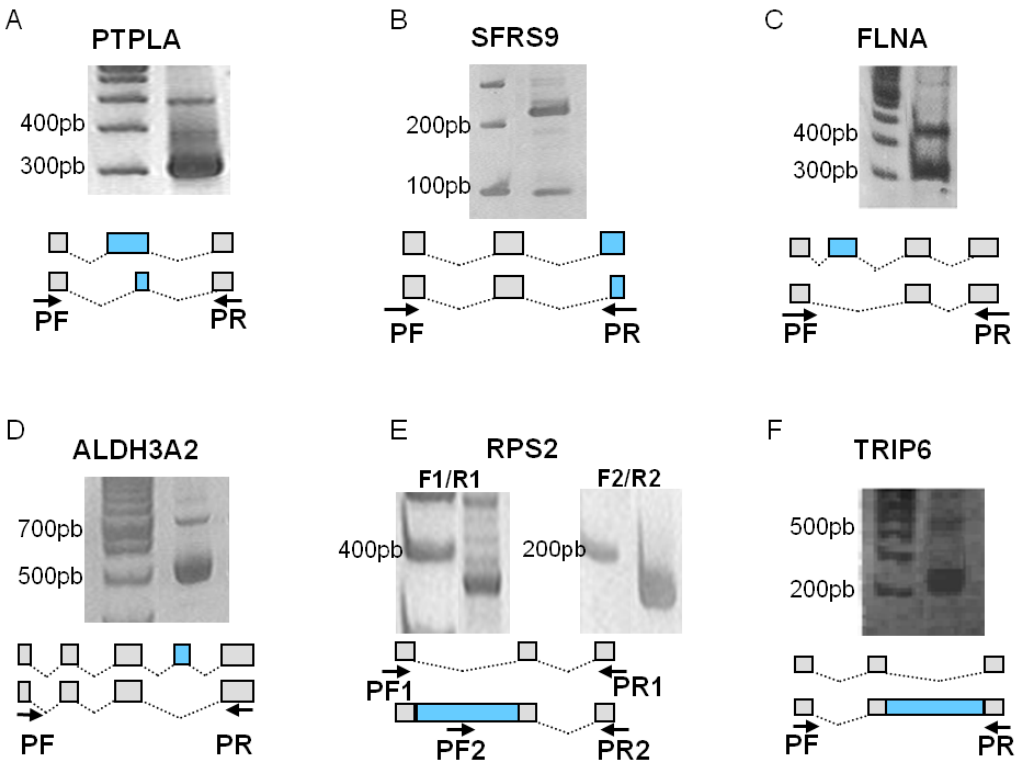


Figura 26: Validação do heteroduplexes para 6 ASSETs. O resultado da amplificação para cada gene é mostrado separadamente. Os produtos de RT-PCR foram avaliados em gel de agarose 1%. No esquema estão representados os éxons correspondentes as ASSETs e as variantes e indicado a localização dos iniciadores *foward* (PF) e *reverse* (PR). Os éxons alternativos estão coloridos em azul. A – Gene *PTPLA*. B – Gene *SFRS9*. C – Gene *FLNA*. D – Gene *ALDH3A2*. E – Gene *RPS2*. Para este gene um segundo par de iniciadores foi desenhado, para amplificação específica da variante com retenção de íntron. F – Gene *TRIP6*.

Destes seis transcritos, dois são transcritos alternativos de *splicing* novos dos genes *PTPLA* e *TRIP6*. O gene *PTPLA* é membro da família *protein tyrosine phosphatase-like* que inclui proteínas similarves as tirosinas fosfatases, mas que contêm o aminoácido prolina no lugar do aminoácido arginina no domínio catalítico (UWANOGO et al., 1999). Este gene está localizado no cromossomo 10 e contém 7 éxons. A variante nova identificada resulta do uso de um sítio de *splice* 5' alternativo que aumenta o tamanho do éxon 5 em 117 nucleotídeos (Figura 27). A proteína PTPLA de 288 aminoácidos contém um domínio proteína tirosina fosfatase-*like* que abrange toda proteína; um sítio ativo tirosina fosfatase entre os aminoácidos 127 a 139; um sítio conservado chaperonina (Cpn60) entre os aminoácidos 41 a 52 e três domínios transmembrânicos nas posições 75aa a 95aa, 205aa a 225aa

e 248aa a 268aa. O transcrito alternativo resulta em uma menor sequência aberta de leitura (208 aa), por inserir um códon de parada prematuro na posição 662 nts do RNAm. De acordo com as predições do banco de dados *InterPro*, esta proteína apresenta os mesmos domínios proteicos que a variante conhecida, no entanto, a localização dos 3 domínios transmembrânicos são alteradas para 76aa a 98aa; 126aa a 146aa e 167aa a 187aa (Figura 27).

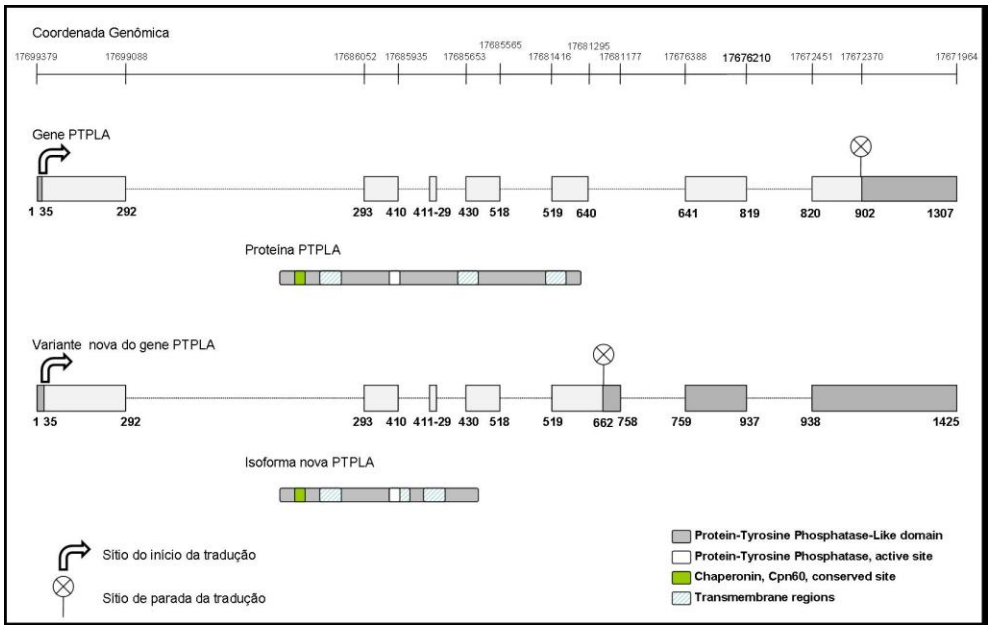


Figura 27: Caracterização da nova variante do gene *PTPLA*. Na parte superior da figura estão colocadas as coordenadas genômicas referente às localizações dos éxons do gene *PTPLA*. A estrutura do gene *PTPLA* e da nova variante identificada estão esquematizadas por retângulos representando os éxons e linhas pontilhadas representando os íntrons. As porções escuras nos éxons são as regiões 3'e 5'não traduzidas. Abaixo de cada esquema está representada a proteína com seus domínios proteicos.

O gene *TRIP6* é um *thyroid hormone receptor interactor 6*. Este gene está localizado no cromossomo 7 e contém 9 éxons. O transcrito novo identificado resulta da retenção do último íntron (Figura 28). A proteína TRIP6 apresenta 476 aminoácidos e apresenta 3 dominíos proteicos LIM, que são domínios do tipo dedo de zinco, envolvidos com ligação ao DNA, RNA e proteínas. A nova isoforma proteica apresenta 446 aminoácidos (30 aminoácidos a menos), e também insere um códon de parada prematuro na proteína sem, no entanto, interferir com os domínios proteicos (Figura 28).

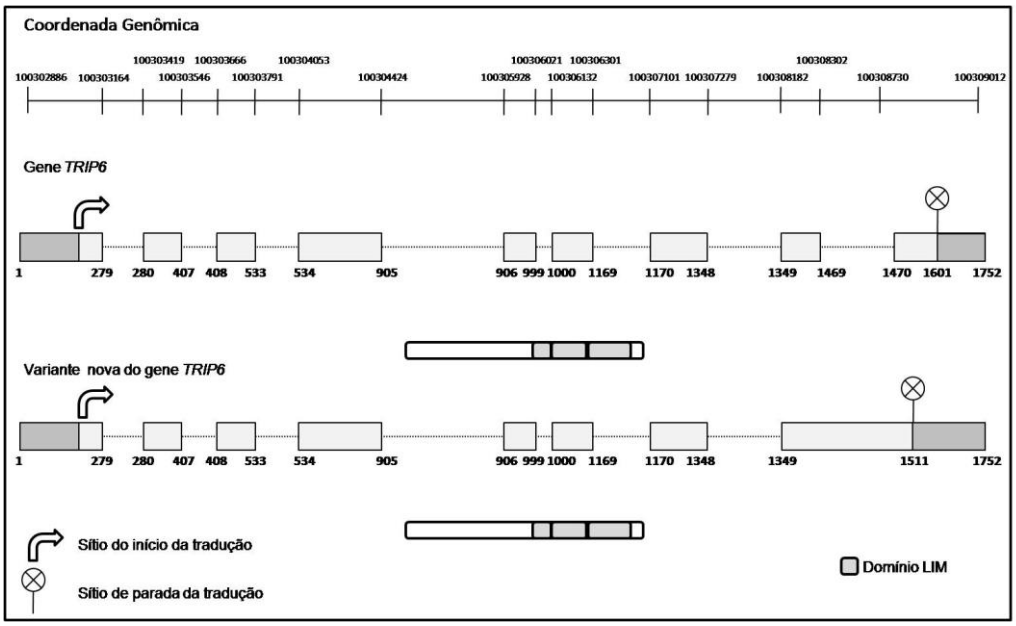


Figura 28: Caracterização da nova variante do gene *TRIP6*. Na parte superior da figura estão colocadas as coordenadas genômicas referente às localizações dos éxons do gene. A estrutura do gene *TRIP6* e da nova variante identificada estão esquematizadas por retângulos representando os éxons e linhas pontilhadas representando os íntrons. As porções escuras nos éxons são as regiões 3´e 5´não traduzidas. Abaixo de cada esquema está representada a proteína com os domínios proteicos.

O fato de não terem sido identificados transcritos alternativos para os demais 11 genes selecionados deve ser devido a uma expressão diferencial entre as variantes, onde a amplificação da variante mais expressa é favorecida em relação à amplificação da variante menos expressa. Dentre estes 11 genes, 8 apresentam transcritos reportados pelos bancos de dados públicos.

Apesar de as duas bibliotecas BES01 e BES02 terem sido originadas a partir de fontes de RNA similares, uma linhagem celular de mama com características tumorais com alta expressão do gene *ERBB2* e um grupo de amostras tumorais de mama que também apresentam alta expressão de ERBB2, a porcentagem de sobreposição de resultados foi em torno de 10%. Assim para investigar se este fato foi devido à baixa cobertura das bibliotecas (aproximadamente 1000 sequências de cada), ou resultado de diferenças entre linhagens celulares e amostras, foram analisadas por RT-PCR se as ASSETs identificadas por uma biblioteca seriam também expressas na fonte

de RNA utilizada para construção da outra biblioteca. Estas verificação foi denominada de validação cruzada. Quatro ASSETs das 6 (66,7%) identificadas pela biblioteca BES01 foram expressas no cDNA representativo da biblioteca BES02, e todas as 10 ASSETS identificadas pela biblioteca BES02 foram expressas no cDNA da linhagem C5.2, representativo da biblioteca BES01 (Tabela 8). Este fato sugere que a pequena sobreposição de ASSETs identificadas por ambas as bibliotecas seria devido ao pequeno número de sequências geradas.

4.1.6 Regulação das variantes de *splicing* pela expressão diferencial de *ERBB2*

Para avaliar o perfil de expressão das variantes de *splicing* entre amostras com diferente nível de expressão do gene *ERBB2*, foi proposto inicialmente analisar de forma quantitativa o nível de expressão das variantes através de experimentos de RT-PCR quantitativo.

No primeiro momento, foi avaliado o perfil de expressão das variantes do gene *SFRS9* com a utilização de iniciadores desenhados nas regiões específicas das duas variantes validadas (Figura 29).

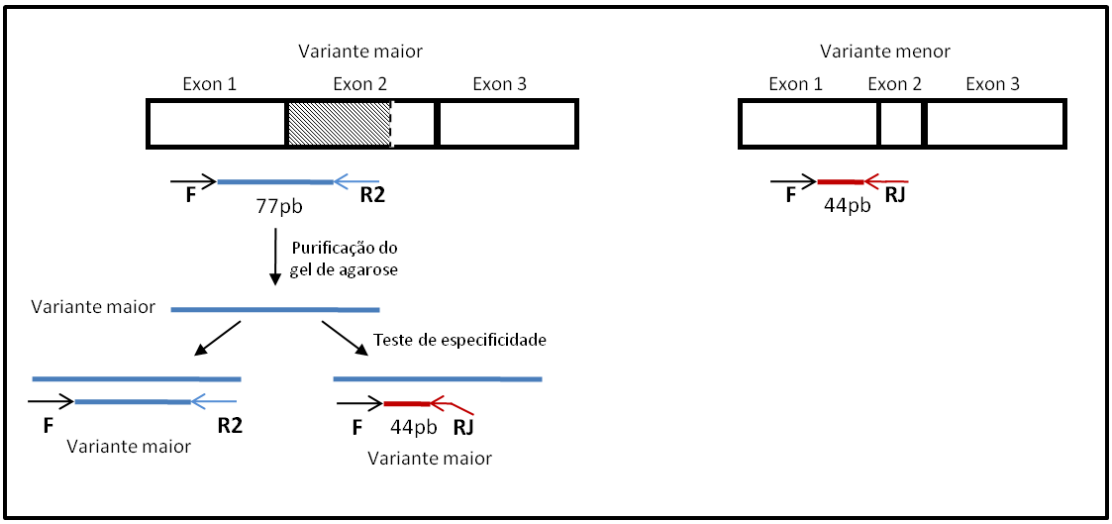


Figura 29: Desenho dos iniciadores para validação por RT-PCR quantitativo. À direita esta o esquema de amplificação da variante menor com a localização dos oligonucleotídeos específicos resultando em um produto de 44pb. À esquerda está o esquema de amplificação da variante maior, resultando em um produto de 77pb. Para o teste de especificidade, a variante maior foi purificada por gel de agarose e utilizada como molde para reação de teste de especificidade com os oligonucleotídeos F e RJ para a amplificação da variante menor.

As variantes reportam o uso alternativo de sítio aceptor de *splice* alterando o tamanho do éxon 2. Assim, um par de iniciadores específicos para a variante maior foi desenhado, denso um deles no éxon 1 (F), comum as duas variantes, e o outro na região do éxon 2 exclusiva da variante maior (R2). Para a variante menor foi utilizado o mesmo iniciador no éxon 1 comum (F) e um segundo iniciador desenhado na junção entre o éxon 1 e éxon 2 reduzido (RJ), a qual é específica da variante menor. O iniciador RJ apresenta 4pb complementares ao éxon flanqueador a montante, éxon 1, podendo utilizar a variante maior como molde resultando em amplificação inespecífica. Para verificar esta possibilidade, a variante maior foi amplificada com os oligonucleotídeos F e R2 e isolada através de purificação de gel de agarose. Em seguida, foi utilizada como molde para uma reação de RT-PCR com o par de iniciadores F e RJ (Figura 30). Como controle positivo da reação foi utilizado cDNA da linhagem C5.2.

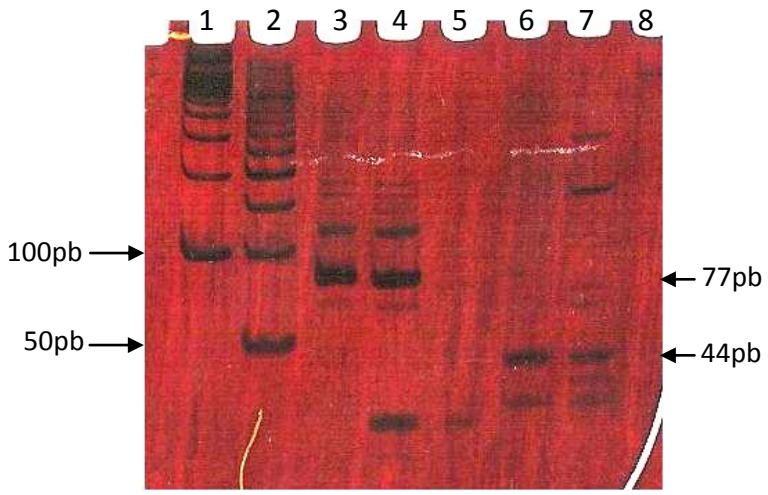


Figura 30: Teste de especificidade do gene *SFRS9*. 1- Marcador 100pb. 2- Marcador 50pb. 3- Amplificação da variante com a região alternativa utilizando cDNA da linhagem C5.2 como molde (77pb). 4- Amplificação da variante com a região alternativa utilizando produto da variante maior purificado como molde (77pb). 5- Controle negativo da reação com a região alternativa, sem molde. 6- Amplificação da variante sem a região alternativa utilizando cDNA da linhagem C5.2 como molde (44pb). 7- Amplificação da variante sem a região alternativa utilizando produto da variante maior purificado como molde (44pb). 8- Controle negativo da reação sem a região alternativa, sem molde. Gel de acrilamida 8%.

Todos os controles utilizados apresentaram resultados esperados. A amplificação da variante maior a partir de cDNA da linhagem C5.2 ou a partir do produto de RT-PCR purificado resultou em uma banda de 77pb. Além disso, a amplificação da variante menor com cDNA da linhagem C5.2 apresentou amplificação de produto de 44pb. No entanto, houve amplificação da variante menor, utilizando o par de iniciadores F e RJ, ao utilizar o produto de RT-PCR purificado correspondente a variante maior (Figura 30 – poço 7). Este fato mostra que mesmo contendo apenas 4pb de sobreposição, o iniciador RJ específico da variante menor, foi suficiente para promover amplificação a partir da variante maior. Como o tamanho do amplificado é o mesmo, independentemente se a amplificação utilizou como molde a variante que contém a região alternativa ou que não contém a região alternativa do éxon 2, não há como avaliar o nível de expressão específico da variante menor.

Devido a dificuldade em obter amplificação específica de cada variante pela impossibilidade de utilizar oligonucleotídeos com regiões comuns a

ambas variantes, as análises de perfil de expressão das variantes de *splicing* baseadas em experimentos de RT-PCR quantitativo foram inviabilizadas.

Uma das abordagens alternativas propostas foi a utilização da metodologia de eletroforese em *chip,* que realiza uma eletroforese em microfluidos e permite maior sensibilidade na detecção e quantificação de difententes produtos de RT-PCR gerados em uma mesma reação, sendo uma excelente alternativa para análise do nível de expressão das variantes de *splicing* sem a necessidade do desenho de oligonucleotídeos específicos para cada variante (VENABLES et al., 2008a).

Para verificar uma possível influência da alta expressão do oncogene ERBB2 na regulação da expressão das diferentes variantes de *splicing* foram realizadas novas reações de RT-PCR para as 17 ASSETs validadas, utilizando cDNA da linhagem C5.2 e HB4a, que apresenta expressão basal de *ERBB2*. Além das 17 ASSETs e variantes de *splicing* validadas foi avaliado o nível de expressão do gene *GAPDH*, como fator normalizador.

Para avaliar a diferença de expressão foram comparados os valores de concentração (ng/µl) dos fragmentos de tamanhos específicos de cada ASSET para as linhagens HB4a e C5.2. O valor obtido na amplificação do gene *GAPDH* foi de 17,6 ng/µl na linhagem HB4a e de 22,2 ng/µl na linhagem C5.2. Estes valores foram utilizados como fatores de normalização, sendo que os dados obtidos para todas as ASSETs com cDNA da linhagem HB4a foram divididos por 1,76 e os valores obtidos com cDNA da linhagem C5.2 foram divididos por 2,22.

Inicialmente foram analisadas as 11 ASSETs validadas apenas na etapa 1 (validação da ASSET). Todos transcritos avaliados apresentaram expressão similar nas linhagens HB4a e C5.2, com valores de razão variando entre -1,89 a 1,39 (Tabela 9), sugerindo que estes transcritos não são modulados pelo nível de expressão de *ERBB2*.

Tabela 9: Análise de expressão das ASSETs entre as linhagens HB4a e C5.2 por eletroforese em *chip*. A tabela apresenta o nome dos genes, o tamanho esperado de amplificação da ASSET, os valores de concentração e concentração normalizada e a razão de expressão na linhagem C5.2 em relação a linhagem HB4a.

Gene	Tamanho ASSET	Linhagem celular	Concentração (ng/ul)	Concentração Normalizada	Razão C5.2/HB4a
ATP1A1	417pb	HB4a	6,79	3,86	-1,19
		C5.2	7,18	3,23	
ITGB5	143 pb	HB4a	0,22	0,13	-1,07
		C5.2	0,26	0,12	
RBM10	184 pb	HB4a	13,61	7,73	-1,17
		C5.2	14,67	6,61	
COL7	387 pb	HB4a	1,27	0,72	1,39
		C5.2	2,22	1,00	
AOF2	391 pb	HB4a	19,13	10,87	1,24
		C5.2	29,87	13,45	
FN1	293 pb	HB4a	1,74	0,99	-1,89
		C5.2	1,16	0,52	
EIF4A3	132 pb	HB4a	1,51	0,86	-1,47
		C5.2	1,3	0,59	
GSTP1	334 pb	HB4a	14,8	8,41	-1,18
		C5.2	15,83	7,13	
KRT18	179 pb	HB4a	11,15	6,34	1,02
		C5.2	14,29	6,44	
PSMC2	172 pb	HB4a	15,02	8,53	-1,04
		C5.2	18,17	8,18	
CTCL	390 pb	HB4a	24,85	14,12	-1,22
		C5.2	25,63	11,55	

Em seguida, foi avaliado o padrão de *splicing* das seis ASSETs que apresentaram amplificação de um segundo transcrito para o mesmo gene. Neste caso foi calculado inicialmente a razão entre o valor de expressão normalizado da ASSET e do transcrito variante, para em seguida ser calculada a diferença de expressão entre a linhagem C5.2 e HB4a, o que permitiu analisar um possível desbalanço de expressão das variantes entre as linhagens celulares de mama. Para 3 genes, *RPS2*, *PTPLA* e *ALDH3A2,* não foi encontrada diferença no balanço de expressão das variantes entre as linhagens celulares (Tabela 10). O gene *RPS2* apresentou expressão similar entre as variantes, no entanto, ambas foram mais expressas na linhagem

celular HB4a em relação à linhagem C5.2, indicando uma possível diminuição da expressão do gene como um todo, influenciada pelo aumento de expressão do gene *ERBB2*. As variantes do gene *PTPLA* e do gene *ALDH3A2* não apresentaram nenhuma alteração no nível de expressão entre as linhagens. Tanto na linhagem HB4a quanto na linhagem C5.2 a variante menor (ASSET) do gene *PTPLA* foi cerca de 90 vezes mais expressa do que a variante maior. De forma similar a variante menor do gene *ALDH3A2* (ASSET) foi cerca de 10 vezes mais expressa do que a variante maior nas duas linhagens investigadas.

Foi identificada alteração no balanço de expressão das variantes dos genes *SFRS9*, *FLNA* e *TRIP6* entre as linhagens HB4a e C5.2 (Tabela 10; Figura 31). Nos três casos foi observado menor expressão da ASSET na linhagem C5.2 em relação à linhagem HB4a, acompanhada por uma expressão aumentada ou inalterada da variante envolvida na formação do heteroduplex na linhagem C5.2 em relação à HB4a, resultando em um desbalanço na expressão das variantes entre as duas linhagens de mama. Este desbalanço no nível de expressão das variantes pode ser decorrente da influência do nível de expressão do oncogene *ERBB2*. A variante maior (ASSET) do gene *SFRS9* apresentou diferença de expressão de 36,17 vezes em relação à variante menor na linhagem HB4a. Esta diferença foi reduzida para apenas 7,48 vezes na linhagem C5.2, resultando em um desbalanço de quase 5 vezes na diferença de expressão entre as variantes (Tabela 10). Um desbalanço similar foi encontrado entre as variantes do gene *FLNA* sendo que a diferença de expressão de 95,69 vezes entre a variante maior (ASSET) e a variante menor encontrada na linhagem HB4a foi reduzida para 20,01 vezes na linhagem C5.2, a qual foi decorrente tanto de um aumento da expressão da variante menor quanto da diminuição da expressão da variante maior (Tabela 10). A variante menor (ASSET) do gene *TRIP6* apresentou redução de expressão na linhagem C5.2 quando comparada com o valor de expressão encontrado na linhagem HB4a (cerca de 4,6 vezes). Esta redução não foi encontrada para a variante maior, que manteve valores de expressão praticamente iguais entre as linhagens (1,35 vezes de diferença), resultando

em um desbalanço de expressão entre as variantes de mais de 3 vezes nas linhagens avaliadas (Tabela 10).

Tabela 10: Análise de expressão das ASSETs e variantes entre as linhagens de mama HB4a e C5.2 pela eletroforese em *chip*. A tabela apresenta o nome dos genes, o tamanho esperado do produto de amplificação da ASSET e da variante (resultante da validação do heteroduplex), os valores de expressão normalizados e a razão de expressão na linhagem C5.2 em relação à linhagem HB4a.

Gene	Linhagem celular	Fragmento	Tamanho (pb)	Concentração (ng/ul)	Concentração Normalizada	ASSET/ variante	C5.2/ HB4a
SFRS9	HB4a	ASSET	232	6,51	3,70	36,17	-4,84
		variante	100	0,18	0,10		
	C5.2	ASSET	232	6,43	2,90	7,48	
		variante	100	0,86	0,39		
FLNA	HB4a	ASSET	500	24,88	14,14	95,69	-4,78
		variante	377	0,26	0,15		
	C5.2	ASSET	500	23,01	10,36	20,01	
		variante	377	1,15	0,52		
ALDH3A2	HB4a	ASSET	470	20,04	11,39	10,07	1,02
		variante	610	1,99	1,13		
	C5.2	ASSET	470	24,61	11,09	10,25	
		variante	610	2,4	1,08		
TRIP6	HB4a	ASSET	203	12,60	7,16	8,87	-3,40
		variante	636	1,42	0,81		
	C5.2	ASSET	203	3,47	1,56	2,61	
		variante	636	1,33	0,60		
PTPLA	HB4a	ASSET	324	10,8	6,14	83,08	1,18
		variante	456	0,13	0,07		
	C5.2	ASSET	324	16,72	7,53	98,35	
		variante	456	0,17	0,08		
RPS2	HB4a	ASSET	187	13,68	7,77	1,42	1,12
		variante	390	12,18	5,49		
	C5.2	ASSET	187	1,79	1,02	1,59	
		variante	390	1,42	0,64		

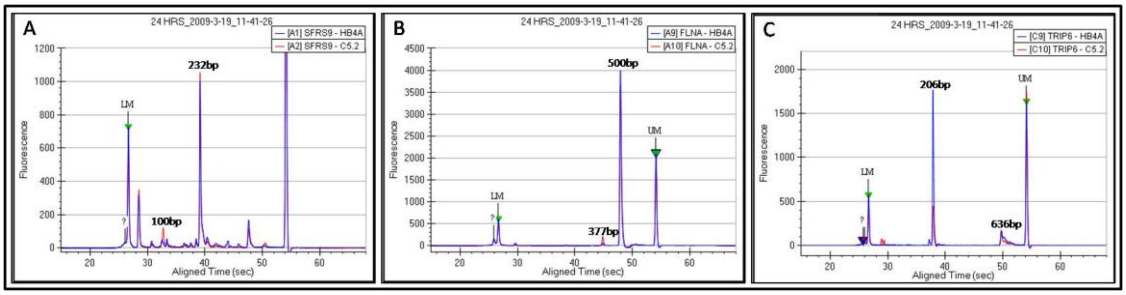


Figura 31: Eletroforese em *chip*. A linha azul correponde as amplificações utilizando cDNA da linhagem HB4a como molde. A linha vermelha correpodende as amplificações utilizando cDNA da linhagem C5.2 como molde. As setas em verde indicam os marcadores internos, LM (*lower marker*) e UM (*upper marker*), usados como controle do próprio equipamento. A – Amplificação de dois transcritos alternativos (100 e 232pb) do gene *SFRS9*. B – Amplificação de dois transcritos alternativos (377 e 500pb) do gene *FLNA*. C – Amplificação de dois transcritos alternativos (206 e 636pb) do gene *TRIP6*.

Estes resultados sugerem que alterações no balanço de expressão de variantes de *splicing* podem ser mediadas pela expressão diferencial do gene *ERBB2*. Para confirmar as alterações encontradas pela eletroforese capilar no balanço de expressão das variantes dos genes *SFRS9*, *FLNA* e *TRIP6* foi utilizada uma segunda abordagem baseada na ligação de sondas variante-específicas e amplificação por PCR. Nesta estratégia dois pares de sondas foram desenhadas para cada gene, sendo cada par específico para cada variante de interesse. Cada par de sonda foi desenhado no limite éxon/éxon específico da variante. Uma das sondas, sonda esquerda, contém na extremidade 5' uma sequência conhecida que serviu para o anelamento do iniciador *foward* na reação de PCR, seguida de uma região de 38 nucleotideos que servem apenas para aumentar o tamanho da sonda, e por fim uma região complementar ao limite éxon/éxon do éxon a montante. A outra sonda, sonda direita, é fosforilada na sua extremidade 5' a qual contém sequência complementar ao limite éxon/éxon do éxon a jusante, seguida de uma região de sequência conhecida que serviu para o anelamento do iniciador *reverse* na reação de PCR. As sondas foram incubadas com cDNA, e na presença da variante de interesse ocorreu ligação entre a sonda esquerda e a sonda direita, pela presença do grupo fosfato, gerando uma sonda única. Esta sonda unida foi utilizada para amplificação por PCR (Figura 32).

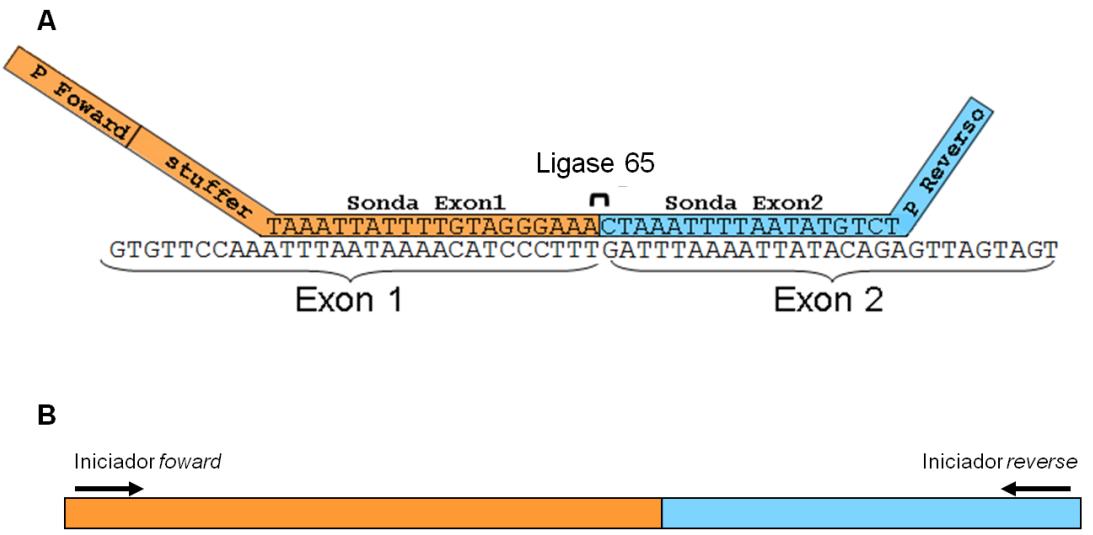


Figura 32: Estratégia de avaliação do nível de expressão das variantes de *splicing* baseada no desenho de sondas-específicas e amplificação por PCR. A – Hibridação das sondas ao cDNA. A sonda esquerda está representada em laranja e a sonda direita em azul. P Foward - Sequência para o anelamento do iniciador *foward*. P Reverso - Sequência para o anelamento do iniciador *reverse*. Stuffer – Sequência de preenchimento. As regiões das sondas complementares ao cDNA estão representadas pelas bases complementares, sendo possível verificar o posicionamento das sondas em regiões adjacentes no cDNA no limite entre dois éxons distintos. Para união da sonda esquerda a sonda direita foi utilizada a enzima Ligase 65. B – Amplificação da sonda resultante, com iniciadores nas extremidades.

Para esta estratégia foram sintetizados cDNA a partir de RNA total das linhagens HB4a e C5.2 utilizando oligonucleotídeos específicos para cada gene. Esses oligonucleotídeos foram desenhados na extremidade 3’ dos genes e possibilitam a síntese de cDNA de todas as variantes de cada gene, diminuindo a possibilidade de hibridações inespecíficas das sondas com outros transcritos. O produto de PCR foi analizado por eletroforese em gel de acrilamida 8% (Figura 33).

A diferença no balanço de expressão entre as variantes de *splicing* dos 3 genes (*FLNA*, *SFRS9* and *TRIP6*) foi confirmada pela estratégia de ligação de sondas variante-específicas, fortalecendo a sugestão de influência da expressão diferencial de *ERBB2* na regulação do *splicing* alternativo para estes genes.

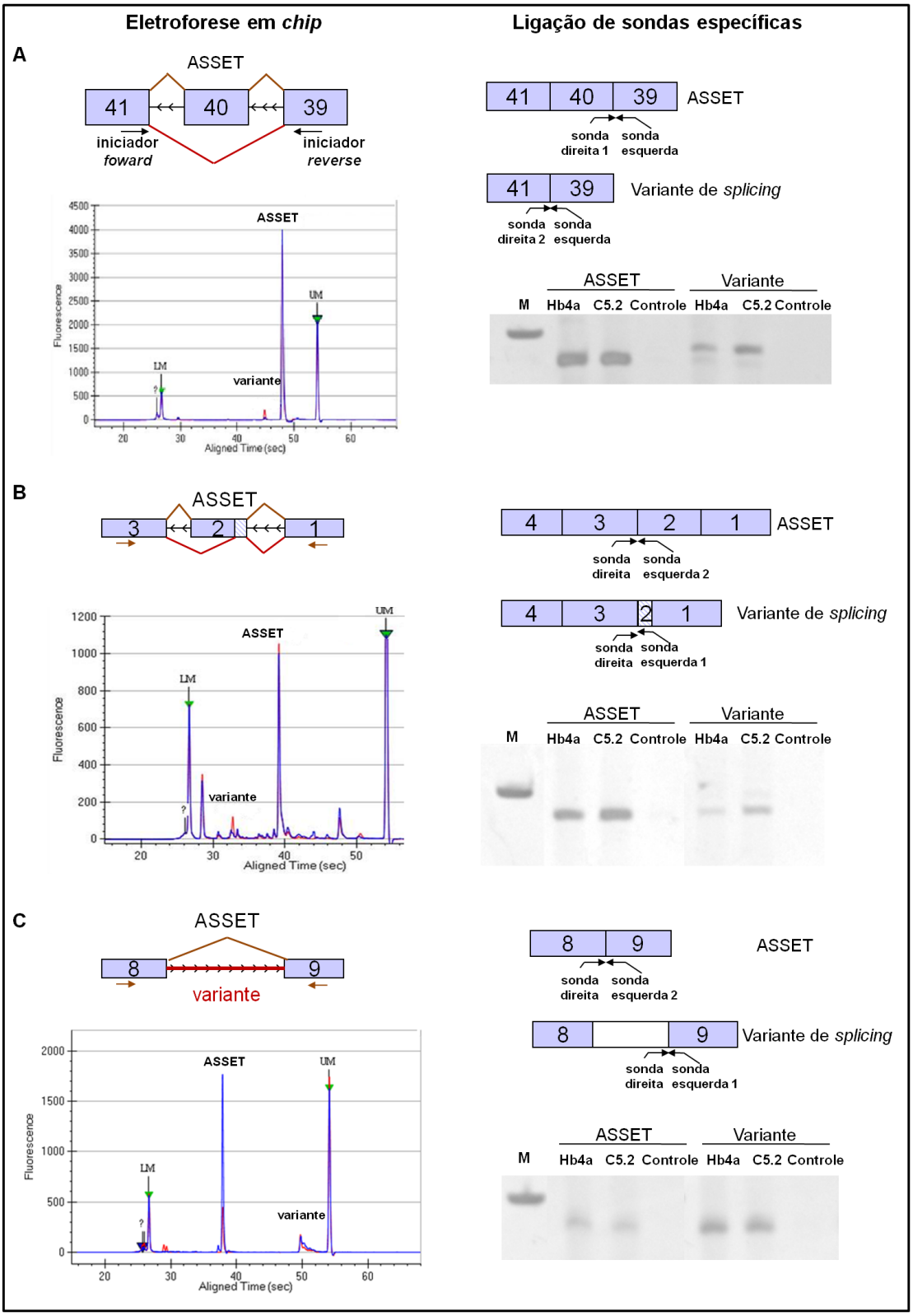


Figura 33: Análise do perfil de expressão das variantes de *splicing*. Alterações no balanço de expressão entre as linhagens de mama HB4a e C5.2 foram avaliadas pelas metodologias de eletroforese capilar e sondas-específicas. A – Gene *SFRS9*. B – Gene *FLNA*. C – Gene *TRIP6*.

4.1.7 Anotação funcional das variantes de *splicing*

Para avaliar os aspectos funcionais dos genes que apresentam *splicing* alternativo, os 142 genes identificados pelas duas bibliotecas foram classificados de acordo com categorias funcionais de Processos Biológicos. Através do banco de dados *Gene Ontology*, os genes foram classificados e agrupados em 8 categorias, sendo que genes pertencentes a mais de uma categoria foram categorizados de maneira hierárquica na seguinte ordem: Ciclo celular, *Splicing*, Transcrição, Tradução, Transdução de Sinal, Metabolismo Proteico, Metabolismo, Adesão e Migração Celular (Tabela 11). Os genes pertencentes a categorias distintas as mencionadas foram agrupados em "Miscelânia".

Tabela 11: Classificação funcional dos genes em processos biológicos.

Processos Biológicos	Número de genes	Genes
Ciclo Celular	3	*CCNB1, KRT18, NAP1L1*
Splicing	5	*ASCC3L1, EIF4A3, INTS9, RBMX, SFRS9*
Transcrição	7	*AOF2, CREB3, ELF3, FLNA, HDAC2, PHF19, SETD2*
Tradução	11	*EEF2, FARS2, GSPT1, MRPL45, PAIP1, RPL11, RPL28, RPL6, RPS2, RPS4X, RPS5*
Transdução de Sinal	9	*CAMK2G, CDC42SE1, GNAS, GNB3, GRK6, INPP1, PTPLA, RANBP1, STK25*
Metabolismo Proteico	9	*CTSH, DDB2, DNAJC10, PPIB, PMSC2, PSMD6, ROCK2, ST13, UQCRC1*
Metabolismo	7	*ACLY, ALDH3A2, ATP1A1, GAPDH, MAN1B1, MAN2A1, OSBPL8*
Adesão e Migração Celular	4	*COL7A1, FN1, ITGB5, TRIP6*
Miscelânia	16	*ATP5A1, ATXN10, C6ORF108, CD320, CDK5RAP2, CLCT, DDEF1, GABARAP, GDF9, GNPTAB, PTPRA, RBM10, SEC61G, SGSM2, SLC4A2, XPO1*
Não categorizados	7	*C7ORF55, DENND4C, KIAA0090, KIAA0152, MYO1C, RNF149, THSD1*

Foi observado enriquecimento significativo nas categorias Elongação da Tradução (p=7E-7), Tradução (p=1,6E-5), Processo Metabólico de Proteínas Celulares (p=1,9E-5), Processo Metabólicos de Proteínas (p=2,4E-5) e Processos Metabólicos de Macromoléculas Celulares (p=3,5E-5) (Figura 34).

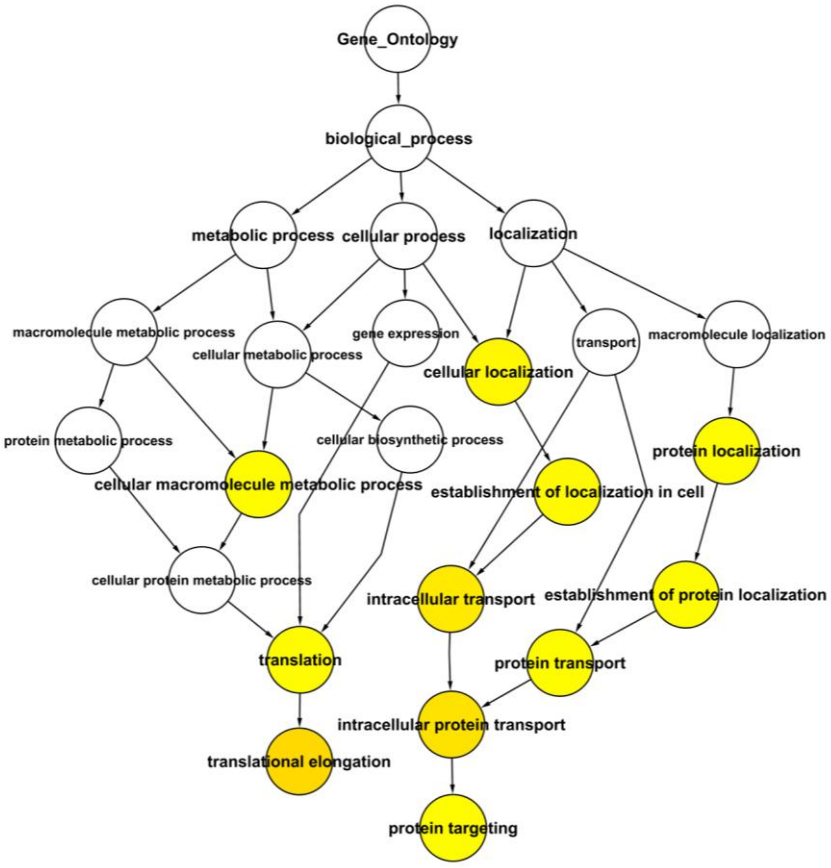


Figura 34: Anotação funcional dos genes. Os círculos em amarelo correspondem a categorias estatisticamente significantes.

4.2 Bibliotecas de cDNA para análise do transcriptoma completo

4.2.1 Estabelecimento da metodologia de construção de bibliotecas para análise do transcriptoma completo das linhagens HB4a e C5.2

Para a construção das bibliotecas para análise do transcriptoma completo foram utilizados RNA poli A^+ das linhagens HB4a e C5.2 para síntese de cDNA utilizando um oligonucleotídeo dT(18) (Figura 35 – I). A síntese da segunda fita do cDNA ocorreu pelo tratamento com as enzimas *RNase*H, DNA *Polymerase*I, *E.coli* ligase e T4 DNA Polimerase (Figura 35 – II). Em seguida as amostras foram clivadas com a enzima de restrição *Dpn*II, gerando fragmentos menores com extremidades coesivas que permitem a ligação de adaptadores específicos (Figura 35 – III). Após a digestão com a enzima *Dpn*II a amostra de cDNA apresentou tamanho mais homogêneo, concentrado principalmente entre 300 pb e 800 pb (Figura 36). Os adaptadores utilizados contêm uma região de complementaridade ao sítio coesivo gerado (CTAG) seguido por uma região de 4pb específica para cada amostra, denominada *tag*, e uma região de 16pb comum a todos adaptadores (Figura 35 – IV). Após a ligação de adaptadores as amostras foram purificadas e amplificadas por 20 ciclos com enzima de alta fidelidade.

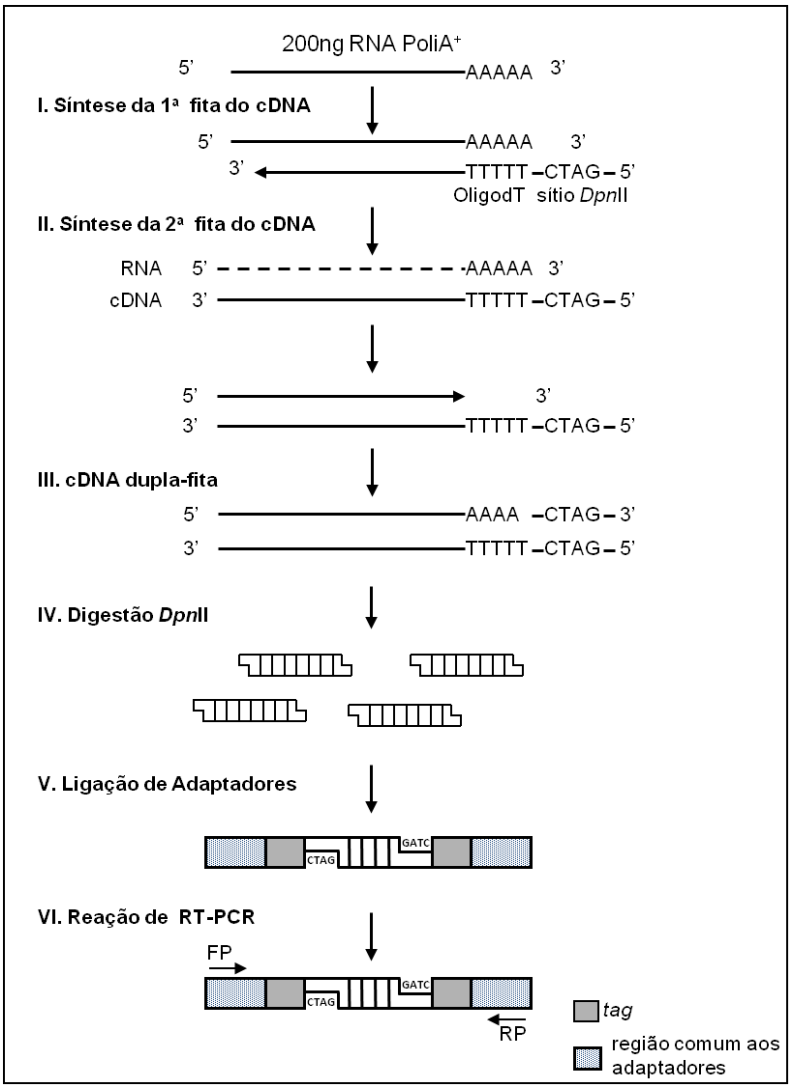


Figura 35: Esquema da metodologia de construção das bibliotecas para análise do transcriptoma completo. I – Síntese da primeria fita de cDNA com oligodT. II – Degradação do RNAm e síntese da segunda fita de cDNA. III – cDNA dupla-fita. IV – Digestão com *Dpn*II formando extremidades coesivas no cDNA. V – Ligação de adaptadores. VI – Amplificação dos fragmentos pela reação de RT-PCR.

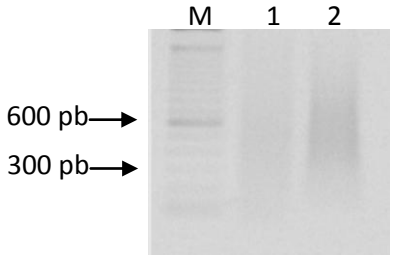


Figura 36: Digestão com a enzima *Dpn*II. M – Marcador 100pb. 1 – cDNA dupla-fita da linhagem HB4a antes da digestão. 2 – cDNA dupla-fita da linhagem HB4a após digestão com *Dpn*II.

Para cada linhagem foram realizadas 5 reações de amplificação e o produto de todas reações foi misturado e quantificado, tendo rendimento de 4,5 µg para a amostra HB4a e 3,7 µg para a amostra C5.2. Cinco por cento do total de cada amostra foi aplicado em gel de agarose (Figura 37). Aproximadamente 25 ng de cada amostra foram utilizados para clonagem e sequênciamento com a metodologia Sanger para validação das bibliotecas previamente ao sequênciamento na plataforma *Genome Sequencer FLX System 454 Roche-Life Sciences*.

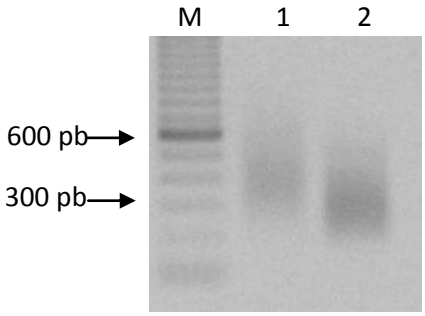


Figura 37: Amplificação por PCR das amostras da biblioteca HB4a (1) e C5.2 (2). Gel de agarose 1%. M – marcador 100bp.

Foram geradas 500 sequências para validação, que foram avaliadas quanto a qualidade e presença das regiões correspondentes aos adaptadores e principalmente as *tags*. Após a validação, 2 µg de produto de PCR purificado de cada biblioteca foram misturados e utilizados para sequênciamento na plataforma *Genome Sequencer FLX System 454 Roche-Life Sciences*.

4.2.2 Análise das sequências geradas pelo sequênciamento em larga escala das bibliotecas das linhagens HB4a e C5.2

No total foram geradas 802.214 sequências da mistura das bibliotecas construídas a partir das linhagens HB4a e C5.2 de tamanho médio de 197 pb. Essas sequências foram depositadas publicamente no *Sequence Read Archive* (SRA) sob número de acesso SRA012436.2.

As sequências foram filtradas para exclusão de RNAs de origem mitocondrial ou ribossomais. Foram também descartadas sequências que não apresentaram as regiões correspondentes aos adaptadores. As 731.628 sequências resultantes destes filtros foram alinhadas contra a sequência do genoma humano, resultando em 651.058 sequências, das quais 614.434 alinharam em uma região única (Figura 38). Em seguida, as sequências foram comparadas com sequências de transcritos de 15 bancos de dados disponíveis pela Universidade de Santa Cruz na Califórnia (UCSC), incluindo o banco de dados de sequências não redundantes (RefSeq). Foram representados no total 11.366 genes distintos, correspondendo a 17.887 transcritos, que representam 40.74% do repertório gênico humano (27.827 genes).

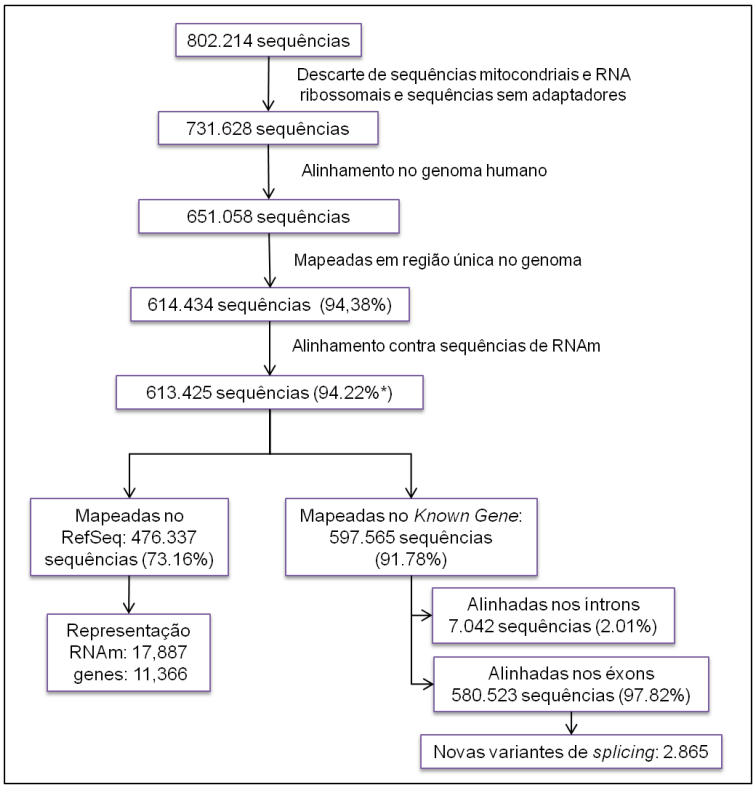


Figura 38: Fluxograma das análises de bioinformática para busca por variantes de *splicing* a partir das bibliotecas para análise de transcriptoma completo.

Em relação a cobertura por transcrito, tanto a região 5' quanto 3' dos transcritos foram bem representadas, apresentando um enriquecimento na porção central dos transcritos (Figura 39A). O fato de ter sido observado boa

coberura da região 5’ independentemente do tamanho dos transcritos representados (Figura 39B), indica que a metodologia utilizada para obtenção do cDNA dupla-fita é apropriada para representação de transcritos completos, sem viés para transcritos de tamanhos menores.

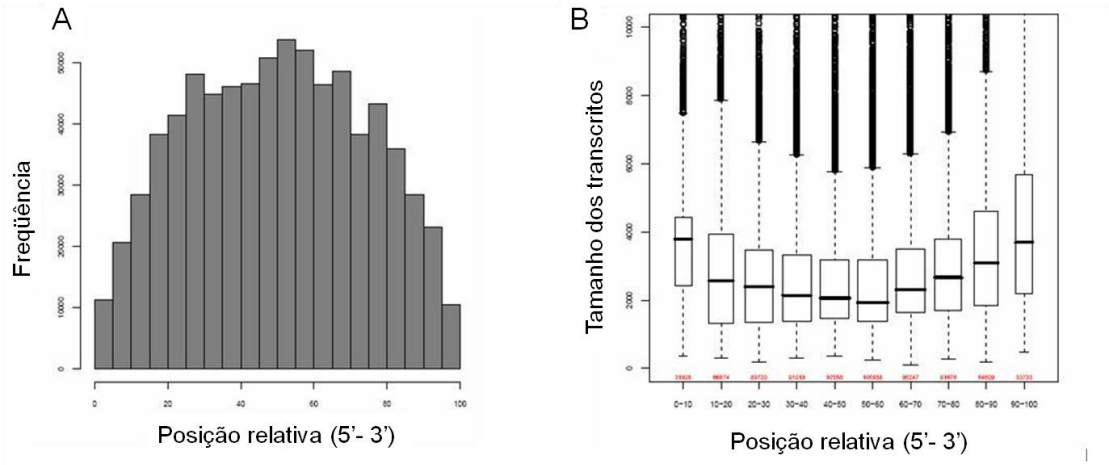


Figura 39: Distribuição relativa das sequências em relação aos transcritos RefSeq. A posição relativa das sequências ao longo dos transcritos é dada em forma proporcional, sendo que 1% corresponde a extremidade 5’ e 100% corresponde a extremidade 3’. A – Posição relativa em relação a frequência. B – Posição relativa em relação aos tamanho dos transcritos.

4.2.3 Identificação de novas variantes de *splicing* das bibliotecas de análise de transcriptoma completo

As 597.565 sequências mapeadas contra o banco de dados *Known Gene* foram alinhadas contra a sequência do genoma humano, utilizando a ferramenta BLAT, e as coordenadas de alinhamento dos limites éxon/íntron foram anotadas. Em seguida, a identificação de putativas novas variantes de *splicing* foi feita pela comparação das coordenadas das sequências das bibliotecas com as coordenadas de alinhamento de transcritos conhecidos.

Foram identificadas 2.875 potenciais novas variantes de *splicing*, considerando a presença de sítios de *splice* conservados. As novas variantes de *splicing* foram categorizadas em uso de sítios de *splice* alternativos, retenção de íntrons ou uso alternativo de éxons. Em relação ao uso alternativo de éxons, foram identificados 487 eventos de exclusão de éxons

(Figura 40A) e 651 eventos de inclusão de éxons, sendo que destes 89 eventos apresentam os dois éxons flanqueadores conhecidos e os demais 562 eventos apresentaram apenas um éxon flanqueador mapeado contra transcritos conhecidos (Figura 40B). Foram identificados 530 eventos de uso de sítio de *splice* alternativo, incluindo sítio doador e aceptor alternativo na mesma categoria (Figura 40C). Por fim foram identificados 1.207 eventos de retenção de íntron, incluindo casos onde o íntron retido estava flanqueado por 2 éxons conhecidos (165 eventos) e 1.042 eventos em que apenas um dos éxons flanqueadores foi mapeado (Figura 40D).

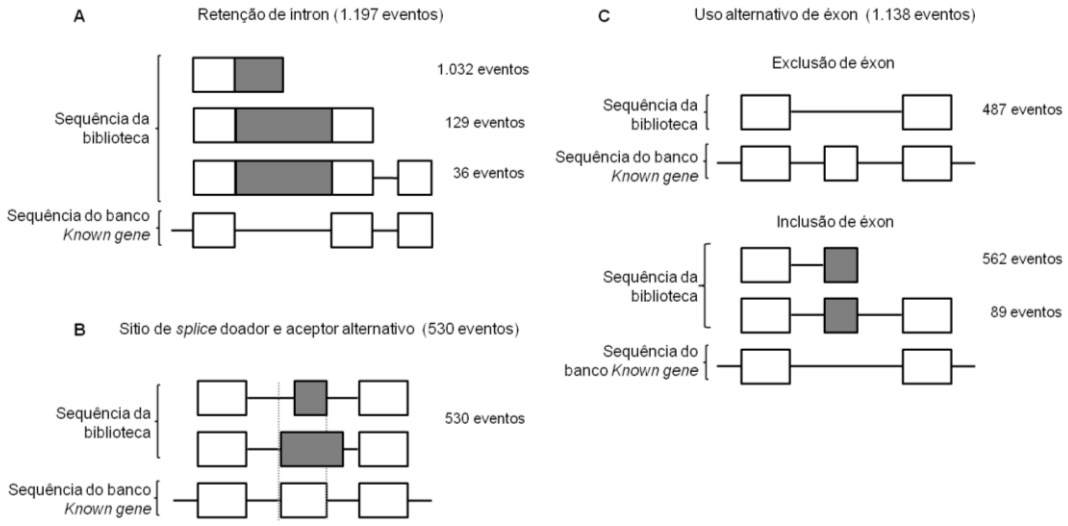


Figura 40: Identificação de novas variantes de *splicing*. As novas variantes de *splicing* estão distribuídas de acordo com o tipo de evento reportado. Os retângulos brancos representam os éxon constitutivos e os retângulos em cinza os éxons alternativos. A – Retenção de íntron com ou sem a identificação de éxons flanqueadores conhecidos. B – Sítios de *splice* doador ou aceptor alternativos. C – Uso alternativo de éxons. Esta classe esta dividida em exclusão de éxons e inclusão de éxons, com ou sem a identificação dos éxons flanqueadores conhecidos.

4.2.4 Validação de eventos de *splicing* alternativo identificados pelas bibliotecas de cDNA de transcriptoma completo

Com o intuito de identificar novas variantes de *splicing* em câncer de mama que possam ser utilizadas como marcadores moleculares ou alvos

terapêuticos foram selecionados 20 eventos para validação por RT-PCR dentre o conjunto de 53 putativas novas variantes que apresentaram a inclusão de um novo éxon e foram reportadas com exclusividade por sequências oriundas da biblioteca da linhagem C5.2 (Tabela 12). Para a reação de RT-PCR os iniciadores foram desenhados no éxon novo e no éxon adjacente e o cDNA da linhagem C5.2 foi usado como molde.

Tabela 12: Caracterização das variantes selecionadas para validação. As variantes estão representadas pelo símbolo dos genes correspondentes. A localização dos éxons em relação a sequência codificante (CDS) ou região não traduzida (5' UTR e 3'UTR). A presença de sequências no banco de dados de sequências expressas EST foi verificada.

GENE	Éxon novo		
	Localização no RNAm	**Mapeado entre os éxons**	**Confirmado por EST**
LRP11	-	1 e 2	não
APEX1	CDS	1 e 3	não
BC039445	-	-	não
CLTC	CDS	25 e 26	sim
CSRP2BP	CDS	8 e 9	não
CUEDC2	5'UTR	1 e 2	não
FLJ00150	CDS	2 e 3	não
FTSJ3	CDS	13 e 14	sim
KIAA1033	CDS	26 e 27	não
KIF2A	CDS	6 e 7	não
NR_002599	-	1 e 2	sim
NR2C1	CDS	2 e 3	sim
PAWR	CDS	4 e 5	sim
PDE6D	CDS	4 e 5	não
PPP2R2A	CDS	4 e 5	sim
PRCC	CDS	5 e 6	não
RPLP1	CDS	1 e 2	não
RPS19	CDS	3 e 4	não
RWDD1	CDS	1 e 2	não
ZNF567	CDS	2 e 3	não

Dezoito dos 20 éxons selecionados foram validados por RT-PCR (90%). A estrutura dos novos éxons com suas respectivas coordenadas genômicas estão detalhadas na figura 41.

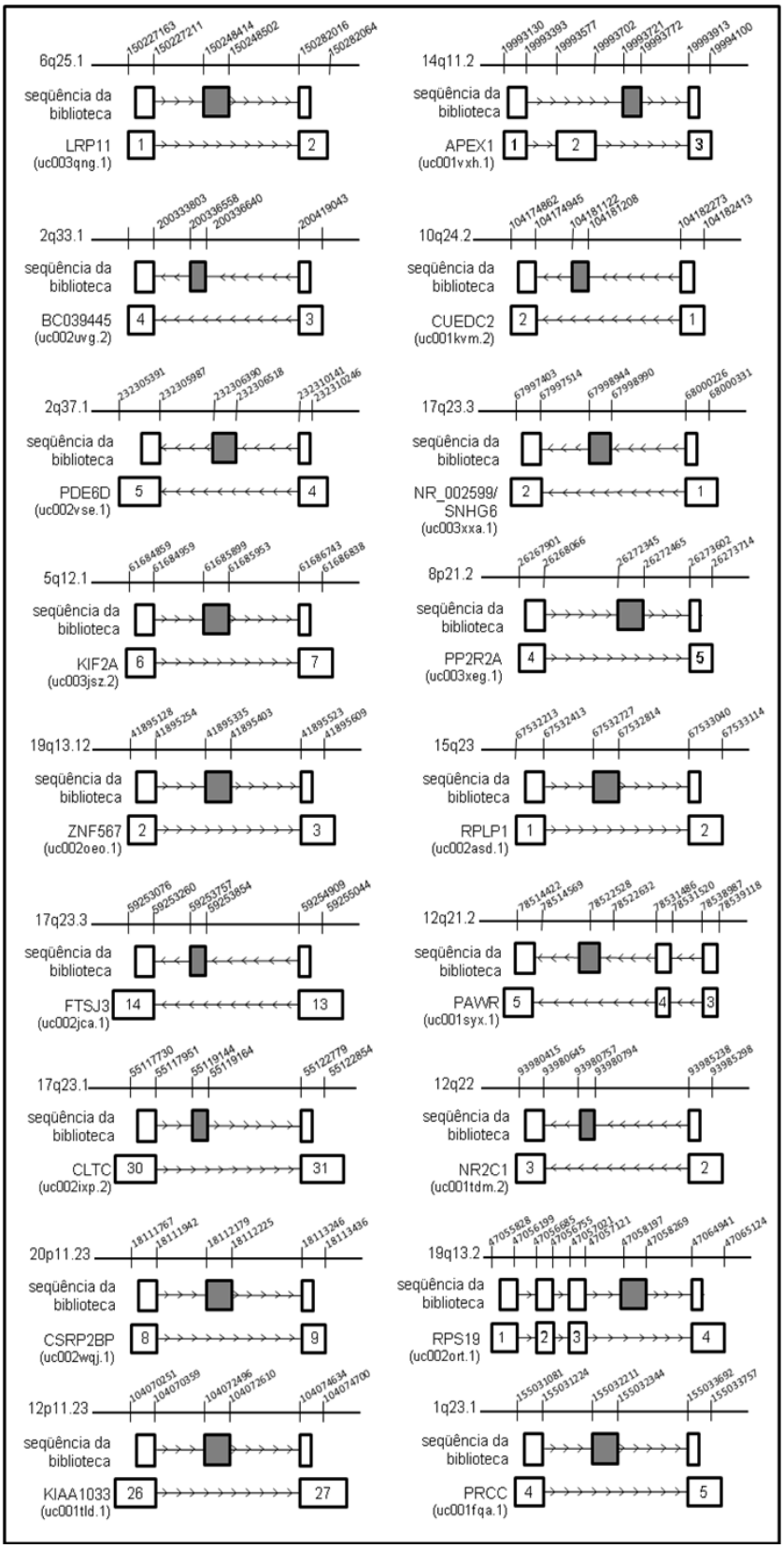


Figura 41: Validação das variantes de *splicing* por RT-PCR. A estrutura gênica das novas variantes e das variantes conhecidas de cada gene estão representadas. As coordenadas genômicas dos limites éxon/íntron foram anotadas. Os retângulos em

branco representam os éxons constitutivos e os retângulos em cinza os éxons alternativos.

Uma vez que 14 dos 18 novos éxons validados estão posicionados na região codificante do gene (CDS) as possíveis alterações na sequência de aminoácidos decorrente da inclusão do novo éxon foram preditas com auxílio da ferramenta ORF*Finder* e a consequente inlfluência nos domínios proteicos preditas com auxílio do banco de predição *InterProScan*.

Metade dos genes apresentou inserção de códon de parada prematuro na variante com a inclusão do novo éxon (*CSRP2BP*, *FTSJ3*, *KIAA1033*, *PAWR*, *PDE6D*, *RPLP1* e *RPS19*), sugerindo provável geração de proteínas truncadas com perdas funcionais. O gene *PRCC* também resultou em provável perda completa de função uma vez que a variante com inserção do éxon não forma sequência aberta de leitura. Outras alterações na sequência aberta de leitura foram identificadas para quatro genes (*APEX1*, *CLTC*, *KIF2A* e *ZNF567*), com a inserção ou deleção de aminoácidos sem, no entanto, resultar em alterações nos domínos proteicos. Por fim, os genes *PP2R2A* e *NR2C1* apresentaram grandes perdas de aminoácido na porção N-terminal, resultando em alterações em alguns domínios proteicos. O gene *PP2R2A* sugere uma possível perda de 107 aminoácidos na região que inclui o domínio N-terminal de proteína serina-treonina fosfatase 2A, subunidade B, e um sítio conservado de subunidade regulatória PR55 da mesma proteína serina-treonina fosfatase além da perda de duas repetições do tipo WD40. O gene *NR2C1* apresenta perda de 177aa na porção N-terminal da proteína com a inserção do novo éxon. Esta perda implica na substituição de domínios proteicos nesta região. A proteína resultante da variante sem o éxon apresenta um domínio do tipo receptor de vitamina D que é substituído por um domínio de receptor de ácido retinóico, além da perda de domínios de receptores hormonais nucleares do tipo dedo de zinco.

4.2.5 Regulação das variantes de *splicing* pela expressão diferencial de *ERBB2*

Para avaliar uma possível influência na regulação na geração de variantes de *splicing* pela expressão diferencial do gene *ERBB2* foram realizadas análises *in silico* e experimentos de RT-PCR quantitativo.

A análise *in silico* avaliou comparativamente o número de novos eventos de *splicing* identificados por cada linhagem celular, HB4a e C5.2. Para esta análise foi utilizado um subgrupo de sequências, dentre as quais não apenas a sequência correta dos adaptadores foi detectada, mas também a sequência correspondente as *tags* específicas das linhagens seguidas das 4 bases correspondentes ao sítio de restrição da enzima *Dpn*II. Este subgrupo contém um total de 410.788 sequências, sendo 222.406 provenientes da amostra C5.2 e 188.382 provenientes da amostra HB4a.

Utilizando este subgrupo, foram identificados 1.927 novos eventos de *splicing* alternativo. Destes, 940 são reportados apenas por sequências da amostra C5.2, 627 são reportados apenas por sequências da biblioteca HB4a e 360 foram identificados por sequênicas das duas linhagens celulares. Após normalizar o número de eventos de cada tipo (retenção de íntron, inclusão de novo éxon, exclusão de éxon e uso alternativo de sítios de *splice*) encontrado pelo número total de sequências geradas para cada biblioteca, foi possível realizar uma comparação entre as linhagens. Foi detectado um enriquecimento de novas variantes com uso alternativo de éxons (tanto inclusão quanto exclusão de éxons) ($p<0,001$) e com o uso alternativo de sítios de *splice* aceptor e doador ($p<0,01$) na amostra C5.2 em relação a amostra HB4a, sugerindo que a alta expressão de *ERBB2* pode modular a regulação do *splicing* alternativo.

Em relação a análise experimental, a diferença de expressão de 8 novos éxons entre as linhagens HB4a e C5.2 foi avaliada por RT-PCR quantitativo. Neste caso, por se tratar de eventos de *splicing* alternativo de inclusão de éxons, o uso da metodologia de RT-PCR quantitativo é apropriada uma vez que há possibilidade da utilização de iniciadores específicos para esta variante. Neste ensaio foram utilizados os mesmos

iniciadores utilizados na etapa de validação dos eventos, onde um dos iniciadores é complementar ao éxon novo, o que resulta em uma quantificação variante específica, e o outro é complementar a um dos éxons adjacentes aumentando a confiabilidade dos resultados de quantificação de transcritos e não de DNA genômico contaminante.

Seis dos 8 éxons avaliados correspondentes aos genes *CSRP2BP*, *PRCC*, *CLTC*, *NR2C1*, *RPS19* e *KIAA1033* apresentaram mais de 2 vezes de diferença de expressão entre as linhagens HB4a e C5.2 após normalização dos dados pelos genes *GAPDH* e *GUSB*.

5. Discussão

A diversidade do repertório transcricional humano derivado do *splicing* alternativo tem sido extensamente investigada. A obtenção de um código de *splicing* será de enorme importância para compreensão de sua influência em doenças humanas complexas como o câncer, além de servir como marcadores moleculares mais acurados.

Neste trabalho, o repertório transcricional gerado pelo *splicing* alternativo de amostras de mama sob a influência da alta expressão do gene *ERBB2* foi investigado através de duas estratégias distintas baseadas na construção de bibliotecas de cDNA. *ERBB2* é um oncogene com papel importante em câncer de mama, uma vez que cerca de 30% desses tumores apresentam alta expressão desse receptor, o que confere maior agressividade e maior taxa de crescimento ao tumor e pior prognóstico à paciente (SLAMON et al., 1987; SLAMON et al., 1989).

As abordagens terapêuticas disponíveis. como o anticorpo monoclonal traztuzumab e a molécula lapatinib (HYNES; LANE, 2005), são extremamente eficientes e apresentam alta taxa de resposta (VOGEL et al., 2002). No entanto, alguns tumores apresentam resistência ao tratamento com piora na sua evolução e os mecanismos que conferem essa resistência não estão bem estabelecidos. Assim, a caracterização das alterações transcricionais mediada pela ativação de *ERBB2*, pode auxiliar na compreensão da biologia de grande percentual dos tumores de mama e pode auxiliar na identificação de marcadores moleculares de prognóstico e de resposta a tratamento. A ativação da sinalização por *ERBB2* pode modular todo o transcriptoma, incluindo a regulação do *splicing* alternativo (MUKHERJI et al., 2006).

Com o objetivo de explorar a variabilidade transcricional gerada pelo *splicing* alternativo diferentes metodologias têm sido aplicadas, sendo o uso de bibliotecas de cDNA uma das alternativas mais promissoras, uma vez que pouco apresentam resultados artefatuais, não dependem de um conhecimento prévio da estrutura dos genes e permitem a identificação de todos os tipos de eventos de *splicing* alternativo com a mesma acurácia.

Neste trabalho foram utilizadas duas estratégias diferentes de construção de bibliotecas de cDNA para avaliação de variantes de *splicing* alternativo. A primeira metodologia proposta foi a construção de bibliotecas de cDNA enriquecidas para *splicing* alternativo, para a qual a metodologia de captura de moléculas de heteroduplexes reportada por Watahiki e colaboradores (2004) foi combinada a metodologia de amplificação de RNAm. A segunda estratégia utilizada foi a construção de bibliotecas de cDNA para avaliação do transcriptoma total, a qual faz uso da grande capacidade de geração de sequências possibilitada pelos novos equipamentos de sequenciamento de nucleotídeos de alto desempenho.

5.1 Biblioteca de cDNA enriquecida para *splicing* alternativo

A metodologia de construção de bibliotecas de cDNA enriquecidas para *splicing* alternativo foi inicialmente proposta por Watahiki e colaboradores em 2004. O trabalho de Watahiki e colaboradores utiliza duas bibliotecas de cDNA completos, como material inicial para a construção da biblioteca enriquecida. A utilização de RNA total, proposta nesse trabalho, não apenas simplificou a metodologia como resultou em uma diminuição de tempo e custos expressivos.

A quantidade de RNA total recuperada de tecido tumoral para estudos moleculares é muitas vezes limitante, devido ao pequeno tamanho de amostras de tecido congelado provenientes de biópsias ou devido a microdissecção a laser. Nestes casos, é necessário o emprego da metodologia de amplificação do RNA mensageiro, para garantir quantidade suficiente de molécula para a realização dos experimentos. Essa metodologia promove uma amplificação linear do RNA mensageiro (RNAm) e evidências mostram que os transcritos são igualmente amplificados, independentemente do seu nível de expressão, quando avaliados por experimentos de microarranjos de DNA (WANG et al., 2000; GOMES et al., 2003; SARAIVA et al., 2005) e RT-PCR (FERREIRA, et al., 2010). A metodologia de

amplificação de RNAm é amplamente utilizada pelo nosso grupo e tem permitido a avaliação global do transcriptoma de amostras tumorais em situações em que a quantidade do RNA total é restrita. Por exemplo, utilizando amplificação de RNAm em experimentos de microarranjos de DNA nosso grupo tem investigado o transcriptoma de regiões individualizadas do tumor de Wilm's, que são compostas por diferentes populações de células e que apontam implicações distintas na clínica e na evolução desse tumor (MASCHIETTO et al., 2008). Em outro estudo, populações homogêneas de células capturadas por microdissecção a laser dos componentes *in situ* e invasivo do câncer de mama também foram investigadas individualmente, proporcionando maior especificidade aos achados. Além disso, o emprego de outras metodologias mais sensíveis para avaliação transcricional, como RT-PCR quantitativo (FERREIRA et al., 2010) e biblioteca subtrativa de cDNA (PINEDA, 2008), combinadas com metodologias de amplificação de RNA tem possibilitado a validação dos achados obtidos por metodologias de grande escala (CASTRO et al., 2008; ROZENCHAN et al, 2009), além da identificação de transcritos raros (PINEDA, 2008), diferencialmente regulados durante a progressão do câncer de mama. Assim, o emprego e a adaptação do método de amplificação de RNAm combinado a outras metodologias de investigação transcricional têm sido cada vez mais factível em nosso grupo. Sua utilização tem aplicação inquestionável, principalmente na área de câncer. Uma vez que os tumores são entidades altamente heterogêneas, avaliações de expressão gênica baseadas em células, ao invés do tumor como um todo, fornecem dados mais acurados em relação às células de interesse.

Dessa forma, nesse presente estudo, como nossa proposta foi o estabelecimento de uma metodologia que possibilitasse a identificação de variantes de *splicing* em amostras tumorais com quantidades restritas de RNA, a utilização da etapa de amplificação do RNAm foi necessária, mesmo utilizando RNA de linhagens celulares no seu estabelecimento. A incorporação da etapa de amplificação do RNAm baseada no oligonucleotídeo *template switch* (TS) (MATZ et al., 1999) e transcrição *in vitro* amplia a potencialidade da utilização dessa estratégia para qualquer

quantidade inicial de RNA total, inclusive provenientes de amostras microdissecadas a laser.

Durante o estabelecimento de nossa metodologia outra abordagem de construção de biblioteca de cDNA enriquecida para *splicing* alternativo foi proposta (ASEtrap), a qual é também baseada no uso de RNA total (THILL et al., 2006) para a síntese de cDNA. Após a síntese de cDNA, Thill e colaboradores utilizaram a estratégia de *SMART PCR* para amplificação do material. A tecnologia *SMART* também é baseada na inserção do oligo *template switch* na extremidade 5' dos transcritos e oligodt na extremidade 3'. No entanto, ao invés da utilização da transcrição *in vitro*, essa estratégia realiza uma amplificação em cadeia da polimerase utilizando iniciadores complementares aos oligos TS e oligodT. Outra diferença importante é o fato de Thill e colaboradores realizaram clivagem enzimática do cDNA previamente a etapa de formação das moléculas de heteroduplexes, ao contrário da nossa estratégia apresentada, na qual a desnaturação e renaturação ocorrem antes da fragmentação. Esse fato pode reduzir a possibilidade de identificação de variantes, em casos em que a região não comum entre as variantes possui sítios de reconhecimento da enzima da clivagem. Além disso, outras pequenas diferenças entre as metodologias podem ser apontadas, como: diferenças nas enzimas de restrição utilizadas (*RsaI*, que gera fragmentos blunt ao invés da *Dpn*II); purificação das estruturas de heteroduplexes por proteínas de ligação a moléculas simples-fita ao invés de oligo randômico biotinilado; e no fato da metodologia de ASEtrap realizar três *rounds* de seleção. No entanto, se essas diferenças metodológicas resultam em dados dostintos, não pode ser avaliado.

As duas bibliotecas de cDNA enriquecidas para *splicing* alternativo, utilizando a linhagem celular C5.2 ou o grupo de amostras tumorais de mama, apresentaram redundância muito próximas de 84,25% e 86,75%, respectivamente. Apesar de serem altamente redundantes quando comparadas a bibliotecas de cDNA para o sequenciamento de ESTs e bibliotecas de cDNA completos, esse valor foi similar ao encontrado na literatura para bibliotecas enriquecidas (75,4% - THILL et al., 2004) e provavelmente são resultantes da etapa de amplificação por PCR utilizada

por ambos os métodos. O valor de redundância um pouco mais elevado, neste trabalho em relação ao da literatura, pode ser devido à diferença na fonte de RNA utilizadas. Thill e colaboradores utilizaram RNA total de tecido de placenta humano, o qual é considerado um tecido com alta diversidade transcricional, provavelmente maior que o tecido mamário humano que foi utilizado no nosso trabalho.

A eficiência na identificação de variantes de *splicing* com o uso de bibliotecas de cDNA enriquecidas para *splicing* alternativo foi verificada pela comparação na capacidade de identificação de variantes entre bibliotecas enriquecidas e bibliotecas de ESTs não enriquecidas, e a estratégia de enriquecimento de heteroduplexes resultou em uma eficiência cerca de 10 vezes maior (THILL, 2006).

Com a aplicação da abordagem de construção de bibliotecas de cDNA enriquecidas para *splicing* alternativo, foram identificados 39 transcritos para os quais uma variante pode ser identificada nos bancos de dados que possa ter participado da formação do heteroduplex. Além desses, 36 transcritos que não possuem variantes conhecidas nos bancos de dados foram identificados. Uma vez que dois transcritos dentre seis dessa segunda classe foram validados, podemos estimar que, ao menos, uma parcela desses 36 transcritos foram resultantes da hibridação com variantes de *splicing* ainda não caracterizadas e não artefatos experimentais.

Em nenhum caso as duas variantes de *splicing* distintas, envolvidas na formação dos heteroduplexes, foram reportadas apenas pelas sequências das bibliotecas, pois as sequências agrupadas no mesmo *cluster* sempre reportaram a mesma variante de *splicing*. Este fato poderia ser devido ao pequeno número de sequências geradas para cada biblioteca. No entanto, uma análise mais detalhada do método e dos dados obtidos nos permitem sugerir que o fato de ter sido utilizado adaptadores de fit-dupla de DNA não fosforilados pode ter prejudicado a representação das duas fitas de cDNA constituintes dos heteroduplexes.

Adaptadores de DNA dupla-fita não fosforilados são amplamente utilizados por diversas metodologias de construção de bibliotecas de cDNA (DIATCHENKO et al., 1996; GURSKAYA et al., 1996). Nestas abordagens,

ocorre o processo denominado *fill-in* durante a etapa de amplificação por PCR, que preenche as extremidades dos transcritos, possibilitando amplificação dos fragmentos (Figura 43). Em maiores detalhes, a ligação dos adaptadores de fita-dupla não fosforilados ocorre de maneira parcial, pois apenas uma das fitas do adaptador, a fita cuja extremidade 3’ liga-se à extremidade 5’ fosforilada do fragmento de cDNA, é incorporada ao fragmento de cDNA através de uma ligação fosfodiéster. A outra fita do adaptador permanece ligada apenas por complementariedade, sendo perdida na etapa de desnaturação inicial da reação de PCR. Assim, no primeiro ciclo da reação de PCR, ocorre a hibridização das regiões de cDNA complementares e a enzima *Taq polimerase* preenche as extremidades 3’ de ambas sequências do cDNA dupla-fita que foram perdidas, possibilitando anelamento dos iniciadores e prosseguimento da reação (Figura 43).

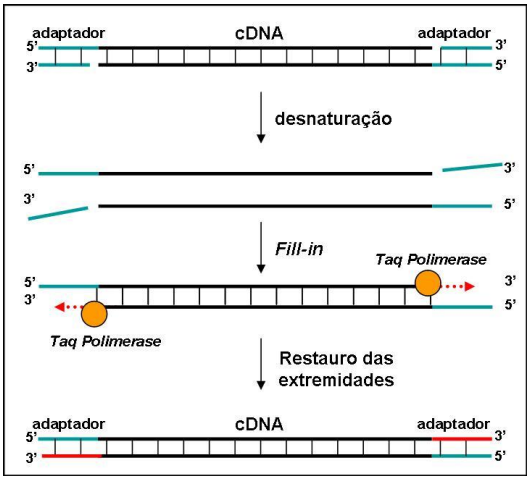


Figura 42: Esquema do processo de *fill-in*. A enzima Taq Polimerase preenche as extremidades do cDNA referentes à sequência do adaptador.

Esse procedimento é considerado eficiente, quando aplicado a moléculas de cDNA de dupla-fita inteiramente complementares. No entanto, no caso da biblioteca enriquecida para *splicing* alternativo as fitas de cDNA não são inteiramente complementares devido à região alternativa, o que pode prejudicar a hibridização no primeiro ciclo da PCR, sendo apenas uma das extremidades do fragmento preenchida pela *Taq polimerase*. Assim apenas uma das fitas do fragmento de cDNA apresenta sequência de adaptadores em ambas as extremidades, necessário para sua amplificação. Uma

alternativa para aumentar a possibilidade de identificação do par de variantes na biblioteca de cDNA enriquecida para *splicing* alternativo seria a utilização de adaptadores fosforilados.

A metodologia de construção de bibliotecas de cDNA enriquecidas para *splicing* alternativo foi aplicada para a linhagem celular de mama C5.2, que apresenta características tumorais devido à transfecção de quatro cópias do oncogene *ERBB2* sob controle do promotor do vírus “long terminal repeat” (MMTVLTR) e sinais de poliadenilação SV40, e para um grupo de cinco amostras tumorais de carcinoma de mama invasivo que apresentam alta expressão de ERBB2 determinada por imunohistoquímica. As linhagens celulares de mama são consideradas um ótimo modelo experimental para o estudo do câncer de mama devido à fácil obtenção, cultivo e manipulação das células. O fato de as células das linhagens serem cultivadas em condições controladas, sem a interferência do meio ou de outros tipos celulares, permite analisar as alterações específicas da célula epitelial. Além disso, possibilitam analisar o papel de genes ou de proteínas específicos, através da observação das alterações celulares decorrentes da inserção ou da inibição de moléculas de interesse, ou ainda verificar o modo de ação de determinadas drogas.

Por outro lado, devido à alta heterogeneidade do câncer de mama, a obtenção de um modelo experimental que recapitule a complexidade dessa doença é extremamente difícil. Nesse sentido, estudos que utilizam tecido tumoral permitem melhor correlação entre os diversos eventos moleculares da célula epitelial e do meio ambiente com o surgimento e a progressão do câncer de mama. Assim, é esperado que alguns comportamentos encontrados nas linhagens sejam divergentes daqueles encontrados nos tumores primários.

Neste trabalho foi encontrada uma sobreposição de 6 a 10% dos transcritos encontrados em as ambas bibliotecas, uma vez que 16 dos 164 clusters com sequências de cDNA e EST dos bancos de dados apresentaram sequências provenientes das duas bibliotecas e 5 dos 79 ASSETs com múltiplos éxons utilizados para busca por variantes de *splicing* foram identificados tanto na linhagem quanto nas amostras tumorais. No entanto,

84,6% dos transcritos avaliados na validação cruzada (11 de 13) foram expressos nas amostras provenientes das duas bibliotecas, sugerindo que este valor de sobreposição não reflete diferenças na regulação de *splicing* entre linhagens e tecido tumoral, mas sim deve ser resultado da pequena cobertura das bibliotecas, isto é, do número de clones sequenciados.

5.2 Biblioteca de cDNA para análise de transcriptoma completo

A estratégia de construção de bibliotecas de cDNA para avaliação do transcriptoma total tornou-se extremamente atrativa, nos últimos 5 anos, com o advento das novas tecnologias de sequenciamento de nucleotídeos de alto desempenho. Esses equipamentos reduziram em muito os custos de cada base sequenciada, diminuíram o tempo de obtenção das sequências e aumentaram enormemente a capacidade de bases sequenciadas por minuto. No mais, uma das suas principais vantagens é a dispensa da etapa de clonagem da biblioteca, a qual foi substituída pela PCR em emulsão ou amplificação clonal em plataforma sólida. Esses avanços possibilitaram investigar o transcriptoma completo das células por estratégias baseadas em sequências, fornecendo informações importantes dos aspectos qualitativos e quantitativos como a identificação de mutações, fusões gênicas, variantes de *splicing*, novos genes e o nível de expressão dos transcritos, em um único experimento. Nesse caso, a ausência de uma estratégia de enriquecimento para variantes de *splicing* é de certa forma suprida pela grande quantidade de sequências geradas, aumentando a cobertura e a possibilidade de identificação de novas variantes, variantes específicas de um tecido e ainda variantes raras.

A investigação em larga escala de variantes de *splicing* tem contribuído para uma melhor caracterização do transcriptoma humano. Dados recentes estimam a ocorrência de *splicing* alternativo em 92% a 95% dos genes humanos com múltiplos éxons (WANG et al., 2009; PAN et al., 2008). Além disso, análises comparativas sugerem que as maiores diferenças na

regulação do *splicing* ocorrem entre diferentes tecidos, ao invés de entre indivíduos (WANG et al., 2009). No mais, essas abordagens têm se mostrado eficientes tanto para identificação de novas variantes como também para a identificação do balanço de expressão das variantes e suas alterações. No entanto, o tamanho das sequências geradas por dois desses novos sequênciadores são um grande obstáculo para o mapeamento de sequências que representam junções entre éxons no genoma humano e necessitam de abordagens bioinformáticas complexas com a perda de informações devido à incapacidade de alinhamento confiável.

Neste trabalho utilizamos a plataforma *Genome Sequencer FLX System 454 Roche-Life Sciences*, a qual fornece sequências de tamanhos maiores em torno de 400 bases, evitando problemas de alinhamento múltiplo. Apesar de ter sido utilizada uma estratégia diferente na construção da biblioteca de cDNA em relação à biblioteca enriquecida nesta etapa do trabalho, o objetivo de investigar o transcriptoma mamário influenciado pela alta expressão do gene *ERBB2* foi mantido. Nesse caso, devido ao grande número de sequências geradas foi possível a identificação de mais de duas mil novas variantes de *splicing*. Além disso, análises comparativas mostraram um enriquecimento de variantes de *splicing* que apresentam uso alternativo de éxons influenciado pela alta expressão do oncogene *ERBB2*.

Análises para a identificação de variantes de *splicing* diferencialmente reguladas entre grupos de amostras é extremamente interessante, uma vez que a literatura mostra que, muito mais do que uma regulação tudo ou nada, a regulação do *splicing* resulta em alterações mais sutis no balanço de expressão entre as variantes. No entanto, análises computacionais para esse fim apresentam alta complexidade, pois devem levar em conta o número de sequências em relação ao tamanho dos transcritos e a cobertura do sequenciamento como também considerar uma distribuição não homogênea das sequências ao longo dos transcritos (HOWARD; HEBER, 2010; TRAPNELL et al., 2010). Apesar do grande aumento no número de sequências geradas nestas bibliotecas, em relação às bibliotecas enriquecidas, a cobertura das variantes não é suficiente para permitir análises de diferença de regulação mais complexas.

Ainda assim, foi possível realizar uma análise comparativa entre os eventos de *splicing* nas duas linhagens sequênciadas que mostrou um enriquecimento das variantes de *splicing* com o uso alternativo de éxons na linhagem C5.2 em relação à linhagem HB4a, sugerindo uma possível influência da modulação da expressão do gene *ERBB2* na regulação do *splicing*. Foram selecionados para validação eventos de *splicing* reportados apenas por sequências da linhagem C5.2 que poderiam ter um padrão de expressão associado a essa linhagem. De fato, seis variantes de *splicing* com inserção de um novo éxon foram mais expressas na linhagem C5.2 em relação à linhagem HB4a pela análise de RT-PCR quantitativo. Para verificar se de fato essas variantes estão mais expressas na linhagem C5.2 e não são resultados da maior expressão do gene como um todo, foi verificado o número total de sequências referentes a cada gene gerado pelo sequenciamento em larga escala das duas linhagens. Apenas um gene (*RPS19*) apresentou maior expressão na linhagem C5.2 em relação a HB4a, outros quatro genes não apresentaram diferença de expressão entre as linhagens e um gene apresentou maior expressão na linhagem HB4a. Portanto, a alta expressão de *ERBB2* parece ter influenciado de forma específica e positiva apenas a nova variante de *splicing* reportada que apresenta a inserção do novo éxon.

Tabela 13: Análise de expressão dos 6 genes nas linhagens HB4a e C5.2. O número de sequências correspondentes a cada gene obtidas pelo sequenciamento em larga escala das linhagens HB4a e C5.2 foram contabilizados. A razão de expressão foi obtida pela divisão do número de sequências da linhagem HB4a em relação ao número de sequências da linhagem C5.2. O sinal negativo indica menor expressão na linhagem C5.2

Gene	Número de sequências na linhagem HB4a	Número de sequências na linhagem C5.2	Razão de Expressão
CLTC	115	111	-1,2
CSRP2BP	27	18	-1,7
NR2C1	7	6	-1,4
RPS19	191	1071	4,8
PRCC	49	48	-1,2
KIAA1033	7	4	-2

5.3 Comparação da eficiência das duas abordagens para construção de bibliotecas de cDNA

Um dos desafios do uso de bibliotecas de cDNA para identificação de variantes de *splicing* é a obtenção de transcritos completos (*full-legth*) que representem as porções 5', 3' e a região codificante dos genes como também a representatividade de variantes de baixa expressão. Nesse trabalho, a obtenção de cDNA dupla-fita a partir de RNA total ou RNA PoliA$^+$ foi realizada através de métodos distintos, no entanto, nos dois casos pudemos observar boa cobertura em relação aos transcritos, com uma preferência na porção central, sem nenhum viés para a região 3'. É especialmente importante ressaltar que, mesmo no caso da utilização da metodologia de amplificação do RNAm, não foi verificada perda na representatividade da porção 5'dos transcritos em relação a porção 3'. Além disso, a representatividade dos transcritos foi independente do tamanho. Em relação à biblioteca enriquecida, a média do tamanho dos transcritos representados foi de 2.836 nucleotídeos, o qual é muito similar ao tamanho médio dos transcritos conhecidos (3.098 nucleotídeos), indicando que não houve preferência na representatividade de transcritos de tamanho pequeno. Em relação à biblioteca de sequenciamento do transcriptoma completo a média do tamanho dos transcritos foi de 2.000 a 4.000 sem detecção de diminuição dessa média na região 5'.

No mais, podemos afirmar que as duas metodologias para construção de bibliotecas de cDNA estabelecidas foram eficientes na identificação de variantes de *splicing* influenciadas pela expressão do gene *ERBB2*. A alta taxa de validação das variantes acima de 90% ressalta a acurácia dos dados gerados e ausência de resultados artefatuais.

Bibliotecas de cDNA enriquecidas para *splicing* alternativo são uma excelente ferramenta para obtenção do catálogo de variantes presentes em determinados tipos celulares. A associação dessa metodologia com o sequenciamento de alto desempenho possibilitaria a investigação profunda e comparativa das espécies de transcritos presentes em diversas amostras de interesse. Por outro lado, a metodologia aplicada para construção de

bibliotecas para análise do transcriptoma completo possibilita além de uma anotação dos transcritos presentes, uma análise quantitativa do padrão de expressão das variantes nos diferentes tecidos. No entanto, exige maior profundidade do sequenciamento e demanda o uso de ferramentas computacionais mais complexas.

5.4 Métodos de avaliação quantitativos de expressão de variantes específicas

A metodologia de RT-PCR quantitativo é considerada o padrão ouro para avaliação da expressão relativa de diferentes genes e micro RNAs em diferentes amostras. No entanto, sua aplicabilidade para detectar diferenças de expressão de variantes de *splicing* específicas depende de um cuidadoso desenho experimental para garantir que o valor de expressão obtido seja referente a uma única variante de interesse. O principal obstáculo é a seleção das regiões específicas de cada variante que permita o desenho de iniciadores para amplificação restrita de uma única variante. Em especial, os eventos de *splicing* do tipo exclusão de éxons, onde apenas uma junção éxon-éxon é exclusiva à variante, e o uso alternativo de sítios doador/aceptor, em que apenas uma pequena porção é única à variante, demandam os maiores cuidados. Apesar de a literatura sugerir a utilização de iniciadores nas junções éxon-éxon como uma solução para determinação de expressão variante-específica (SHULZHENKO et al., 2003), verificamos a ocorrência de amplificação inespecífica de uma variante diferente da variante de interesse, pelo anelamento das quatro bases da extremidade 3’ do iniciador. Portanto, devido ao pareamento parcial do iniciador em variantes distintas, descartamos o uso da metodologia de RT-PCR quantitativo para validação das variantes de *splicing* identificadas pelas bibliotecas de cDNA enriquecidas para *splicing* alternativo, uma vez a maioria das variantes reportavam eventos do tipo uso alternativo de éxons e éxon *skipping*.

Como alternativa ao RT-PCR quantitativo foram utilizadas duas abordagens para quantificação de variantes específicas. A metodologia baseada na ligação de sondas específicas possibilita alta confiabilidade aos resultados, uma vez que as sondas são desenhadas em éxons distintos, em posição limítrofe a cada um deles, requerendo que ocorra antes da amplificação a ligação das duas sondas pela enzima Ligase 65, conferindo alta especificidade ao processo. Utilizando essa metodologia, foi verificado desbalanço na expressão de variantes de um mesmo gene entre duas linhagens que apresentam expressão diferencial de *ERBB2*.

A segunda metodologia proposta consistiu no uso de RT-PCR convencional, associada a eletroforese capilar que é uma técnica capaz de identificar difererentes produtos gerados em uma mesma reação e de informar o tamanho e a quantidade de cada um deles, independentemente. Assim, para a reação de RT-PCR foram utilizados iniciadores simples, complementares a regiões comuns a mais de uma variante. Os produtos da amplificação foram aplicados no equipamento de eletroforese capilar e, em poucos minutos, foi obtida quantificação específica de diferentes variantes de um mesmo gene, comprovando os resultados obtidos anteriormente. Essa é uma ferramenta simples e rápida sendo extremamente interessante para análise de variantes de *splicing*. Essa estratégia foi complementar à utilização da abordagem baseada em sondas específicas, aumentando a confiabilidade dos resultados obtidos.

Por outro lado, a metodologia de RT-PCR quantitativo foi aplicada para investigação da diferença de expressão de variantes de *splicing* das bibliotecas para análise de transcriptoma completo que apresentaram a inclusão de um novo éxon. A existência do novo éxon, único a essas variantes, permitiu o uso de iniciadores específicos para essas variantes e resultaram em dados acurados de diferença de expressão.

Pela exploração de três metodologias diferentes, foram identificados 9 variantes de *splicing* diferencialmente reguladas entre duas linhagens celulares de mesmo conteúdo genético que apresentam apenas diferença de expressão do gene *ERBB2*. Esses resultados sugerem que esse oncogene apresenta influência na regulação do *splicing* alternativo.

5.5 Análise das variantes de *splicing* influenciadas pela expressão diferencial de *ERBB2*

O gene *ERBB2,* também conhecido como *HER-2/neu,* codifica um receptor transmembrânico da família de receptores de tirosina quinase, formada por quatro membros (HYNES; MACDONALD, 2009). Apesar de não apresentar um ligante específico, essa proteína age através da dimerização podendo ativar diferentes vias de sinalização intracelular dependendo da homodimerização ou interação com outros membros da família (heterodímeros) (DOUGALL et al., 1994). No mais, *ERBB2* é um oncogene altamente expresso em 20 a 30% dos tumores de mama associado com pior prognóstico. Amplificação de *ERBB2* nestes tumores é considerada uma alteração causal que desempenha papel central na tumorigênese (DI FIORE et al., 1987). Apesar de ser altamente estudado, os mecanismos envolvidos com a transformação e progressão dos tumores de mama ERBB2 positivos ainda não foram completamente elucidados. Neste trabalho identificamos três genes (*FLNA*, *SFRS9* e *TRIP6*) cujo balanço de expressão entre variantes de *splicing* influenciado pela expressão diferencial de *ERBB2* pela biblioteca de cDNA enriquecida para *splicing* alternativo; e seis variantes (do genes - *CLCT*, *CSRP2BP*, *KIAA1033*, *NR2C1, PRCC* e *RPS19*) com aumento de expressão influenciado pela alta expressão de *ERBB2*, pela biblioteca de análise de transcriptoma completo.

FLNA (ENSG00000196924) é um membro da família de proteínas de ligação a actina envolvida com a organização dos filamentos de actina, principalmente com a formação das junções ortogonais, sendo importante em diversos processos celulares como desenvolvimento embrionário e locomoção celular (STOSSEL et al., 2001). Essa proteína também foi identificada no núcleo, onde interage com o receptor de hormônio andrógeno inibindo a transcrição celular mediada por este receptor (LOY; SIM; YONG, 2003). Para ser direcionada ao núcleo, Filamina A é previamente clivada em dois peptídeos menores, e apenas o peptídeo de tamanho correspondente a 110 kDa é translocado para o núcleo. A fosforilação de um aminoácido serina

na posição 2152 (codificada pelo exon 39) inibe a clivagem da proteína, aumentando os níveis citoplasmáticos de filamina A. Como consequência, pode ocorrer aumento da motilidade celular, bem como a diminuição de repressão transcricional do receptor de andrógeno. Altos níveis de FLNA citoplasmáticos foram associados a tumores de próstata metastáticos enquanto tumores benignos ou localizados apresentaram localização predominantemente nuclear desta proteína (BEDOLLA et al., 2009). A variante identificada, neste trabalho, apresenta perda do éxon 39 que resulta na perda de 41aa entre as posições 2127aa e 2167aa, onde está localizada o aminoácido serina envolvido com a regulação da clivagem da proteína (BEDOLLA et al., 2009). Esta variante apresentou um aumento de 3,5 vezes de expressão na linhagem C5.2 em relação à linhagem HB4a. Se a regulação diferenciada dessa variante está envolvida com as características tumorais da linhagem C5.2 devido à alta expressão de *ERBB2*, não foi verificada.

SFRS9 (ENSG00000111786) codifica uma proteína que age como fator de *splicing* integrante da família de proteínas que se ligam a RNA e são ricas nos resíduos serina/argenina. Essa proteína está envolvida tanto com o *splicing* constitutivo quanto com o *splicing* alternativo, modulando a seleção de sítios de *splice* (SIMARD;CHABOT, 2002). A fosforilação de SFRS9 foi detectada em linhagens celulares de mama e ovário que apresentam alta expressão de *ERBB2*. Além disso, a inibição da atividade de ERBB2 pelo tratamento com anticorpo monoclonal herceptina reduziu significativamente a fosforilação de SFRS9, sugerindo que a atividade dessa e de outras proteínas com domínio de ligação a RNA é regulada pelo receptor (MUKHERJI et al., 2006). Em relação a uma participação em processos neoplásicos, SFRS9 parece favorecer a migração celular, uma vez que uma diminuição de migração após bloqueio dos transcritos de SFRS9 por RNA de interferência foi detectada em linhagens celulares de ovário (MUKHERJI et al., 2006). A predição da sequência proteica, resultante da variante identificada, apresenta redução de 43 aa, no entanto não altera a sequência aberta de leitura e não interrompe domínios proteicos segundo predição da ferramenta *InterProScan* e também não altera o sítio de fosforilação de

SFRS9 por ERBB2. Essa variante apresentou aumento da expressão de quase cinco vezes na linhagem C5.2 em relação à linhagem HB4a.

TRIP6 (ENSG00000087077) codifica uma proteína que interage com o receptor do hormônio de tireóide. Esta proteína está localizada em sítios de adesão focal e ao longo de fibras de actina (YI 1998; WANG et al., 1999). A proteína TRIP6 está envolvida com migração celular induzida pelo ácido lisofosfatídico, através da ligação direta entre os domínios do tipo LIM de TRIP6 e a região carboxi-terminal do receptor LPA2 (XU et al., 2004). Além disso, essa proteína interage com MAGI-1b/PTEN através do domínio PDZ desestabilizando os complexos de junção célula-célula formados pelas proteínas β-catenina e E-caderinas, aumentando a motilidade celular (CHASTRE et al., 2009). A estrutura do novo transcrito identificado foi descrito em detalhes neste trabalho (Figura 28) e a proteína resultante apresenta um códon de parada prematuro, localizado a mais de 50 bases a montante da junção éxon-éxon, o que pode sinalizar que esta variante seja degradada pela via de sinalização NMD (*non-sense mediated decay*), tendo papel mais relacionado com controle transcricional que alteração funcional da proteína. Essa variante não apresentou diferença de nível de expressão entre as linhagens.

O sistema de degradação *non-sense mediated decay* tem como principal função eliminar transcritos potencialmente danosos que apresentem códon de parada prematuro, devido a mutações ou alterações no *splicing*, e que resultariam em proteínas truncadas não funcionais (LEWIS et al., 2003). A degradação de transcritos via NMD mediada pelo *splicing* alternativo é um mecanismo comum que afeta 35% das variantes de *splicing* (GREEN et al., 2003). Durante o processamento do RNAm, o spliceossomo deposita um complexo de cerca de 20 a 24 proteínas chamado complexo de junção de éxon (*exon junction complex* - EJC), em cada um dos limites éxon-éxon, que auxiliam o transporte dos RNAm do núcleo para o citoplasma (LE HIR et al., 2001). No citoplasma, os RNAm são direcionados para os ribossomos para que ocorra a síntese proteica, durante a qual, os ribossomos removem os complexos EJC das moléculas de RNAm. No entanto, se houver um códon de parada 50 nucleotídeos a montante da última junção éxon-éxon, um ou

mais complexos EJC permanecerão acoplados ao RNAm (NAGY; MAQUAT, 1998) e irão recrutar enzimas que promoverão a degradação do transcrito.

O gene *CLTC* (ENSG00000141367) codifica a proteína claritina de cadeia pesada, a qual é um componente importante da porção citoplasmática de organelas intracelulares, chamadas vesículas cobertas (em inglês, *coated vesicles*) ou depressões revestidas (em inglês, *coated pits*) (ROYLE, 2006). Essas organelas especializadas estão envolvidas com tráfego intracelular de receptores e com a endocitose de diversas moléculas. Mais recentemente, essa proteína foi reportada como parte integrante do fuso mitótico e sua depleção foi associada à desregulação da segregação de cromossomos podendo contribuir para o aumento da instabilidade genética encontrada em muitos tumores (ROYLE, 2006). A variante nova identificada, com expressão aumentada na linhagem C5.2, resulta na inserção de 7 aminoácidos na porção N-terminal da proteína, na posição 1612 a 1619, sem alterar os domínios proteicos.

A proteína codificada pelo gene *CSRP2BP* (ENSG00000149474) não possui função bem definida, podendo estar envolvida com a formação de grandes complexos proteicos por interação com a proteína CSRP2. A proteína CSRP2 contém repetições ricas em glicina e dois domínios do tipo LIM, que são motivos de dedos de zinco duplos, encontrados em proteínas que agem como proteínas adaptadoras, que permitem a interação entre duas ou mais proteínas, resultando em grandes complexos proteicos (WEISKIRCHEN; GRESSNER, 2000). A isoforma proteica resultante da inserção de um novo éxon entre os éxon 9 e 10 altera a região C-terminal da proteína com a inserção de um códon de parada prematuro, resultando em uma proteína truncada que possui 121aa a menos que a isoforma selvagem. Portanto, essa variante pode ser reconhecida pela via de degradação NMD (*non-sense mediated decay*), pelos mesmos motivos descritos acima.

O gene *KIAA1033* (ENSG00000136051) codifica uma proteína de função pouco conhecida. Recentemente, foi reportada interação entre KIAA1033 com a molécula adaptadora AP2, envolvida na formação de vesículas revestidas de claritina (SCHMID et al., 2006), e com a proteína WAFL envolvida no transporte de endossomos e na interação com

microtúbulos e fialmentos de actina (PAN et al., 2010). Portanto, KIAA1033 pode ser uma proteína importante, envolvida com endocitose e tráfego intracelular. A variante nova identificada, com expressão aumentada na linhagem C5.2, apresenta a inclusão de um novo éxon que resulta na inserção de um códon de parada prematuro, gerando uma proteína truncada, com perda de 229 aminoácidos, sendo possivelmente degradada pela via NMD (*non-sense mediated decay*).

O gene *NR2C1* (ENSG00000120798) codifica um receptor nuclear hormonal também chamado de receptor nuclear órfão TR2 pela ausência de um ligante conhecido. Esse gene é caracterizado pela presença de um domínio conservado de ligação a DNA, uma região variável e uma porção carboxi-terminal com domínio de interação com ligante, que são domínios tipicamente encontrados nos receptores nucleares de hormônios esteróides e hormônio da tireóide (LEE; LEE; CHANG, 2002). Na presença de ligantes, esses receptores atuam como fatores de transcrição regulando a expressão de diversos genes. A variante nova encontrada nesse trabalho apresenta a inclusão de um éxon na região codificante referente à porção N-terminal da proteína (entre os éxons 2 e 3), sugerindo que a variante pode influenciar na regulação da expressão gênica. A porção N-terminal dos receptores nucleares é altamente variável em sequência de aminoácidos e tamanho e esta relacionada com a interação com outras proteínas ativadoras da expressão gênica (LEE; LEE; CHANG, 2002).

O gene *PRCC* (ENSG00000143294) ou *papillary renal cell carcinoma* tem sido implicado com o surgimento de um subtipo de carcinoma renal devido a uma fusão que esse gene sofre com o gene *TFE3* (MEDENDORP et al., 2009). A região N-terminal da proteína PRCC está associada ao aumento da capacidade de trasativação da proteína PRCCTFE3 fusionada quando comparada à proteína TFE3 selvagem (WETERMAN et al., 2001a). Essa região também é importante para interação de PRCC com a proteína MAD2B, a qual é membro da família de proteínas envolvidas com pontos de checagem do ciclo celular (WETERMAN et al., 2001B). A variante nova identificada, com expressão aumentada na linhagem C5.2, apresenta uma inserção de um códon de parada prematuro que diminui a proteína em 36

aminoácidos, o que pode sinalizar que essa variante seja degradada pela via de sinalização NMD (*non-sense mediated decay*), pelos mesmos motivos citados acima.

O gene *RPS19* codifica uma proteína integrante da subunidade 40S do ribossomo. Mutações nesse gene foram associados com à causa de anemia do tipo *Diamond-Blackfan* (CAMPAGNOLI et al., 2008), caracterizada pela ausência ou baixa quantidade de precursores de eritróides, sugerindo uma possível função extra ribossômica para esse gene. A variante nova identificada reporta a inclusão de um novo éxon, com expressão aumentada na linhagem C5.2, que resulta na inserção de um códon de parada prematuro, gerando uma proteína truncada de apenas 75 aminoácidos, 70 aminoácidos a menos que a proteína selvagem. O códon de parada prematuro pode sinalizar a degradação da variante pela via NMD.

Como visto, cinco variantes (55%) apresentaram inserção de um códon de parada prematuro no transcrito, que podem ser alvos de degradação pela via *non-sense mediated decay*, inibindo a produção de proteínas truncadas. No entanto, pelo fato de não termos avaliado os transcritos completos e os seus respectivos quadros abertos de leitura, não podemos assegurar que as alterações geradas pelas variantes de *splicing* levem a inserção de códons de parada prematuro.

As diferentes abordagens experimentais e computacionais descritas neste trabalho se mostraram eficientes para explorar diversidade transcricional resultante de *splicing* alternativo. O número de novas variantes de *splicing* identificadas na abordagem de sequenciamento global sugere que, apesar de grandes progressos obtidos nas últimas décadas, ainda estamos longe de ter uma definição completa do padrão de *splicing* alternativo e sua regulação nos diversos contextos biológicos, sendo portanto, necessário o desenvolvimento de metodologias para exploração da diversidade transcricional resultante desse processo. A determinação da modulação do nível de expressão do transcrito individual e não do gene como um todo, poderá contribuir não somente para a melhor compreensão da biologia dos tumores, mas principalmente para a identificação de marcadores moleculares mais precisos.

6. Conclusões

1) A metodologia de construção de bibliotecas de cDNA enriquecidas para *splicing* alternativo, combinada à amplificação do RNA foi estabelecida, permitindo a identificação de variantes de *splicing* alternativo com alta taxa de validação (94,4%), evidenciando a eficiência do método.

2) A exploração de dados de sequenciamento em larga escala de bibliotecas de cDNA para análise global do transcriptoma permitiu identificação de novos eventos de *splicing* alternativo, com alta taxa de validação (90%).

3) A identificação de novas variantes de *splicing* pelas duas abordagens utilizadas, sugere que o repertório de variantes de *splicing* não está totalmente definido.

4) A expressão diferencial de *ERBB2* influencia a regulação do *splicing* alternativo, alterando o nível de expressão e/ou o balanço de expressão das variantes.

Referências Bibliográficas

ADAMS, M.D. et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. **Science**, v. 252, p.1651-1656, 1991.

AMREIN, H., GORMAN, M., NÖTHIGER, R. The sex-determining gene tra-2 of Drosophila encodes a putative RNA binding protein. **Cell**, v. 55, p.1025-1035, 1988.

ANDERSON, E. The role of oestrogen and progesterone receptors in human mammary development and tumorigenesis. **Breast Cancer Research**, v. 4, p.197-201, 2002.

AST, G. How did alternative splicing evolve? **Nature Review Genetics**, v. 5, p.773-782, 2004.

BARASH, Y. et al. Deciphering the splicing code. **Nature**, v. 465, p. 53-59, 2010.

BASELGA, J., et al. Phase II study of weekly intravenous recombinant humanized anti-p185HER2 monoclonal antibody in patients with HER2/neu-overexpressing metastatic breast cancer. **Journal of Clinical Oncology**, v. 14, p. 737-744, 1996.

BEDOLLA, R. G. et al. Nuclear versus cytoplasmic localization of filamin A in prostate cancer: immunohistochemical correlation with metastases. **Clinical Cancer Research**, v. 15, p.788-796, 2009.

BENNETT, S. T. et al. Toward the 1,000 dollars human genome. **Pharmacogenomics**, v. 6, p. 373-382, 2005.

Benson, D.A. et al. GenBank. **Nucleic Acids Research**, v. 36, p. 25-30, 2008

BERNS, K. et al. A functional genetic approach identifies the PI3K pathway as a major determinant of trastuzumab resistance in breast cancer. **Cancer Cell**, v. 12, p. 395-402, 2007.

BLACK , D. L. Mechanisms of alternative pre-messenger RNA splicing. **Annual Review of Biochemistry**, v. 72, p. 291-336, 2003.

BLENCOWE, B. J. Alternative splicing: new insights from global analyses. **Cell**, v. 126, p. 37-47, 2006.

BOISE, L.H. et al. bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. **Cell**, v. 74, p. 597-608, 1993.

BRETT, D. et al. EST comparison indicates 38% of human RNAms contain possible alternative splice forms. **FEBS Letters**, v. 474, p. 83-86, 2000.

BRINKMAN, B. M. N. Splice variants as cancer biomarkers. **Clinical Biochemistry**, v. 37, p.584-594, 2004.

BOGUSKI, M. S., LOWE, T. M., TOLSTOSHEV, C. M. dbEST--database for "expressed sequence tags". **Nature Genetics**, v. 4, p. 332-333, 1993.

BOTTILLO, I. et al. Functional analysis of splicing mutations in exon 7 of NF1 gene. **BMC Medical Genetics**, v. 8, p.4, 2007.

BURKE, J. et al. Alternative gene form discovery and candidate gene selection from gene indexing projects. **Genome Research**, vol. 8, p. 276-290, 1998.

BUTT, S. Parity and age at first childbirth in relation to the risk of different breast cancer subgroups. **International Journal Cancer**, v. 125, p.1926-1934, 2009.

C. ELEGANS SEQUENCING CONSORTIUM. Genome sequence of the nematode C. elegans: a platform for investigating biology. **Science**, v. 282, p. 2012-2018, 1998.

CAMPAGNOLI, M. F. et al. RPS19 mutations in patients with Diamond-Blackfan anemia. **Human Mutation**, v. 29, p. 911-920, 2008.

CAPUTI, M. et al. hnRNP A/B proteins are required for inhibition of HIV-1 pre-RNAm splicing. **The Embo Journal**, v. 18, p. 4060-4067, 1999.

CAREY, L. A. Through a glass darkly: advances in understanding breast cancer biology, 2000-2010. **Clinical Breast Cancer**, v. 10, p. 188-195, 2010.

CARNINCI, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. **Nature Genetics**, v. 38, p. 626-635, 2006.

CARTEGNI, L. et al, ESEfinder: a web resource to identify exonic splicing enhancers. **Nucleic Acid Research**,v. 31, p. 3568-3571, 2003.

CASTRO, N. P. Evidence that molecular changes in cells occur before morphological alterations during the progression of breast ductal carcinoma. **Breast Cancer Research**, v. 10, R87, 2008.

CHASTRE, E. et al. TRIP6, a novel molécular partner of the MAGI-1 scaffolding molecule, promotes invasiveness. **The FASEB Journal**, v. 23, p. 916-928, 2009.

CHEN, F. C. et al. Identification and evolutionary analysis of novel exons and alternative splicing events using cross-species EST-to-genome comparisons in human, mouse and rat. **BMC Bioinformatics**, v. 7, p.136, 2006.

COBLEIGH, M.A. et al. Multinational study of the efficacy and safety of humanized anti-HER2 monoclonal antibody in women who have HER2-overexpressing metastatic breast cancer that has progressed after chemotherapy for metastatic disease. **Journal of Clinical Oncology**, v. 17, p. 2639-2648, 1999.

CUPERLOVIC-CULF, M. et al. Microarray analysis of alternative splicing. **OMICS**, v. 10, p. 344-357, 2006.

DA SILVA, S. D. et al. Clinicopathological significance of ubiquitin-specific protease 2a (USP2a), fatty acid synthase (FASN), and ErbB2 expression in oral squamous cell carcinomas. **Oral Oncology**, v. 45, p.134-139, 2009.

DEL GATTO-KONCZAK, F. et al. hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer. **Molecular and Cellular Biology**, v. 19, p. 251-260, 1999.

DI FIORE, P. P. et al. erbB-2 is a potent oncogene when overexpressed in NIH/3T3 cells. **Science**, v. 237, p. 178-182, 1987.

DIATCHENKO, L. et al. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. **Procedings of the National Academy of Sciencies of the U S A.**, v. 93, p. 6025-6030, 1996.

DIFEO, A.; NARLA, G.; MARTIGNETTI, J.A. The role of KLF6 and its splice variants in cancer therapy. **Drug Resistance Updates**, v. 12, p. 1–7, 2009.

DOS SANTOS, M. L. et al. Transcriptome characterization of human mammary cell lines expressing different levels of ERBB2 by serial analysis of gene expression. **International Journal of Oncology**, v. 28, p. 1441-1461, 2006.

DOUGALL, W. C. et al. The neu-oncogene: signal transduction pathways, transformation mechanisms and evolving therapies. **Oncogene**, v. 9, p. 2109-2123, 1994.

DRALYUK, I. et al. ASDB: database of alternatively spliced genes. **Nucleic Acids Research**, v. 28, p. 296-297, 2000.

DU, L., GATTI, R. A. Progress toward therapy with antisense-mediated splicing modulation. **Current Opinion in Molecular Therapeutics**, v.11, p. 116-123, 2009.

DUFFY, M.J. Predictive markers in breast and other cancers: a review. **Clinical Chemistry**, v. 51, p. 494-503, 2005.

EISENBERG, A. L. M.; KOIFMAN, S. Câncer de Mama : Marcadores Tumorais (revisão de literatura) 2001. **Revista Brasileira de Cancerologia**, v. 47, p. 377-388, 2001.

ELSTON, C. W.; ELLIS, I. O. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. **Histopathology**, v. 19, p. 403-410, 1991.

ERICKSON, J. W.; QUINTERO, J. J. Indirect effects of ploidy suggest X chromosome dose, not the X:A ratio, signals sex in Drosophila. **PLoS Biology**, v. 5, e332, 2007.

EWING, B. et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Research**, v. 8, p. 175-185, 1998.

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. **Genome Research**, v.8, p. 186-194, 1998.

FAN, W. et al. A statistical mehod for predicting splice variants between two groups of samples using GeneChip expression array data. **Theoretical Biology and Medical Modelling**, v.3, p. 19 – 27, 2006.

FERREIRA, E. N. et al. Heteroduplex formation and S1 digestion for mapping alternative splicing sites. **Genetics and Molecular Research**, v. 7, p. 958-969, 2008.

FERREIRA, E. N., et al. Evaluation of quantitative rt-PCR using nonamplified and amplified RNA. **Diagnostic and Molecular Pathology**, v. 19, p. 45-53, 2010.

FERREIRA, E. N., et al. Alternative splicing: a bioinformatics perspective. **Molecular Biosystems**, v. 3, p. 473-477, 2007.

FETTIPLACE, R.; FUCHS, P. A. Mechanisms of hair cell tuning. **Annual Review of Physiology**, v. 61, p. 809-834, 1999.

FLOREA, L. et al. A computer program for aligning a cDNA sequence with a genomic DNA sequence. **Genome Research**, v. 9, p. 967-974, 1998.

FRASOR, J. et al. Profiling of estrogen up- and down-regulated gene expression in human breast cancer cells: insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype. **Endocrinology,** v. 144, p. 4562-4574, 2003.

FU, X. D. Towards a splicing code. **Cell**, v. 119, p. 736-738, 2004.

GAIL, M. H. et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. **Journal of the National Cancer Institute**, v.81, p. 1879-1886, 1989.

GARDINA , P.J. et al. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. **BMC Genomics**, v. 7, p. 325, 2006.

GILBERT, W. Why genes in pieces? **Nature**, v.271, p.501, 1978.

GLISIN, V.; CRKVENJAKOV, R.; BYUS, C. Ribonucleic acid isolated by cesium chloride centrifugation. **Biochemistry**, v. 13, p. 2633-2637, 1974.

GORLOV, I. P., SAUNDERS, G. F. A method for isolating alternatively spliced isoforms: isolation of murine Pax6 isoforms. **Analytical Biochemistry**, vol. 308, p. 401-404, 2002.

GRAVELEY, B.R. Sorting out the complexity of SR protein functions. **RNA**, v. 6, p. 1197-1211, 2000.

GRAVELEY, B.R. Alternative splicing: increasing diversity in the proteomic world. **Trends in Genetics**, v. 17, p. 100-107, 2001.

GREEN, R. E. et al. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. **Bioinformatics**, v. 19, p. 118-121, 2003.

GOMES, L.I. et al. Comparative analysis of amplified and nonamplified RNA for hybridization in cDNA microarray. **Analytical Biochemistry**, v. 321, p. 244-251, 2003.

GUPTA, S. et al. Genome wide identification and classification of alternative splicing based on EST data. **Bioinformatics**, v. 20, p. 2579-2585, 2004.

GURSKAYA, N. G. et al. Equalizing cDNA subtraction based on selective suppression of polymerase chain reaction: cloning of Jurkat cell transcripts induced by phytohemaglutinin and phorbol 12-myristate 13-acetate. **Analytical Biochemistry**, v. 240, p. 90-97, 1996.

GUSTERSON, B.A. et al. Prognostic importance of c-erbB-2 expression in breast cancer. **Journal of Clinical Oncology**, v. 10, p.:1049-1056, 1992.

HAILE, R. W. et al. BRCA1 and BRCA2 mutation carriers, oral contraceptive use, and breast cancer before age 50. **Cancer Epidemiology, Biomarkers & Prevention**, v. 15, p. 1863-1870, 2006.

HARRIS, L. et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. **Journal of Clinical Oncology**, v. 25, p. 5287-5312, 2007.

HARRIS, R.A. et al. New model of ErbB-2 over-expression in human mammary luminal epithelial cells. **International Journal of Cancer**, v. 80, p. 477-484, 1999.

HARROW, J. et al. GENCODE: producing a reference annotation for ENCODE. **Genome Biology**, v. 7, p. 1-9, 2006.

HARTMUTH, K. et al. Protein composítion of human prespliceossomes isolated by a tobramycin affinity-selection method. **Procedings of the National Academy of Sciencies of the U S A.**, v. 99,p. 16719-16724, 2002.

HEINZEN, E. L. et al. Alternative ion channel splicing in mesial temporal lobe epilepsy and Alzheimer's disease. **Genome Biology**, v. 8, R32, 2007.

HOLSTE, D. et al. Hollywood: a comparative relational database of alternative splicing. **Nucleic Acids Research**, v. 34, D56-D62, 2006.

HOSHIJIMA, K. et al. Control of doublesex alternative splicing by transformer and transformer-2 in Drosophila. **Science**, v. 252, p. 833-836, 1991.

HOWARD, B. E.; HEBER, S. Towards reliable isoform quantification using RNA-SEQ data. **BMC Bioinformatics**, v. 11, p. 3-6, 2010.

HSU, F. et al. The UCSC Known Genes. **Bioinformatics**, v. 22, p. 1036-1046, 2006.

HSU, F. R. et al. AVATAR: a database for genome-wide alternative splicing event detection using large scale ESTs and mRNAs. **Bioinformation**, v.1, p. 16-18, 2005.

HU, G.K. et al. Predicting splice variant from chip expressiond data. **Genome Research**, v. 11, p. 1237-1245, 2001.

HUNTER, S. et al. InterPro: the integrative protein signature database. **Nucleic Acids Research**, v. 37, p. 224-228, 2009.

HWANG, E. S. et al. Patterns of chromosomal alterations in breast ductal carcinoma in situ. **Clinical Cancer Research**, v. 10, p. 5160-5167, 2004.

HYMAN, B. T.; AUGUSTINACK, J. C.; INGELSSON, M. Transcriptional and conformational changes of the tau molecule in Alzheimer's disease. **Biochimica et Biophysica Acta**, v. 17, p. 150–157, 2005.

HYNES, N. E.; MACDONALD, G. ErbB receptors and signaling pathways in cancer. **Current Opinion in Cell Biology**, v. 21, p. 1-8, 2009.

HYNES, N. E.; LANE, H. A. ERBB receptors and cancer: the complexity of targeted inhibitors. **Nature Review Cancer**, v. 5, p. 341-354, 2005.

IZQUIERDO, J. M. et al. Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. **Molecular Cell**, v.19, p. 475-484, 2005.

JIANG, Z. et al. Aberrant splicing of tau pre-RNAm caused by intronic mutations associated with the inherited dementia frontotemporal dementia with Parkinson linked cromossome 17. **Molecular and Cellular Biology**, v. 20, p. 4036 – 4048, 2000.

JOHNSON, J. M. et al. Genome-wide survey of human alternative pre-RNAm splicing with exon junction microarrays. **Science**, v. 302, p. 2141-2144, 2003.

KALNINA, Z. et al. Alterations of pre-RNAm splicing in cancer. **Gene Chromossomes Cancer**, v. 42, p. 342-357, 2005.

KAN, Z.et al. Evolutionarily conserved and diverged alternative splicing events show different expression and functional profiles. **Nucleic Acids Research**, v. 33, p. 5659-5666, 2005.

KAN, Z.; ROUCHKA, E. C.; GISH, W. R. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. **Genome Research**, v. 5, p. 889-900, 2001.

KENT, W. J. BLAT--the BLAST-like alignment tool. **Genome Research**, v. 12, p. 656-664, 2002.

KHOURY, M. P.; BOURDON, J. C. The isoforms of the p53 protein. **Cold Spring Harbor Perspectives in Biology**, v. 2, a000927, 2010.

KIM, N. et al. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. **Nucleic Acids Research**, v.35, p.D1–D6, 2007.

KIM, N.; SHIN, S.; LEE, S. ASmodeler: gene modeling of alternative splicing from genomic alignment of RNAm, EST and protein sequences. **Nucleic Acids Research**, v. 32, p.181–186, 2004.

KIM, E.; MAGEN, A.; AST, A. Different levels of alternative splicing among eukaryotes. **Nucleic Acids Research**, v. 35, p. 125–131, 2007.

KIRSCHBAUM-SLAGER, N., et al. Splicing factors are differentially expressed in tumors. **Genetics and Molecular Research**, v. 3, p. 512-520, 2004.

KIRSCHBAUM-SLAGER, N., et al. Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. **Physiological Genomics**, v. 21, p. 423-432, 2005.

KONECNY, G.E. et al. Association between HER-2/neu and vascular endothelial growth factor expression predicts clinical outcome in primary breast cancer patients. **Clinical Cancer Research**, v.10, p. 1706-1716, 2004.

KORNBLIHTT, A. R. Chromatin, transcript elongation and alternative splicing. **Nature Structural & Molecular Biology**, v. 13, p. 5-7, 2006.

KRAMER, A. The structure and function of proteins involved in mammalian pre-RNAm splicing. **Annual Review of Biochemistry**, v. 65, p. 367-409, 1996.

KRAWCZAK, M.; REISS, J.; COOPER, D. N. The mutational spectrum of single base-pair substitutions in RNAm splice junctions of human genes: causes and consequences. **Human Genetics**, v. 90, p. 41-54, 1992.

LAM, B. J.; HERTEL, K. J. A general role for splicing enhancers in exon definition. **RNA**, v. 8, p. 1233-1241, 2002.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, p. 860-921, 2001.

LAUFER, S. D.; RESTLE, T. Peptide-mediated cellular delivery of oligonucleotide-based therapeutics in vitro: quantitative evaluation of overall efficacy employing easy to handle reporter systems. **Current Pharmaceutical Design**, v.14, p. 3637-3655, 2008.

LE HIR, H. Et al. The exon-exon junction complexprovides a binding platform for factors involved in mRNA export andnonsense-mediated mRNA decay. **The EMBO Journal**, v. 20, p. 4987-4997, 2001.

LEE, Y. F.; LEE, H. J.; CHANG, C. Recent advances in the TR2 and TR4 orphan receptors of the nuclear receptor superfamily. **The Journal of Steroid Biochemistry Molecular Biology**, v. 81, p. 291-308, 2002.

LEWIS, B.P., GREEN, R.E. E BRENNER, S.E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated RNAm decay in humans. Proc. Natl Acad. Sci. USA, 2003: 189–192.

LODISH, H. et al. RNA processing, nuclear transport, and post-transcriptional control. Em **Molecular Cell Biology** – 3a edição, p. 415-420, 2000.

LÓPEZ-BIGAS, N. et al. Are splicing mutations the most frequent cause of hereditary disease? **FEBS Letters**, v.579, p. 1900-1903, 2005.

LOY, C. J.; SIM, K. S.; YONG E. L. Filamin-A fragment localizes to the nucleus to regulate androgen receptor and coactivator functions. **Procedings of the National Academy of Sciencies of the U S A.**, v. 100, p. 4562-4567, 2003.

LÜTZEN, A. et al. Functional analysis of HNPCC-related missense mutations in MSH2. **Mutation Research**. 2008 Oct 14;645(1-2):44-55.

Ma, X. J. et al. Gene expression profiles of human breast cancer progression. **Procedings of the National Academy of Sciencies of the U S A.**, v. 100, p. 5974-5979, 2003.

MACKAY, A. et al. cDNA microarray analysis of genes associated with ERBB2 (HER2/neu) overexpression in human mammary luminal epithelial cells. **Oncogene**, v. 22, p. 2680-2688, 2003.

MAERE, S.; HEYMANS, K.; KUIPER, M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. **Bioinformatics**, v. 21, p. 3448-3449, 2005.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature**, v. 437, p. 376–380, 2005.

MASCHIETTO, M. et al. Molecular role for the Wnt signaling pathway in kidney and tumor development. **Oncology**, v. 75, p. 81-91, 2008.

MATLIN, A. J.; CLARK, F.; SMITH, C. W. Understanding alternative splicing: towards a cellular code. **Nature Reviews Molecular Cell Biology**, v. 6, p. 386-398, 2005.

MATTER, N.; HERRLICH, P.; KÖNIG, H. Signal-dependent regulation of splicing via phosphorylation of Sam68. **Nature**, v. 420, p. 691-695, 2002.

MATTICK, J. S.; MAKUNIN, I. V. Non-coding RNA. **Human Molecular Genetics**, v. 15, p. R17–R29, 2006

MATZ, M. et al. Amplification of cDNA ends based on template-switching effect and step-out PCR. **Nucleic Acids Research**, v. 27, p. 1558-1560, 1999.

MCDONNELL, D. P. et al. Analysis of estrogen receptor function in vitro reveals three distinct classes of antiestrogens. **Molecular Endocrinology**, v. 9, p. 659–669, 1995.

MCPHERSON, K.; STEEL, C. M.; DIXON, J. M. ABC of breast diseases. Breast cancer—epidemiology, risk factors, and genetics. **BMJ**, v.321, p. 624-628, 2000.

MEDENDORP, K. et al. The renal cell carcinoma-associated oncogenic fusion protein PRCCTFE3 provokes p21 WAF1/CIP1-mediated cell cycle delay. **Experimental Cell Research**, v.315, p. 2399-2409, 2009.

MERCATANTE, D. R. et al. Modification of alternative splicing of Bcl-x pre-mRNA in prostate and breast cancer cells. analysis of apoptosis and cell death. **The Journal of Biological Chemistry**, v. 276, p. 16411-16417, 2001.

MIRONOV, A. A.; FICKETT, J. W.; GELFAND, M. S. Frequent alternative splicing of human genes. **Genome Research**, v. 12, p. 1288-1293,1999.
MIRZA, M. et al. Osteopontin-c is a selective marker of breast cancer. **International Journal of Cancer**, v. 122, p. 889-897, 2008.

MODREK, B. et al. Genome-wide detection of alternative splicing in expressed sequences of human genes. **Nucleic Acids Research**, v. 29, p. 2850-2859, 2001.

MODREK, B; Lee, C. J. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. **Nature Genetics**, v. 34, p. 177-180, 2003.

MOLINA, R. et al. Tumor markers in breast cancer- European Group on Tumor Markers recommendations. **Tumour Biology**, v. 26, p. 281-293, 2005.

MUKHERJI, M. A phosphoproteomic analysis of the ErbB2 receptor tyrosine kinase signaling pathways. **Biochemistry**, v. 45, p. 15529-15540, 2006.

NAGY, E.; MAQUAT, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. **Trends in Biochemical Science**, v. 23, p. 198-199, 1998.

NAOR, D., Nedvetzki, S., Golan, I., Melnik, L., Faitelson, Y. CD44 in cancer. **Critical Reviews in Clinical Laboratory Science**, v. 39,p. 527-579, 2002.

NILSEN, T. W. The spliceossome: the most complex macromolecular machine in the cell? **Bio Essays**, v. 25, p. 1147- 1149, 2003.

PAIK, S. et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. **New England Journal of Medicine**, v. 351, p. 2817-2826, 2004.

PAJARES, M. J. et al. Alternative splicing: an emerging topic in molecular and clinical oncology. **Lancet Oncology**, v. 8, p. 349-357, 2007.

PAN, Q. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. **Nature Genetics**, v. 40, p. 1413-1415, 2008.

PAN, Q. et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. **Molecular Cell,** v.16, p. 929-941, 2004.

PAN, Y. F. et al. The ulcerative colitis marker protein WAFL interacts with accessory proteins in endocytosis. **International Journal of Biological Sciences**, v. 29, p. 163-171, 2010.

PEROU, C. M. et al. Molecular portraits of human breast tumours. **Nature**, v. 406, p. 747-752, 2000.

PETTIGREW, C. et al. Evolutionary conservation analysis increases the colocalization of predicted exonic splicing enhancers in the BRCA1 gene with missense sequence changes and in-frame deletions, but not polymorphisms. **Breast Cancer Research**, v. 7, p. 929-939, 2005.

PIND, M. T.; WATSON, P. H. SR protein expression and CD44 splicing pattern in human breast tumours. **Breast Cancer Research and Treatment**, v. 79, p. 75-82, 2003.

PINEDA, P. H. B. **Identificação de genes supostamente envolvidos com o processo de invasão tumoral em câncer de mama através da técnica RaSH**. 2007. Dissertação (Mestrado em Oncologia) - Fundação Antonio Prudente, São Paulo, 2007.

POSPISIL, H. et al. EASED: Extended Alternatively Spliced EST Database. **Nucleic Acids Research**, v. 32, p. 70-74, 2004.

PRIFTI, E. et al. FunNet: an integrative tool for exploring transcriptional interactions **Bioinformatics**, v. 24, p. 2636-2638, 2008.

PRUITT, K.D.; TATUSOVA, T.; MAGLOTT, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. **Nucleic Acids Research**, v. 35, p. 61-65, 2007.

QUEVILLON, E. et al. InterProScan: protein domains identifier. **Nucleic Acids Research,** v. 33, p. 116-120, 2005.

RADICE, D.; REDAELLI, A. Breast cancer management: quality-of-life and cost considerations. **Pharmacoeconomics**, v. 21, p. 383-396, 2003.

RANGEL, M. C. R. **Identificação de marcadores moleculares em câncer de mama através da técnica de microarray utilizando uma plataforma de exons tumor-associados.** 2008. 165f. Tese (Doutorado em Oncologia) - Fundação Antonio Prudente, São Paulo, 2008.

REIS-FILHO, J. S.; LAKHANI, S. R. The diagnosis and management of pre-invasive breast disease: genetic alterations in pre-invasive lesions. **Breast Cancer Research**, v.5, p.313-319, 2003.

RING, A.; DOWSETT, M. Mechanisms of tamoxifen resistance.**Endocrine-Related Cancer**, v. 11, p. 643–658, 2004.

RELÓGIO, A. et al. Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. **Journal of Biological Chemistry**, v. 280, p. 4779-4784, 2005.

RONAGHI, M.; UHLÉN, M.; NYRÉN, P. A sequencing method based on real-time pyrophosphate. **Science**, v. 281, p. 363-365, 1998.

ROYLE SJ. The cellular functions of clathrin. **Cell Molecular Life Sciences**, v. 63, p. 1823-1832, 2006.

SALZ, H. K.; ERICKSON, J. W. Sex determination in Drosophila: The view from the top. **Fly**, v. 4, p. 60-70, 2010.

SAMBROOK, J.; RUSSELL, D. **Molecular Cloning: A Laboratory Manual**. 3a. edição, 2001.

SARAIVA, T.F. et al. Effects of Oligo dT-T7 RNA Primer in RNA Amplification from Paraffin-Embedded Tissue for Microarray Experiments. **Applied Cancer Research**, v. 26, p. 236-242, 2006.

SCHELL, T.; KULOZIK, A. E.; HENTZE, M. W. Integration of splicing, transport and translation to achieve mRNA quality control by the nonsense-mediated decay pathway. **Genome Biology**, v. 3, p. 1006,2002.

SCHMID, E. M. et al. Role of the AP2 beta-appendage hub in recruiting partners for clathrin-coated vesicle assembly. **PLoS Biology**, v. 4, p. 262, 2006.

SCHNITT, S. J. The transition from ductal carcinoma in situ to invasive breast cancer: the other side of the coin. **Breast Cancer Research**., v. 11, p. 101, 2009.

SCHWARTZ, S.; HALL, E.; AST, G. SROOGLE: webserver for integrative, user-friendly visualization of splicing signals. **Nucleic Acids Research**, v. 37, p. 189-192, 2009.

SHARP, P. A. Split genes and RNA splicing. **Cell**, 77, p. 805-815, 1994.

SHENDURE, J.; CHURCH, G. M. Computational discovery of sense-antisense transcription in the human and mouse genomes. **Genome Biology**, v. 3, p.44, 2002.

SHENDURE, J. et al. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. **Science**, v. 309, p. 1728-1732, 2005.

SHULZHENKO, N. et al. Specificity of alternative splice form detection using RT-PCR with a primer spanning the exon junction. **Biotechniques**, v. 34, p. 1244-1249, 2003.

SIMARD, M. J.; CHABOT, B. SRp30c is a repressor of 3' splice site utilization. **Molecular Cell Biology**, v.22, p. 4001-4010, 2002.

SIWICKI, K. K.; KRAVITZ, E. A. Fruitless, doublesex and the genetics of social behavior in Drosophila melanogaster. **Current Opinion in Neurobiology**, v. 19, p. 200-206, 2009.

SLAMON, D.; PEGRAM, M. Rationale for trastuzumab (Herceptin) in adjuvant breast cancer trials. **Seminars in Oncology**, v. 28, p. 13-19, 2001.

SLAMON, D. J. et al. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. **Science**, v. 235, p. 177-182, 1987.

SLAMON, D. J. et al. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. **Science**, v. 244, p. 707-712, 1989.

SLAMON, D.J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. **New England Journal of Medicine**, v. 344, p. 783-792, 2001.

SMITH, P. J., Zhang, C., Wang, J. Chew, S. L., Zhang, M. Q. and Krainer, A. R. 2006. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. **Human Molecular Genetics**, v. 15, p. 2490-2508, 2006.

SNEATH, R. J.; MANGHAM, D. C. The normal structure and function of CD44 and its role in neoplasia. **Molecular Pathology**, v. 51, p. 191-200, 1998.

SORLIE, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. **Procedings of the National Academy of Sciencies of the U S A.**, v.98, p. 10869-10874, 2001.

SPARANO, J. A.; PAIK, S. Development of the 21-Gene Assay and Its Application in Clinical Practice and Clinical trials. **Journal of Clinical Oncology**, v. 26, p. 721-728, 2008.

SREBROW, A.; KORNBLIHTT, A. R. The connection between splicing and cancer. **Journal of Cell Sciences**, v. 119, p. 2635-2641, 2006.

STAMM, S. et al. ASD: a bioinformatics resource on alternative splicing. **Nucleic Acids Research**, v. 34, p. 46-55, 2006.

STAMPS, A.C. et al. Analysis of proviral integration in human mammary epithelial cell lines immortalized by retroviral infection with a temperature-sensitive SV40 T-antigen construct. **International Journal of Cancer**,v. 57, p. 865-874, 1994.

STEIN, L. D. et al. The generic genome browser: a building block for a model organism system database. **Genome Research**, v. 12, p. 1599-1610, 2002.

STEPHAN, M.et al. Self-alignments to detect mutually exclusive exon usage. **In Silico Biology**, v. 7, p. 613-621, 2007.

STIECKLER, E. et al. Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis. **Oncogene**, v. 18, p. 3574-3582, 1999.

STOSSEL, T. P. et al. Filamins as integrators of cell mechanics and signalling. **Nature Review Molecular Cell Biology**, v. 2, p. 138-145, 2001.

SULTAN, M. et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. **Science**, v. 321, p. 956-960, 2008.

TARN, W.Y.; STEITZ, J. A. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. **Cell**, v. 84, p. 801-811, 1996.

THILL, G. et al. ASEtrap: a biological method for speeding up the exploration of spliceomes. **Genome Research**, v. 16, p. 776-786, 2006.

TIMMS, J. F. et al. Effects of ErbB-2 overexpression on mitogenic signaling and cell cycle progression in human breast luminal epithelial cells. **Oncogene**, v. 21, p. 6573-6586, 2002.

TRAPNELL, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **Nature Biotechnology**, v. 28. p. 511-515, 2010.

TZUKERMAN, M. T. et al. Human estrogen receptor transactivational capacity is determined by both cellular and promoter context and mediated by two functionally distinct intramolecular regions. **Molecular Endocrinology**, v. 8, p. 21–30, 1994.

USUKA, J., ZHU, W., BRENDEL, V. Optimal spliced alignment of homologous cDNA to a genomic DNA template. **Bioinformatics**, v. 16, p. 203-211, 2000.

UWANOGO, D.A. et al. Molecular cloning, chromosomal mapping, and developmental expression of a novel protein tyrosine phosphatase-like gene. **Genomics**, v. 62, p. 406-416, 1999.

VAN DE VIJVER, M. J. et al. A gene-expression signature as a predictor of survival in breast cancer. **New England Journal of Medicine**, v. 347, p. 1999-2009, 2002.

VAN 'T VEER, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. **Nature**, v. 415, p. 530-536, 2002.

VENABLES, J. P.; BURN, J. EASI--enrichment of alternatively spliced isoforms. **Nucleic Acids Research**, v. 34, e103, 2006.

VENABLES, J. P. et al. Identification of alternative splicing markers for breast cancer. **Cancer Research**, v. 68, p. 9525-9531, 2008.

VENABLES, J. P. Enrichment of alternatively spliced isoforms. **Methods in Molecular Biology**, v. 419, p.161-170, 2008.

VENTER, J.C. et al. The Sequence of the Human Genome. **Science**, v. 291, p. 1304-1351, 2001.

VOGEL, C. L. et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. **Journal of Clinical Oncology**, v. 20, p. 719-726, 2002.

VOGEL, V. G. Epidemiology, genetics, and risk evaluation of postmenopausal women at risk of breast cancer. **Menopause: The Journal of The North American Menopause Society**, v. 15, p. 782-789, 2008.

WANG, Y. et al. Characterization of mouse Trip6: a putative intracellular signaling protein. **Gene**, v. 234, p. 403-409, 1999.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, p. 57-63, 2009.

WANG, E. et al. High-fidelity mRNA amplification for gene profiling. **Nature Biotechnology**, v.18, p.457-459, 2000.

WATAHIKI, A. et al. Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. **Nature Methods**, v. 3, p. 233-239, 2004.

WEISKIRCHEN, R.; GRESSNER, A. M. The cysteine- and glycine-rich LIM domain protein CRP2 specifically interacts with a novel human protein (CRP2BP). **Biochemical and Biophysical Research Communications**, v. 274, p. 655-663, 2000.

WETERMAN, M.A. et al. Transformation capacities of the papillary renal cell carcinoma-associated PRCCTFE3 and TFE3PRCC fusion genes. **Oncogene**, v. 20, p.1414-1424, 2001.

WETERMAN, M. A. et al. Impairment of MAD2B-PRCC interaction in mitotic checkpoint defective t(X;1)-positive renal cell carcinomas. **Procedings of the National Academy of Sciencies of the U S A.**, v. 98, p. 13808-13813, 2001.

WHITE, E. S. et al. Control of fibroblast fibronectin expression and alternative splicing via the PI3K/Akt/mTOR pathway. **Experimental Cell Resesearch**, v.5, 2010.

WILLIAMS, R. et al. Amplification of complex gene libraries by emulsion PCR. **Nature Methods**, v. 3, p. 545 – 550, 2006.

XING, D.; LI, Q. Q. Alternative polyadenylation: a mechanism maximizing transcriptome diversity in higher eukaryotes. **Plant Signaling & Behavior**, v. 4, p. 440-442, 2009.

XU, J. et al. TRIP6 enhances lysophosphatidic acid-induced cell migration by interacting with the lysophosphatidic acid 2 receptor. **The Journal of Biological Chemistry**, v. 279, p. 10459-10468, 2004.

YI, J., BECKERLE, M. C. The human TRIP6 gene encodes a LIM domain protein and maps to chromosome 7q22, a region associated with tumorigenesis. **Genomics**, v. 49, p.314-316, 1998.

ZHANG, Z. et al. A greedy algorithm for aligning DNA sequences. **Journal of Computational Biology**, v. 7, p. 203-214, 2000.

ZHOU, Z. et al. Comprehensive proteomic analysis of the human spliceosome. **Nature**, v. 419, p.182-185, 2002.

Anexo A

Alternative splicing enriched cDNA libraries identify breast cancer-associated transcripts

Elisa N Ferreira[1,2], Maria C R Rangel [1], Pedro F Galante[3], Jorge E de Souza[3], Gustavo C Molina[1], Sandro J de Souza[3], Dirce M Carraro[2§]

[1]Laboratory of Genomics and Molecular Biology, Hospital A.C. Camargo, Fundação Antonio Prudente, Rua Prof Antonio Prudente, São Paulo, Brazil
[2]Department of Genetics and Evolutionary Biology, Institute of Biosciences, University of São Paulo, Rua do Matão, São Paulo, Brazil
[3] Laboratory of Computational Biology, Ludwig Institute for Cancer Research, São Paulo, Brazil

Abstract

Background. ***Alternative splicing (AS) is a central mechanism in the generation of genomic complexity and is a major contributor to transcriptome and proteome diversity. Alterations of the splicing process can lead to deregulation of crucial cellular processes and have been associated with a large spectrum of human diseases. Cancer-associated transcripts are potential molecular markers and may contribute to the development of more accurate diagnostic and prognostic methods and also serve as therapeutic targets. Alternative splicing-enriched cDNA libraries have been used to explore the variability generated by alternative splicing. In this study, by combining the use of trapping heteroduplexes and RNA amplification, we developed a powerful approach that enables transcriptome-wide exploration of the AS repertoire for identifying AS variants associated with breast tumor cells modulated by ERBB2 (HER-2/neu) oncogene expression.*** Results. ***The human breast cell line (C5.2) and a pool of 5 ERBB2 over-expressing breast tumor samples were used independently for the construction of two AS-enriched libraries. In total, 2,048 partial cDNA sequences were obtained, revealing 214 alternative splicing sequence-enriched tags (ASSETs). A subset with 79 multiple exon ASSETs was compared to public databases and reported 138 different AS events. A high success rate of RT-PCR validation (94.5%) was obtained, and 2 novel AS events were identified. The influence of ERBB2-mediated expression on AS regulation was evaluated by capillary electrophoresis and probe-ligation approaches in two mammary cell lines (Hb4a and C5.2) expressing different levels of ERBB2. The relative expression balance between AS variants from 3 genes was differentially modulated by ERBB2 in this model system.*** Conclusions. ***In this study, we presented a method for exploring AS from any RNA source in a transcriptome-wide format, which can be directly easily adapted to next generation sequencers. We identified AS transcripts that were differently modulated by ERBB2-mediated expression and that can be tested as molecular markers for breast cancer. Such a methodology will be <u>useful</u> for completely deciphering the cancer cell transcriptome diversity resulting from AS and for finding more precise molecular markers.***

Background

More than 30 years ago, Gilbert predicted the existence of protein variants due to the alternative use of exon-intron borders in eukaryotic cells [1]. This prediction has been continually confirmed as a common feature of many species, including humans. Recent estimations, based on high-throughput sequencing, suggest that 90-95% of multiple-exon human genes undergo alternative splicing (AS) [2; 3], producing an average of six distinct transcripts from each gene [4]. This phenomenon enormously impacts the repertoire of proteins, since 80% of AS events occur within the coding region [5], thus interfering in the functional aspects of the cells.

AS regulates important processes, such as embryonic development, cellular differentiation and apoptosis, by the generation of different protein isoforms among distinct tissues, developmental stages and pathological conditions [6; 7;8].
Alterations of the splicing process, such as the loss of expression balance between variants and aberrant splicing, can lead to the deregulation of crucial cellular processes and are consequently associated with a large spectrum of human diseases [9], including cancer [10; 11; 12].

The development of methodologies to explore transcriptome diversity resulting from AS has been shown to be a potent tool, not only for improving the biological basis of cancer but also for searching for more precise molecular markers for diagnostic, prognostic and therapeutic purposes [13; 14]. Different strategies for large-scale AS variant exploration have been used with different goals. Sequence and microarray-based approaches have been used for defining the AS repertoire of human cells. The former includes several computational analyses concerning genomic and transcriptome alignments of human ESTs (expressed sequence tags) and mRNA databases [11; 15; 16; 17] and cross-species alignment from closely related organisms [18; 19]; the latter includes genomic and exon-intron junction microarray platforms [20; 21; 22; 23]. Both approaches have contributed to the investigation of the expression pattern of AS variants and also facilitated the identification of novel AS variants. Nonetheless, both approaches are impaired in detecting low-abundance AS transcripts. In this sense, AS-enriched cDNA libraries is one of the most interesting approaches because it combines the convenience of cDNA direct sequencing with the advantage of detecting low-abundance transcript variants. The methodology is based on one enrichment step, consisting of the trapping of heteroduplex molecules formed by the hybridization of two distinct AS variants from the same gene [24]. The heteroduplex can be captured by molecules that recognize the heteroduplex structure [25; 26], generating a vast number of AS events without previous knowledge of them. In this study, to explore AS variants associated with breast tumor cells, we established a powerful approach that enabled the direct exploration of an AS repertoire by combining the use of trapping heteroduplex and RNA amplification. To favor the trapping of splicing variants associated with breast tumor cells that over-expresses the *ERBB2* (*HER-2/neu*) oncogene, a human breast cell line (C5.2) and a pool of 5 ERBB2 over-expressing breast tumor samples were used. Two AS-enriched libraries were constructed, generating a set of 2,048 partial cDNA sequences, named here as alternative splicing sequence-enriched tags (ASSETs), as suggested by Watahiki and collaborators [25]. A subset with 79 ASSETS representing distinct multiple exon sequences was explored in this analysis and reported 138

different AS events. A high rate of validation by RT-PCR (94.5%) was obtained, and 2 novel AS events were identified. Moreover, the balance in the expression level of the AS variants from 3 genes was influenced by *ERBB2*-mediated expression.

The approach presented here certainly will contribute to the identification of the AS repertoire of cancer cells, especially as it is potentially applicable to any cell type from any tumor tissue, since a small amount of total RNA is required with no previous cDNA library construction. Furthermore, it is completely suitable for using with next-generation sequencers, substantially increasing its potential in deciphering the AS diversity in cancer cell transcriptome.

Results

Alternative splicing libraries

Two distinct AS libraries were constructed (Lib_1 and Lib_2) using 5 µg of total RNA as the starting material. Library 1 (Lib_1) was prepared from the human breast cell line C5.2, which over-expresses the oncogene *ERBB2*, and library 2 (Lib_2) was prepared from a pool of 5 invasive breast carcinoma samples that stained positively for ERBB2 according to immunohistochemistry analysis (Table 1). The strategy for AS library construction was based on the methodology described by Watahiki and collaborators [25] with some modifications. One significant difference was the use of total RNA instead of parental full-length cDNA libraries, which simplifies the process and decreases costs. Another important adaptation was the inclusion of an RNA amplification procedure based on T7 RNA polymerase and Template Switch oligo (TS-oligo) [27], which allows the use of small quantities of RNA (Figure 1 I-IV). The amplified RNA was converted into double-stranded cDNA (dscDNA) (Figure 1 V-VI), which was then submitted to denaturation and renaturation steps, promoting the formation of heteroduplex DNA molecules by the hybridization of complementary regions from two distinct splicing variants from the same gene (Figure 1 VII). The remaining single-stranded molecules or overhanging regions were removed with exonuclease VII treatment (Figure 1 VIII), whereas the double-stranded cDNA molecules were cleaved with the *Dpn*II restriction enzyme (Figure 1 IX). This step resulted in double-stranded fragments, constituting homo- and heteroduplex molecules with cohesive ends to bind adaptors. The enrichment of the heteroduplex molecules occurred through the trapping of single-stranded loops by the annealing of random 25-mer biotinilated oligonucleotides captured by streptavidin magnetic particles (Figure 1 X). Complementary overhanging adaptors were then specifically ligated to the cohesive ends of the heteroduplex molecules (Figure 1 XI), generating a recognition site for primer annealing and consequently allowing for PCR amplification (Figure 1 XII), cloning and sequencing.

A total of 2,048 high quality sequences (Phrep > 20) were generated from both libraries. Sequences from each library were clustered using the CAP3 program [28], resulting in 149 consensus sequences for library Lib_1 (96 contigs and 53 singlets) and 146 consensus sequences for library Lib_2 (74 contigs and 74 singlets) (Table 2 and Figure 2A). The number of consensus sequences obtained revealed, as expected, a high redundancy within the libraries (Table 1), since no normalization procedure was implemented in our approach.

All consensus sequences were then aligned to the human genome (NCBI build #36.1) using BLAST [29] and Sim4 [30], where only the best hit was considered. Based on criteria for identity (≥ 93%) and coverage (≥ 55%), 214 consensus sequences were aligned on the human genome, 93 and 121 of them reporting multiple and one-exon(s) sequences, respectively (Figure 2B). The consensuses were termed ASSETs, as previously proposed [25; 26]. Furthermore, to check whether our library construction approach enables full-length representation, including the 5' end of transcripts, we verified the relative position of the ASSETs throughout the length of full-lengths (Figure 3). The analysis resulted in a similar representation of 3' and 5'ends with a slightly higher concentration of ASSETs in the central region, indicating that no bias were introduced towards higher representation of full transcript 3'ends. Additionally, the fact that the average size of mRNAs (RefSeq) represented by ASSETs in our libraries was of 2,836 nt, similar to the average size of all mRNA sequences from RefSeq database (3,098 nt) suggested no bias in representation of 5'end from short transcripts (Figure 3).

Detection of alternative splicing events

No distinct splicing variants were observed among the sequences belonging to the same consensuses that would be indicative of putative AS events. Therefore, we searched for AS events through comparisons between ASSETs and full-length or partial cDNA sequences available in public databases.

First, ASSETs were clustered with ESTs from dbEST (8,133,299 ESTs), mRNAs (244,284 sequences) and RefSeqs (26,040 sequences) downloaded from UCSC (September 2007) (Figure 2C). This step resulted in 164 clusters, where 142 contained at least one RefSeq sequence. Sixteen clusters contained sequences from both libraries (Lib_1 and Lib_2), revealing an overlap of approximately 10%.

The 79 clusters containing ASSETs with multiple exons were scanned for AS events through pairwise comparisons of exon/intron boundaries between the ASSET and the reference sequences of each cluster. AS events were searched within the region delimited by the two outermost overlapping regions of each ASSET related cluster. For each ASSET, the corresponding gene and the number and type of related alternative splicing events were annotated.

All 79 multiple exon ASSETs were considered known transcripts since they were represented by sequences at public databases. Moreover, for 39 out of 79 ASSETs (49.4%), an alternatively spliced transcript was described in the public databases. For these 39 ASSETs, 138 AS events were detected, including intron retention (5.8%), exon skipping (9.4%), alternative splice site 3' (39.8%) and alternative splice site 5' (44.9%). The remaining 40 ASSETs, to which no AS event has yet been reported in the public domain, may result from novel AS events not yet reported in public databases (Table 2).

The intronless ASSETs were not used for the AS search, since it is not possible to identify the direction of transcripts in the absence of splice sites. Nonetheless, it is interesting to note that 63 out of 96 intronless sequences (65.6%) aligned to regions involved in AS, according to public databases. This can be considered an indirect sign that these ASSETS are prone AS transcripts.

Gene ontology annotation

For exploring the functional aspects of the genes that harbor AS, the 142 ASSETs were classified within the biological process categories. Using BinGO tools [31], 11 categories revealed a statistically significant enrichment of genes (Table 3) and are represented in a hierarchical form in Figure 4. The most significantly enriched category was translation elongation, due to a great number of ribosomal proteins detected in our data.

Validation of ASSETs and heteroduplexes

Eighteen ASSETs were randomly selected for RT-PCR validation, including 6 and 7 ASSETs exclusively from Lib_1 andLib_2, respectively and 5 ASSETs that were detected in both libraries. The validation process was performed in two steps: *i.* ASSET validation - to confirm the presence of the ASSET in the same RNA used for library construction and *ii.* heteroduplex validation - to search for alternatively spliced transcripts that could have participated in the heteroduplex formation (Table 3). By using a pair of primers that aligned at the extremities of the ASSET sequence, all but one ASSET was validated (17 out of 18, 94.4% validation rate). The 5 ASSETs identified by both libraries were validated in both templates. Secondly, for 6 (*SFRS9*, *FLNA*, *ALDH3A2*, *PTPLA*, *RPS2* and *TRIP6*) out of the 17 validated ASSETs (35.3%), an additional AS variant was identified that could have participated in the heteroduplex formation. Two out of 6 AS variants that were transcribed from the genes *PTPLA* and *TRIP6,* which were not described in public databases, are novel splicing variants. The lack of heteroduplex

validation for the other 11 genes was probably due to a differential expression balance between splicing variants that precluded the amplification of one variant in favor of the most abundant one. The support for this assumption is that for 5 out of 11 (45.5%) genes, an AS variant that could have participated in the heteroduplex formation was available in databases.

For verifying whether the low level of overlap between both libraries was due to the low coverage in terms of the number of sequences generated for each library or due to the specific AS pattern of each RNA source used, we tested if the 13 ASSETs validated in cDNA from the corresponding library were also expressed in the cDNA from the other library. Four ASSETs from the 5 identified by Lib_1 were successfully amplified using the cDNA from the pool of the tumor samples (Lib_2). All 7 ASSETs from Lib_2 were successfully amplified using the cDNA from C5.2 (Lib_1), totaling 91.8% cross-validation (11 out of 12). The validation results are summarized in Table 4.

Novel alternative splicing: characterization of the putative isoforms

The 2 novel AS variants were characterized regarding the putative corresponding protein isoform. The *PTPLA* gene [RefSeq:NM_014241.3] codes for the member *A* of the protein tyrosine phosphatase-like family that contains proline instead of catalytic arginine. This gene contains 7 exons, and the AS variant detected in our study is due to the use of an alternative 5' splice site of intron 5 that elongates exon 5 by 117 nt (Figure 5A). All protein functional domains found for PTPLA were also present in the novel AS detected. However, in the novel AS variant, a premature stop codon was created 96 nt upstream of the exon 5/exon 6 junction, probably leading to regulation by non-sense mediated decay (NMD) [32; 33].

The *TRIP6* gene [RefSeq:NM_003302.2] is a thyroid hormone receptor interactor 6 that contains 9 exons. The novel alternatively spliced transcript reports retention of the last intron (Figure 5B). The protein coded by the *TRIP6* gene localizes to focal adhesion sites and along actin stress fibers. The novel AS variant identified also inserts a premature stop codon in the putative coding protein, without interfering with any protein functional domain.

Evaluation of AS variant regulation by *ERBB2*-mediated expression

Finally, we investigated the putative influence of *ERBB2*-mediated expression on the regulation of AS variants for 17 ASSETS validated using *GAPDH* as a normalization factor, by comparing the expression level of the ASSETs in the C5.2 cell line in relation to the *ERBB2* basal expressed counterpart – the normal breast cell line (Hb4a) through capillary microfluidic electrophoresis (LabChip GX – Caliper Lifesciences) that accurately assesses the size and quantity of each amplification product [34].

For the 11 validated ASSETs, the relative expression levels were analyzed showing a slight influence of *ERBB2* over-expression in all ASSETs (ratio ranging from -1.9 to 1.4) (Supplemental Table 1).

For the 6 ASSETS presenting an additional AS variant, the putative influence of *ERBB2* over-expression in the relative expression balance of the pairs of distinct splicing variants (ASSET and additional AS) was evaluated in both cell lines. We first calculated the expression ratio from ASSET against the variant to each cDNA template and then compared the expression ratio between the C5.2 against Hb4a cell lines.

For 3 out of 6 genes, a decrease in the expression balance of the ASSET and additional AS variants was identified between the tumor and normal cell lines (Figure 6; Table 6). In more detail, the ASSETs corresponding to *SFRS9* [RefSeq: NM_003769.2] and *FLNA* [RefSeq: NM_001456.3] genes were stably expressed between cell lines, while the additional AS variants were more expressed in the C5.2 compared to Hb4a cell line (fold = 4 and 3.5, respectively) leading to a decrease of 4.84 (*SFRS9*) and 4.78 (*FLNA*) in the expression balance between the splice variants (Table 5). The ASSET of the *TRIP6* gene [RefSeq: NM_003302.2] was more expressed in the Hb4a than in the C5.2 (fold=4.6), whereas the additional AS variant presented no expression difference. These results suggested that *ERBB2*-mediated expression differently modulates the alternative splice variants of the genes *SFRS9*, *FLNA* and *TRIP6*. For the other 3 genes (*RPS2*, *PTPLA* and *ALDH3A2*), no difference in the expression balance of the AS variants between the cell lines was observed (Table 5).

To confirm the alteration in the relative expression balance of AS variants mediated by *ERBB2* expression, a different approach based on probe-specific ligation and PCR amplification was applied [35]. In this strategy, 2 pairs of probes were designed for each gene, specifically targeting the variants of interest (Figure 6). The expression balance difference was confirmed for all 3 genes (*FLNA*, *SFRS9* and *TRIP6*) visualized on the acrylamide gel (Figure 6).

Discussion

The diversity of the human transcriptional repertoire caused by AS has been extensively investigated [2-3], and it is agreed that its regulation is an important mechanism for physiological and pathological aspects of cells. Moreover, AS is a major contributor to protein diversity, which, in part, explains the high complexity of mammals compared to much simpler organisms containing a similar numbers of genes [5].

Different approaches have been used to explore the variability caused by this phenomenon, and one of the most promising strategies is the use of AS enriched cDNA libraries [25-26]. This strategy does not require previous knowledge of the variants and permits an AS transcriptome-wide analysis.

Deciphering of the human transcriptional repertoire related to AS variability is an enormous contribution in the comprehension of cancer and in the identification of more precise molecular markers in cancer.

Here we described an AS enriched cDNA library method by combining the use of trapping heteroduplex and RNA amplification procedures. The methodology was initially proposed by Watahiki and collaborators [25] and was applied in this study with some modifications to favor its application in clinically-oriented cancer studies, in which the availability of total RNA recovered from tumor tissues is normally restrictive. Moreover, the methodology established in this study is potentially applicable to RNA purified from a homogenous tumor cell population captured from a complex tissue by laser, which produces transcriptional data more correlated with the tumor cell.

Our strategy showed, in general, minimal artifacts in the identification of ASSETs, since our validation rate by RT-PCR was significantly high (94.5%). Moreover, the fact that the great majority of the AS events found in our AS enriched libraries were present in public databases and that 100% of them harbor conserved splice sites strengths the assumption that we have established a robust methodology for identifying AS in a transcriptome-wide format.

The fact that we could confirm by RT-PCR novel alternatively spliced transcripts for 2 genes to which no AS variant was present in public databases is further support that among the ASSETs with no confirmation of AS events, a high frequency of prone additional AS variants, which could participate in heteroduplex formation, is expected.

The absence of amplification during the validation process of additional AS transcripts for two thirds of the selected genes suggests a significant difference in the expression level of both variants with consequent competition for the same pair of primers in the PCR reaction, avoiding the amplification of low-abundance AS transcripts.

The relatively high redundancy levels encountered in both libraries (84.78% and 86.84%) were somewhat expected. This number is similar to the redundancy reported by Thill and collaborators [26]. In technical terms, this problem can be bypassed by decreasing the number of PCR cycles in the library construction, which is relatively easy to control.

Another potential problem was that no additional alternatively spliced transcripts were identified in sequence data provided by

enriched cDNA libraries alone. This can be indicative of a problem caused by using non-phosphorylated adaptors. In this situation, only one strand (5'-3') of these adaptors was ligated to the 5' end of the *Dpn*II digested molecules that contains a phosphate residue; the other strand (3'-5') was not ligated and, as a consequence, was disconnected from the cDNA molecules at the denaturation step of PCR and was consequently unable to be cloned and sequenced. Usually this region is re-synthesized by polymerase at the first cycle of the PCR reaction through annealing of complementary regions of cDNA molecules, a process known as polymerase fill-in, also seen in some cDNA library approaches [36; 37]. However, in our case where the strands of cDNA molecules are from distinct alternatively spliced transcripts, the fill-in process is probably inefficient due to non-perfect annealing. To avoid this problem, the use of phosphorylated adaptors is a simple solution that would favor the representation of both alternatively spliced transcripts that formed the heteroduplex structure.

ERBB2, or *HER-2/neu*, is an oncogene that is over-expressed in 20-30% of human breast carcinomas and is associated with poor prognosis, independent of the lymph node status [38; 39]. This marker is also associated with chemo resistance to a range of anticancer drugs and a positive response to herceptin [40; 41]. Despite this oncogene being most extensively investigated in clinical and basic oncology, the *ERBB2*-mediated mechanism involved in the transformation and progression of breast tumors has not yet been totally elucidated.

In this study, we proposed to identify alternatively spliced transcripts associated with breast tumors that are under *ERBB2* influence by constructing 2 AS-enriched cDNA libraries using RNA sources representative of *ERBB2* over-expression: the human breast cell line C5.2 that was previously transfected with 4 copies of full-length *ERBB2* [42] and a pool of 5 breast carcinoma samples, which demonstrate strong positivity in ERBB2 immunostaining in tumor cell membranes [43].

For testing if the expression of ASSETs was regulated by *ERBB2*-mediated expression, we evaluated the ASSETs validated by RT-PCR in both cell lines, HB4a and C5.2, the former with basal levels and the latter with over-expression of *ERBB2* mRNA [44]. Both cell lines have been considered a model for the investigation of *ERBB2*-mediated expression, since the only difference between them is the insertion of 4 copies of full-length *ERBB2* in the C5.2 cell line [45; 46]. For the ASSETs in which we could identify an additional AS transcript by RT-PCR, 50% of them (3 out of 6 - *TRIP6*, *FLNA* and *SFRS9*) seemed to be influenced by *ERBB2*-mediated expression, since differences in the relative expression balance between both cell lines were observed.

Although the expression assessment of 2 or more AS variants is a problematic issue concerning accurate quantification the results presented here were confirmed by an alternative methodology, which increased the robustness of the data.

The microfluidic capillary electrophoresis-based strategy relies on amplification of both variants in the same reaction and could introduce bias due to amplification competition between variants. However, this would be expected to equally influence all reactions, independent of the template used. The alternative strategy relies on the specific binding of probes under highly stringent conditions, enabling the evaluation of each variant separately, with high accuracy and is consequently very promising for AS expression assessment. The different expression balance between both cell lines for 3 genes confirmed by 2 different approaches suggests that these genes transcribe AS variants, whose expression is differently influenced by *ERBB2*.

FLNA is a member of the actin-binding protein family that organizes actin filaments and is involved in numerous cellular processes, especially development. Many studies have reported the involvement of this protein in carcinogenesis. Using melanoma cell lines lacking or expressing *FLNA*, Fiori and collaborators [47] have shown that this protein is an important regulator of EGFR members (including *ERBB2*) that ensure efficient ligand-mediated activation of these receptors and, consequently, intracellular trafficking and degradation.

SFRS9 is a RNA-binding protein from the arginine/serine-rich family that acts as a splicing factor regulating constitutive splicing and also modulating the selection of alternative splice sites. It has been suggested that this protein acts downstream of the ERBB2 pathway, since phosphorylation of SFRS9 was detected in *ERBB2*-over expressing breast and ovarian cancer cells and was reduced by monoclonal antibody *Herceptin* treatment. Moreover, a putative role for SFRS9 in cell migration was suggested, since migration was significantly retarded following the depletion of *SFRS9* transcripts in ovarian cancer cell lines [48].

TRIP6 is a thyroid hormone receptor interactor that localizes to focal adhesion sites and along actin stress fibers [49; 50]. This protein enhances lysophosphatidic acid (LPA) -induced cell migration by directly binding to the carboxyl-terminal tail of the LPA2 receptor through its LIM domains [51]. Moreover, TRIP6 might enhance cell migration by binding to PDZ domain of MAGI-1b/PTEN destabilizing membrane β-catenin and E-cadherin junctional complexes, promoting cell motility [52].

The development of strategies to selectively represent the AS transcripts repertoire, requiring small amounts of total RNA, will be important for generating more correlated information between AS transcripts and specific cell types and conditions in a transcriptome-wide format.

In spite of using Sanger sequencing in this study, our approach is completely suitable for using with next-generation sequencers [53], with the possibility of decreasing the number of PCR cycles, and consequently the redundancy level of the library; and assaying multiple barcoded samples with high sequence coverage in a single run. Finally, the use of next generation sequencers would tremendously expand the applicability of our approach toward characterizing cancer cell transcriptome diversity resulting from AS.

Conclusions

In this study we presented a method for exploring AS from any RNA source that generates reliable AS data in a transcriptome-wide format. Additionally, our data identified AS transcript candidates whose expression was influenced by *ERBB2*-mediated expression and can be tested as molecular markers for breast cancer. The association of such methodology with deep sequencing may be helpful for completely deciphering the cancer cell transcriptome and finding more precise molecular markers.

Methods

Samples

The human breast cell line C5.2 is derived from normal luminal cells transfected with four copies of the full-length ERBB2 cDNA (HER-2/neu) gene presenting tumor characteristics [42]. Cells were maintained in RPMI medium supplemented with 100 ml/l fetal bovine serum (FBS), 5 µg/ml insulin, 5 µg/ml hydrocortisone and 1 mmol/l L-glutamine in a humidified incubator containing 50 ml/l CO_2 at 37°C. The medium was changed every 2-3 days, and after 10 days the total RNA was extracted using a CsCl gradient [54]. The yield of extracted total RNA was determined with a Kontron 810 spectrophotometer GeneQuant pro (Amersham Pharmacia Biotech), and the integrity was also verified by electrophoresis through 1% agarose gel upon visualization with ethidium bromide. RNA samples from 5 ductal breast carcinoma samples used in this study were provided by the biorepository bank from A.C. Camargo Hospital (São Paulo, Brazil). These samples were positive for ERBB2 through immunohistochemistry analysis (Table 6), according to the following criteria: weak to moderate complete membrane staining in > 10% of tumor cells or strong complete membrane staining in > 30% of tumor cells.

Alternative splicing enriched cDNA library construction

RNA amplification and double strand cDNA synthesis

For first strand cDNA synthesis, total RNA was incubated with 0.75 µg oligo dT containing the T7 RNA polymerase site (5'AAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGC GCT(24)'3') at 70°C for 10 minutes. The reaction was performed by adding 1X first strand buffer, 0.01 M DTT (Dithiothreitol), 40 U of RNasin (Promega, Madison, WI), 1 mM dNTP (GE Healthcare), 1 µg of Template Switch (TS) DNA Oligo (5'AAGCAGTGGTAACAACGCAGAGTACGCGGG 3') and 400 U of SuperScript II (Invitrogen) in a total volume of 20 µl. The reaction was incubated for 120 min at 42ºC. For the second strand synthesis, the Advantage® cDNA PCR Kit (Clontech) was used as follows: 5X cDNA PCR Reaction Buffer, 1 mM dNTP Mix, 5X Advantage cDNA Polymerase Mix, 1.4 U of RNase H (Invitrogen) in a final volume of 100 µl. The reaction was incubated at 37°C for 10 min, 94°C for 3 min, 65°C for 5 min. and 75°C for 30 min. The stop reaction including 0.25 M of NaOH and 0.5 mM EDTA was added, followed by incubation at 65°C for 10 min. The dscDNA was purified by phenol:chloroform:isoamylic alcohol (25:24:1) pH 8.0 extraction followed by Microcon YM-100 Centrifugal Filter Unit. Double-strand cDNA was in vitro transcribed into RNA with RiboMAX™ Large Scale RNA Production Systems (Promega) as follows: 1X buffer, 3 µM rNTP and 2.5 µl Enzyme T7 Mix. The reaction was incubated at 37°C for 6 hours. Amplified RNA (aRNA) was purified by TRIzol® Reagent (Sigma – Aldrich Corporation). After purification, aRNA was used for double-stranded cDNA synthesis as described above using 1 µg of TS-oligo for the first strand synthesis and 0.5 µg oligo dT(24) for the second strand synthesis.

Denaturation and renaturation

Double-stranded cDNA molecules were heated at 96°C for 20 min and incubated at 42°C for 24 hours in a mixture of 0.2% SDS, 0.5 M NaCl, 0.05 M Tris-HCl pH 7.5 and 30% formamide.

Exonuclease VII cleavage

Exonuclease VII (USB) cleavage was performed in 70 mM Tris-HCl, pH 8.0; 8 mM EDTA, pH 8.0; 10 mM 2-mercaptoethanol; 50 µg/ml BSA and 0.2 U of the enzyme and incubated at 37°C for 30 min. The enzyme was inactivated at 95°C for 10 min.

*Dpn*II digestion

Fifteen units of the restriction enzyme II (New England Biolabs) was used for each 500 ng of cDNA in 1X buffer. The reaction was incubated at 37°C for 3 hours.

Heteroduplex molecule trapping by biotin-streptavidin

The cDNA sample was incubated with 100 pmoles of random 25-mer oligonucleotide biotinilated at the 5' end in 6X SSC and 0.1% SDS at 65 ºC for 16 hours. This mixture was incubated with 1 mg streptavidin magnetic particles (F. Hoffmann-La Roche Ltd.) and 300 µl TEN100 binding buffer (10 mM Tris-HCl; 1 mM EDTA; 100 mM NaCl, pH 7.5) for 30 min at room temperature. The tube was applied to a magnetic separator, and the supernatant was removed and incubated with another aliquot of streptavidin magnetic particle for a second round of purification. Both aliquots of magnetic particles coupled to heteroduplex molecules by the biotinilated random oligonucleotide were mixed and washed 3 times with TEN100 washing buffer (10 mM Tris-HCl; 1 mM EDTA; 1 M NaCl, pH 7.5). The cDNA molecules were then eluted from the magnetic particles by adding 6 M guanidine-HCl and purified by a phenol: chloroform: isoamylic alcohol pH 8.0 extraction.

Ligation of XDPN12 and XDPN14 adaptors

The adaptors were commercially synthesized and contained four bases complementary to the cleavage site of the *Dpn*II enzyme. First, the cDNA molecules were mixed with 1X T4 Ligase Buffer, 400 pmols XDPN12 (5'GATCTCTCGAGT3') and 400 pmols XDPN14 (5'CTGATCACTCGAGA3') and incubated at 55°C for 1 min. Next, the temperature was decreased from 54°C to 28°C at a rate of 2°C every 2 min and from 28°C to 14°C at a rate of 2°C every 4 min to favor a perfect annealing of the oligonucleotides. At last, 2000 units of T4 DNA ligase (Invitrogen) were added, and the reaction was incubated at 14°C for 16 hours. The reaction was purified with a Microcon YM-100 Centrifugal Filter Unit.

Polymerase chain reaction

The RT-PCR reaction was carried out in 1X buffer, 0.1 mM dNTP, 1.5 mM MgCl2, 200 pmols XDPN18 oligonucleotide (5'CTGATCACTCGAGAGATC3'), 2 units GoTaq® DNA Polymerase (Promega) and 10 µl of purified cDNA in a total volume of 20 µl. The reaction was incubated at 95°C for 4 min followed by 35 cycles of 95°C for 45 s, 58°C for 1 min and 72°C for 4 min and a final extension at 72° for 7 min.

Cloning and sequencing

PCR products were inserted into T/A plasmid vector pTZ57R/T using the InsT/Aclone PCR Product Cloning Kit (Fermentas), following the manufacturer's recommendations, in a total volume of 10 µl. The ligation was performed at 22°C for 16 hours. The ligation was dialyzed for 20 min in nitrocellulose membrane (0.025 µM – MILLIPORE), and 3 µl was used for transformation in DH10B E. coli cells by electroporation (2.5 KV, 25 µFD, 200 OHMS). The clone inserts were sequenced with ABI Prism 3100 (Applied Biosystems). The sequencing reaction was performed with M13 reverse primer (5'GTCATAGCTGTTTCCTG3') and BigDye Terminator v3.1 cycle sequencing kit (Applied Biosystems), following the manufacturer's recommendations.

Bioinformatics analysis

The sequences were automatically analyzed, and regions corresponding to vector sequences were trimmed. The quality control was performed in 20 bp windows, where only windows containing at least 15 bp with a Phrep quality score ≥ 20 were considered. The sequences of each library were clustered individually using the CAP3 program, allowing estimation of library's redundancy. The consensus sequences were first aligned against the human genome (NCBI build #36.1) using BLAT [29]. Second, to improve the quality of and specificity of alignment the best hit of each sequence in the genome was selected, and realigned using Sim4 [30]. Third, sequences showing identity ≥ 93% and sequence coverage (percentage of sequence length aligned) ≥ 55% were considered. Lastly, the sequences were clustered with ESTs from dbEST (8,133,299 sequences), mRNAs (244,284 sequences) and RefSeqs (26,040 sequences) downloaded from UCSC (September 2007) (see Galante [55] for more details).

RT-PCR validation

The primers for splice variant validation were designed at the extremities of the ASSET sequence. Twenty nanograms of cDNA from both the total RNA from the C5.2 cell line and the pool of breast cancer samples were used to validate the ASSETs from Lib_1 and Lib_2, respectively. The PCR reaction was performed in a total volume of 20 µl by mixing 1 X reaction buffer (Invitrogen-Life Technology Carlsbad, CA), 2.5 mM $MgCl_2$ (Invitrogen-Life Technology Carlsbad, CA), 0.2 mM dNTP (Amersham Biosciences, Piscataway, NJ), 10 pmoles of each primer and 1 unit Taq Platinum (Invitrogen-Life Technology Carlsbad, CA). PCR reactions were performed with 40 cycles at 95°C for 30 sec, 60°C for 30 sec and 72°C for 30 sec, followed by a final extension at 72°C for 7 min. Amplification products were visualized on a 8% acrylamide gel and subsequently sequenced by ABI3130.

ERBB2 influence on relative expression

For verifying the *ERBB2* influence on gene expression, all ASSETs were amplified using the C5.2 cell line and also the Hb4a cell line, which is a human mammary luminal epithelial cell line. The PCR products were quantified through capillary microfluidic electrophoresis (LabChip GX – Caliper Lifesciences). The expression of the *GAPDH* gene was used as a normalization

factor. The expression ratio was determined by the normalized value of C5.2 divided by the normalized value of Hb4a for each ASSET. Genes were considered to be differently expressed between cell lines for ratios ≥|2|. The differently expressed genes were analyzed in a group of tumor and normal breast samples through a strategy based on specific-probe ligation. The left and right probes were targeted against specific exon junctions of each variant of a gene. The left probe contained at its 5' end a recognition sequence of the forward PCR primer (GGGTAGGCTAAGGGTAGGA) followed by a stuffer sequence of 38 nucleotides (CCGTTGCCAGTCTGCTCAGACCTCCCTCGCGCCATCAG), and the right probe was phosphorylated at its 5' end and contained a recognition sequence of the reverse PCR primer (TCTAGATTGGATCTTGCTGGCAC) at its 3' end. A specific RT primer designed downstream of the probe target sequence was used for cDNA synthesis. The probes were hybridized to pre-heated cDNA from Hb4a and C5.2 at 600°C overnight, and only the probes specifically hybridized to their target sequences were connected by T4 DNA ligase, resulting in one unique probe. As a negative control, ligation and hybridization were performed in the absence of any template for all pairs of probes. The unique probes were PCR amplified. Amplification products were analyzed on 8% acrylamide gel.

Authors' contributions

ENF designed the study, carried out all wet lab assays and wrote the manuscript. MCRR participated in the design of the study and helped with the construction of the libraries. PAG and JES performed the bioinformatics analysis. GCM participated in the validation experiments. SJS conceived the study and coordinated the bioinformatics analysis. DMC conceived, designed and coordinated the study and wrote the manuscript.

Acknowledgements

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (CEPID/FAPESP 98/14335). ENF is supported by grant FAPESP (05/56289-2). We are grateful to the Biobank and the Research and Educational Center at A.C. Camargo Hospital. We thank Dr. Ricardo Renzo Brentani for important comments and corrections on the manuscript.

References

1. Gilbert W. Why genes in pieces? *Nature* 1978, 271: 501.
2. Pan Q, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet.* 2008, 40:1413-1415
3. Wang ET, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008, 456: 470-476
4. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 2006, 7:1-9
5. Zavolan M,.van Nimwegen E. The types and prevalence of alternative splice forms *Curr. Opin. Struct. Biol.* 2006, 16:362–367.
6. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 2003, 72:291-336
7. Yura K, et al Alternative splicing in human transcriptome: functional and structural influence on proteins. *Gene* 2006, 380:63-71.
8. Xing Y & Lee C. Relating alternative *splicing* to proteome complexity and genome evolution. *Adv Exp Med Biol.* 2007, 623:36-49
9. Tazi J, Bakkour N, Stamm S. Alternative splicing and disease. *Biochim Biophys Acta.* 2009, 1792:14-26.
10. Bartel, F., Taubert, H., Harris, L.C. Alternative and aberrant splicing of MDM2 mRNA in human cancer. *Cancer Cell* 2002, 2: 9-15.
11. Hui L, et a. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene* 2004, 23:3013-3023
12. Venables JP, Klinck R, Koh C, Gervais-Bird J, Bramard A, Inkel L, Durand M, Couture S, Froehlich U, Lapointe E, Lucier JF, Thibault P, Rancourt C, Tremblay K, Prinos P, Chabot B, Elela SA. Cancer-associated regulation of alternative splicing. *Nat Struct Mol Biol.* 2009, 16:670-676
13. Venables JP, Klinck R, Bramard A, Inkel L, Dufresne-Martin G, Koh C, Gervais-Bird J, Lapointe E, Froehlich U, Durand M, Gendron D, Brosseau JP, Thibault P, Lucier JF, Tremblay K, Prinos P, Wellinger RJ, Chabot B, Rancourt C, Elela SA. Identification of alternative splicing markers for breast cancer. *Cancer Res.* 2008, 68:9525-9531
14. Brinkman, B. M. N. Splice variants as cancer biomarkers. *Clinical Biochemistry* 2004, 37: 584-594.
15. Hsu FR, Chang HY, Lin YL, Tsai YT, Peng HL, Chen YT, Cheng CY, Shih MY, Liu CH, Chen CF. AVATAR: a database for genome-wide alternative splicing event detection using large scale ESTs and mRNAs. *Bioinformation* 2005, 1:16-8
16. Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 2001, 29: 2850-2859
17. N. Kirschbaum-Slager, R. B. Parmigiani, A. A. Camargo and S. J. de Souza. Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. Physiol. Genomics, 2005, 21, 423–432.
18. Kan, Z., Rouchka, E.C., Gish, W.R. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* 2001, 5, 889-900
19. Chen FC, Chen CJ, Ho JY, Huang TJ. Identificatyion and evolutionary analysis of novel exons and alternative splicing events using cross-species EST-to-genome comparisons in human, mouse and rat. *BMC Bioinformatics* 2006, 7:136
20. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 2003, 302:2141-2144.
21. Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, Schweitzer A, Awad T, Sugnet C, Dee S, Davies C, Williams A, Turpaz Y. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics* 2006, 7: 325.
22. Cuperlovic-Culf M, Belacel N, Culf AS, Ouellette RJ. Microarray analysis of alternative splicing. *OMICS* 2006, 10:344-357
23. Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, Johnson JM. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet.* 2008, 40:1416-1425.
24. Ferreira EN, Rangel MC, Pineda PB, Vidal DO, Camargo AA, Souza SJ, Carraro DM. Heteroduplex formation and S1 digestion for mapping alternative splicing sites. *Genet Mol Res*. 2008, 7:958-969.
25. Watahiki A, Waki K, Hayatsu N, Shiraki T, Kondo S, Nakamura M, Sasaki D, Arakawa T, Kawai J, Harbers M, Hayashizaki Y, Carninci P. Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. *Nat Methods* 2004, 3:233-239
26. Thill G, Casteli V, Pallud S, Salanoubat M, Wincker P, de la Grange P, Auboet D, Schachter V, Weissenbach J. ASEtrap: a biological method for speeding up the exploration of spliceomes. *Genome Res.* 2006, 16:776-786
27. Matz M, Shagin D, Bogdanova E, Britanova O, Lukyanov S, Diatchenko L, Chenchik A. Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res.* 1999, 27: 1558-1560.
28. Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res.*1999, 9: 868-877

29. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002, 12:996-1006
30. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 1998, 9: 967-974
31. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics* 2005, 21: 3448-3449
32. Lewis BP, Green RE, Brenner SE Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* 2003, 7:189-192
33. Green RE, Lewis BP, Hillman RT, Blanchette M, Lareau LF, Garnett AT, Rio DC, Brenner SE. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics* 2003, 19:i118-21.
34. Venables JP, Koh CS, Froehlich U, Lapointe E, Couture S, Inkel L, Bramard A, Paquet ER, Watier V, Durand M, Lucier JF, Gervais-Bird J, Tremblay K, Prinos P, Klinck R, Elela SA, Chabot B **Multiple and specific mRNA processing targets for the major human hnRNP proteins.** *Mol Cell Biol*. 2008, **28**:6033-6043.
35. Nardi A, Pomari E, Zambon D, Belvedere P, Colombo L, Dalla Valle L. **Transcriptional control of human steroid sulfatase.** *J Steroid Biochem Mol Biol.* 2009,
36. Jiang Z, Cote J, Kwon JM, Goate AM, Wu JY. **Aberrant splicing of tau pre- mRNA caused by intronic mutations associated with the inherited dementia frontotemporal dementia with Parkinson linked cromossome 17**. *Molecular and Cellular Biology* 2000, **20**:4036 – 4048.
37. Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD. **Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries**. *Proc Natl Acad Sci U S A* 1996, **93**:6025-6030.
38. Albanell J, Baselga J. **Unraveling resistance to trastuzumab (Herceptin): insulin-like growth factor-I receptor, a new suspect**. *J Natl Cancer Inst* 2001, **93**:1830-183
39. Slamon DJ, Leyland-Jones B, Shak S, et al. **Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.** *N Engl J Med* 2001, **344**:783-92
40. Kumar CC, Madison V. **Drugs targeted against protein kinases**. *Expert Opin Emerg Drugs* 2001, **6**:303-315
41. Slamon D, Pegram M. **Rationale for trastuzumab (Herceptin) in adjuvant breast cancer trials.** *Semin Oncol* 2001, **28**:13-19.
42. Harris RA, Eichholtz TJ, Hiles ID, Page MJ, O'Hare MJ. **New model of ErbB-2 over-expression in human mammary luminal epithelial cells**. *Int J Cancer* 1999, **80** :477-484
43. Press MF, Hung G, Godolphin W, Slamon DJ. **Sensitivity of HER-2/neu antibodies in archival tissue samples: potential source of error in immunohistochemical studies of oncogene expression.** *Cancer Res*. 1994, **54**:2771-2777.
44. Stamps AC, Davies SC, Burman J, O'Hare MJ. **Analysis of proviral integration in human mammary epithelial cell lines immortalized by retroviral infection with a temperature-sensitive SV40 T-antigen construct.** *Int. J. Cancer* 1994, **57**:865-874.
45. Jongeneel CV, Iseli C, Stevenson BJ, Riggins GJ, Lal A, Mackay A, Harris RA, O'Hare MJ, Neville AM, Simpson AJG, Strausberg RL. **Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing.** *PNAS,* 2003, **100**:4701-4705
46. dos Santos ML, Palanch CG, Salaorni S, Da Silva WA Jr, Nagai MA. **Transcriptome characterization of human mammary cell lines expressing different levels of ERBB2 by serial analysis of gene expression.** *Int J Oncol.* 2006, **28**:1441-1461
47. Fiori JL, Zhu TN, O'Connell MP, Hoek KS, Indig FE, Frank BP, Morris C, Kole S, Hasskamp J, Elias G, Weeraratna AT, Bernier M. **Filamin A modulates kinase activation and intracellular trafficking of epidermal growth factor receptors in human melanoma cells.** *Endocrinology* 2009, **150**:2551-2560
48. Mukherji M, Brill LM, Ficarro SB, Hampton GM, Schultz PG. **A phosphoproteomic analysis of the ErbB2 receptor tyrosine kinase signaling pathways.** *Biochemistry* 2006, **45**:15529-15540.
49. Yi J, Beckerle MC. **The human TRIP6 gene encodes a LIM domain protein and maps to chromosome 7q22, a region associated with tumorigenesis.** *Genomics* 1998, **49**:314-316
50. Wang Y, Dooher JE, Koedood Zhao M, Gilmore TD. **Characterization of mouse Trip6: a putative intracellular signaling protein.** *Gene* 1999, **234**:403-409
51. Xu J, Lai YJ, Lin WC, Lin FT. **TRIP6 enhances lysophosphatidic acid-induced cell migration by interacting with the lysophosphatidic acid 2 receptor.** *J Biol Chem*. 2004, **279**:10459-10468.
52. Chastre E, Abdessamad M, Kruglov A, Bruyneel E, Bracke M, Di Gioia Y, Beckerle MC, van Roy F, Kotelevets L. **TRIP6, a novel molecular partner of the MAGI-1 scaffolding molecule, promotes invasiveness.** *FASEB J*. 2009, **23**:916-928
53. Holt RA, Jones SJ. **The new paradigm of flow cell sequencing.** *Genome Res.* 2008, **18**:839-846.
54. Glisin V, Crkvenjakov R, Byus C. **Ribonucleic acid isolated by cesium chloride centrifugation.** *Biochemistry* 1974, **13**:2633-2637
55. Galante PA, Vidal DO, de Souza JE, Camargo AA, de Souza SJ. **Sense-antisense pairs in mammals: Functional and evolutionary considerations.** *Genome Biol.* 2007, **8**:R40.

Figures

Figure 1 - Schematic view of the alternative splicing library construction with amplification of RNA. I. Oligo dT containing T7 RNA Polymerase recognition site was used for first strand cDNA synthesis with Superscript II that adds cytosine residues after reaching the 5`end of mRNAs. II. This c-rich region serves as anchor for TS-oligo alignment, allowing further polymerization to the end of the oligo. III. Second strand cDNA synthesis using TS-oligo. IV. Amplification of mRNA using T7 RNA Polymerase. V. First strand cDNA synthesis using TS-oligo. VI. Second strand cDNA synthesis using oligodT. VII. Denaturation and renaturation resulting in the formation of heteroduplexes molecules by common exons complementarity. VIII. Single-stranded molecules degraded by Exonuclease (dotted line). IX. DpnII digestion resulting in small cohesive fragments. X. 25mer biotinilated random oligos coupled to streptavidin magnetic beads anneal to single-strand loops. XI. Coupling of specific adaptors to the cohesive ends of the captured heteroduplexes. XII. PCR amplification of fragments using adaptors specific oligos (double line).

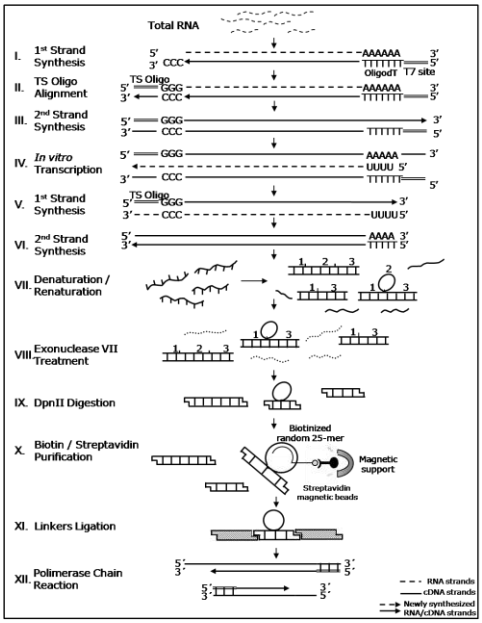


Figure 2 - Flowchart of the bioinformatics pipeline.

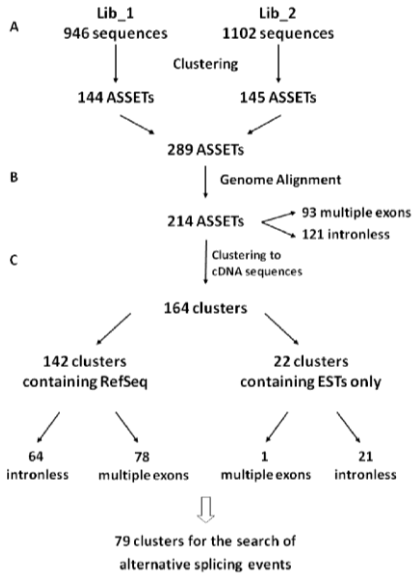


Figure 3 – Relative position of the ASSETs throughout full-length mRNAs. The graphic represents the distribution of the ASSETs along corresponding transcript position. In the x-axis the relative transcript position is shown as a percentage value, where 0 indicates the 5' end and 100 indicates the 3' end. The coverage is the number of ASSETs aligning at each position.

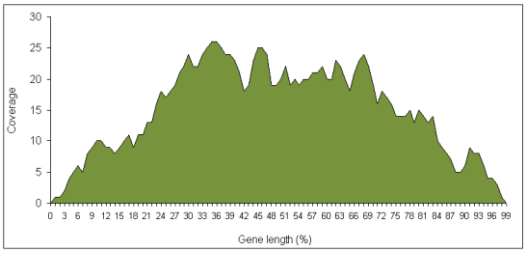


Figure 4 - Graphical view of GO Biological Process overrepresented categories. The graphic is represented in a hierarchical form. The yellow circles correspond to the categories that were statistically significantly enriched.

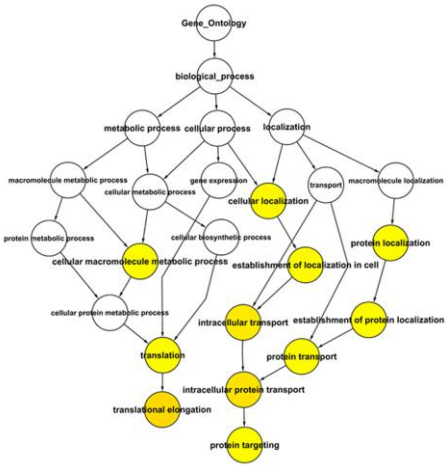


Figure 5 - Characterization of the novel AS variants identified.

The scheme shows the genomic structure and protein domains of the known and putative novel variants. The squares represent the exons, and the lines represent the introns. The dark regions represent the 5'and 3'untranslated regions (UTR), the arrow represents the translational initiation site and the circles represent the stop codons. A – *PTLA* and B – *TRIP6*.

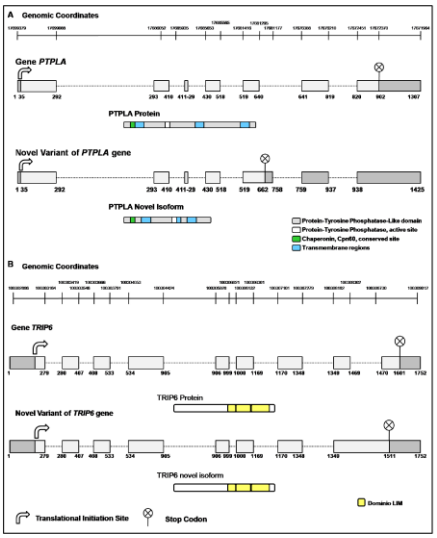
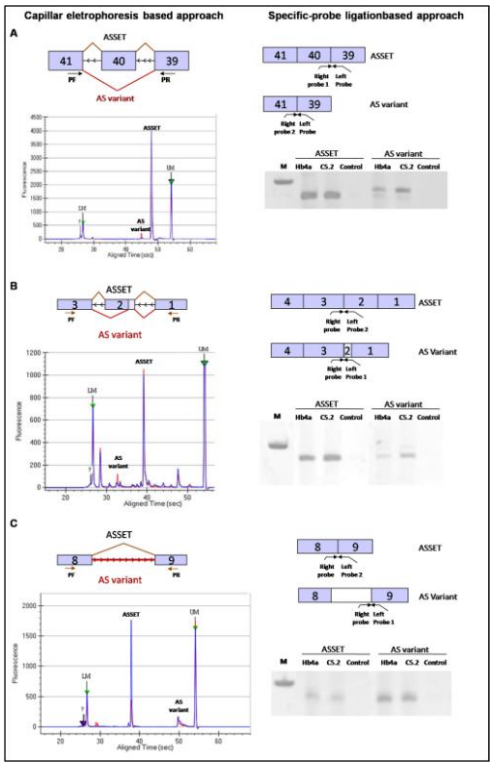


Figure 6 - The influence of *ERBB2*-mediated expression on the regulation of AS variants.

In the left panel, a schema of the microfluidic capillary electrophoresis approach is shown. The exons are represented by numbered squares according to the exons involved in the AS events for each gene. The black arrows represent the primers used for PCR amplification (PF – forward primer; PR – reverse primer). The electropherogram represents the amplification of the AS variants for the Hb4a cell line (blue line) and for the C5.2 cell line (red line). The green arrows indicates internal markers: LM (lower marker) and UM (upper marker). In the right panel, the probe-ligation approach is shown. Each pair of probe is shown for each AS variant separately, with the corresponding PCR products on 8% acrylamide gel. M – 100 bp ladder. A –*SFRS9* gene, B – *FLNA* and C – *TRIP6*.



Tables

Table 1 - Characterization of libraries Lib_1 and Lib_2

Library	# High Quality Sequences	# Contigs	# Singlets	# Consensus	Redundancy
Lib_1	946	96	48	144	84.78%
Lib_2	1102	74	71	145	86.84%
Total	2048	170	119	289	-

Table 2 - Search for AS variants by comparison with sequences from public databases. ***ASSETs selected for RT-PCR validation**

	Presence of alternatively spliced transcripts in databases		No alternatively spliced transcripts in databases	
Lib_1	*ATP1A1**	*MAN1B1*	*CDC42SE1*	*MAN2A1*
	ATP5A1	*NAP1L1*	*CDK5RAP2**	*OSBPL8*
	C6orf108	*PPIB*	*DDB2*	*PSMD6*
	CAMK2G	*RANBP1*	*EEF2*	*PTPRA*
	CD320	*RPL28*	*FARS2*	*188268*
	CTSH	*RPL6*	*GABARAP*	*RBM10**
	ELF3	*RPS4X*	*GNB3*	*RNF149*
	*FLNA**	*SETD2*	*GRK6*	*ROCK2*
	GAPDH	*SFRS9**	*HDAC2*	*RPL11*
	GNAS	*STK25*	*ITGB5**	*THSD1*
	GNPTAB	*UQCRC1*		
Lib_2	*ALDH3A2**		*ACLY*	*MRPL45*
	*AOF2**		*ASCC3L1*	*PHF19*
	CCNB1		*C7orf55*	*PTPLA**
	CREB3		*COL7A1**	*RBMX*
	DNAJC10		*DDEF1*	*SGSM2*
	*FN1**		*DENND4C*	*SLC4A2*
	INTS9		*GDF9*	*TRIP6**
	MYO1C		*KIAA0090*	*XPO1*
	*RPS2**		*KIAA0152*	
	RPS5			
	SEC61G			
	ST13			
Lib_1 & Lib_2	*CLTC**		*ATXN10*	
	*EIF4A3**		*INPP1*	
	*GSPT1**		*PAIP1*	
	*KRT18**			
	*PSMC2**			

Table 3 - Functional classification of genes within the statistically significant biological process categories.

GO-ID Description	Corrected p value	Gene Symbol
Translation Elongation	1.67E+01	*RPL6 RPL21 EEF2 RPL11 RPS4X RPS2 RPS5 RPL28*
Intracellular Protein Transport	4.21E+01	*XPO1 CLTC GABARAP KRT18 YWHAH NUP62 ZFYVE16 KPNA6 RPL11 MRPL45 SEC61G SEC61A1 SRP9*
Intracellular Transport	7.66E+01	*XPO1 MYO1C CLTC GABARAP YWHAH KRT18 NUP62 ZFYVE16 SEC22B KPNA6 RPL11 RANBP1 GNAS MRPL45 SRP9 SEC61G SEC61A1*
Cellular Localization	2.55E+02	*XPO1 MYO1C VIL2 CLTC GABARAP YWHAH KRT18 NUP62 ZFYVE16 SEC22B KPNA6 GNAS RPL11 RANBP1 MRPL45 SRP9 SEC61G SEC61A1*
Establishment of Localization in Cell	3.06E+02	*XPO1 MYO1C CLTC GABARAP YWHAH KRT18 NUP62 ZFYVE16 SEC22B KPNA6 RPL11 RANBP1 GNAS MRPL45 SRP9 SEC61G SEC61A1*
Cellular Macromolecule Metabolic Process	3.81E+02	*PPP6C XPO1 UQCRC1 CAMK2G PTPLAD1 FARS2 DNAJC10 MAN1B1 RPS2 RPL6 PTPLA RPL11 PSMD6 DNAJA3 GLT25D1 STK25 ROCK2 PAIP1 PTPRA ZDHHC7 AXL MOBKL1A EEF2 RPS4X RPS5 RPL28 IFNAR1 CCNB1 MGAT1 ST13 SENP1 HDAC2 GSPT1 PPIB RPL21 PSMC2 DDB2 GRK6 MRPL45 CTSH*
Protein Targeting	3.81E+02	*XPO1 ZFYVE16 KPNA6 RPL11 GABARAP SRP9 SEC61G*
Protein Localization	3.81E+02	*XPO1 VIL2 CLTC GABARAP YWHAH KRT18 NUP62 ZFYVE16 SEC22B KPNA6 RPL11 GNAS MRPL45 SEC61G SEC61A1 SRP9*
Translation Elongation	3.81E+02	*GSPT1 RPL6 RPL21 PAIP1 FARS2 EEF2 RPL11 RPS4X RPS2 MRPL45 RPS5 RPL28*
Protein Transport	3.81E+02	*XPO1 CLTC GABARAP YWHAH KRT18 NUP62 ZFYVE16 KPNA6 SEC22B RPL11 GNAS MRPL45 SEC61G SEC61A1 SRP9*
Establishment of Protein Localization	3.81E+02	*XPO1 CLTC GABARAP YWHAH KRT18 NUP62 ZFYVE16 KPNA6 SEC22B RPL11 GNAS MRPL45 SEC61G SEC61A1 SRP9*

Table 4 – RT-PCR validation. Library Origin: library where the ASSET was captured from; selected ASSETs: number of randomly selected ASSETs; ASSET validation: to confirm the presence of the ASSET in the same RNA used for library construction; Heteroduplex validation: amplification of alternatively spliced transcripts that could have participated in the heteroduplex formation; and cross-validation: amplification of ASSETs specifically captured from one library using the cDNA template from the other library.

Library Origin	Selected ASSETs	ASSET validation	Heteroduplex validation	Cross-validation
Lib_1	6	5	2	4
Lib_2	7	7	4	7
Lib_1 & Lib_2	5	5	0	-

Table 5: Gene expression analysis under the influence of *ERBB2* over-expression.

Gene Symbol	Cell line	Variant	Size (bp)	Concentration (ng/ul)	Normalized Concentration	ASSET/Variant	C5.2/Hb4a
SFRS9	Hb4a	ASSET	232	6.5	3.7	36.2	-4.8
		variant	100	0.2	0.1		
	C5.2	ASSET	232	6.4	2.9	7.5	
		variant	100	0.9	0.4		
FLNA	Hb4a	ASSET	500	24.9	14.1	95.7	-4.8
		variant	377	0.3	0.2		
	C5.2	ASSET	500	23.0	10.4	20.0	
		variant	377	1.2	0.5		
ALDH3A2	Hb4a	ASSET	470	20.0	11.4	10.0	1.0
		variant	610	2.0	1.1		
	C5.2	ASSET	470	24.6	11.0	10.3	
		variant	610	2.4	1.0		
TRIP6	Hb4a	ASSET	203	12.6	7.2	8.9	-3.4
		variant	636	1.4	0.8		
	C5.2	ASSET	203	3.5	1.6	2.6	
		variant	636	1.3	0.6		
PTPLA	Hb4a	ASSET	324	10.8	6.1	83.1	1.2
		variant	456	0.1	0.1		
	C5.2	ASSET	324	16.8	7.5	98.4	
		variant	456	0.2	0.1		
RPS2	Hb4a	ASSET	187	13.7	7.8	1.4	1.1
		variant	390	12.2	5.5		
	C5.2	ASSET	187	1.8	1.0	1.6	
		variant	390	1.4	0.6		

Table 6: Clinical characteristics from the ductal carcinoma samples. Age: age of diagnosis; TNM: classification according to TNM (T – size; N – lymph node status; M – presence of metastasis); LN: involvement of sentinel lymph nodes; grade: grades I, II and III according to SBR; molecular markers: ER – estrogen receptor; PR – progesterone receptor; p53 – protein TP53; and ERBB2 – protein ERBB2.

Sample	Stage	Age	TNM	LN	Grade	Molecular Markers
9T	IIa	55 years	T2N0M0	Negative	Grade I SBR	ER +/ PR +/ p53 -/ ERBB2+ (3+)
20T	IIb	87 years	T2N0M0	Negative	Grade II SBR	ER +/ PR -/ p53 -/ ERBB2+ (3+)
22T	IIb	56 years	T2N1M0	Positive	Grade III SBR	ER +/ PR -/ p53 -/ ERBB2+ (2+/3+)
28T	IIIa	42 years	T2N2M0	Posi*ve*	Grade II SBR	ER +/ PR -/ p53 -/ ERBB2+ (3+)
36T	I	45years	T1N0M0	Negative	Grade III SBR	ER +/ PR -/ p53 -/ ERBB2+ (3+)

Anexo B

Global transcriptome analysis by parallel sequencing for the assessment of *ERBB2*-mediated gene activation in breast cancer

Dirce Maria Carraro[1*], Elisa Napolitano e Ferreira[1,2], Gustavo Molina[1], Eduardo Abrantes[1,2], Renato Puga[1], Adriana Priscila Trapé[4], Diana N Nunes[3], Maria Mitzi Brentani[4], Wadih Arap[3], Renata Pasqualini[3], Helena Brentani[1], Emmanuel Dias-Neto[1,3,5] and Ricardo Renzo Brentani[1].

1 – Centro de Pesquisas do Hospital AC Camargo. Rua Prof. Antonio Prudente, 211 – 01509-900, São Paulo, SP, Brazil. 2 – Instituto de Biociências – Rua do Matão, trav. 14, nº 321 – 05508-090, Cidade Universitária, São Paulo, SP, Brazil. 3 - David H. Koch Center, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, USA. 4 – Departamento de Radiologia, Faculdade de Medicina da Universidade de São Paulo, Av. Dr Arnaldo, 455 – 01246-903, São Paulo, SP, Brazil. 5 – Instituto e Departamento de Psiquiatria, Faculdade de Medicina da Universidade de São Paulo, R. Dr. Ovídio Pires de Campos, 785 - 01060-970, São Paulo, SP, Brazil.

ABSTRACT

Parallel tagged sequencing provides an excellent opportunity to investigate the transcriptome. Herein we report for the first time a pervasive analysis, both quantitatively as well as qualitatively, of the entire transcriptome of two human mammary cell lines, differentially expressing *ERBB2,* an oncogene over expressed in 30% of breast cancer. A total of 1,189,693 reads was obtained by 454-Roche sequencing using a new method based on *Dpn*II restriction and specific 4-nt barcoding adapters for each cell line. A single sequencing run matched 11,366 human genes, corresponding to 40.75% of the complete human repertoire and covering 23.48% of the potential nucleotide sequence. Comprehensive bioinformatics analysis and experimental validation revealed novel genes, splicing variants, Single Nucleotide Polymorphisms (SNPs) and gene fusions. All qualitative findings were corroborated by RT-PCR and sequencing of cDNA and genomic DNA with the help of specific primers. As proof of principle, *ERBB2*-over expression was confirmed by the comparison of tagged sequences of RNA-seq. Quantitative findings were confirmed by RT-PCR of representative genes from high and low *ERBB2*- expressing cell lines or human breast cancer specimens and from cell extracts before or after rapamycin treatment. Alterations in expression levels of genes, such as *LOX*, *ATP5L, GALNT3* and *MME* revealed by RNA seq were confirmed both in cell lines and tumor specimens with different *ERBB2* backgrounds. No increase in inter and intra-chromosomal gene fusion mediated by *ERBB2* over expression was observed indicating no increase in genome instability driven by the oncogene. Enrichment in the alternative exon usage category was observed in *ERBB2* over expressing cells, suggesting changes in alternative splicing regulation mediated by the oncogene. This single-nucleotide resolution approach seems suitable for structural and quantitative analyses of complex transcriptomes, even when amplified RNA is used extending its applicability to laser microdissected samples where the quantity of RNA is restricted.

INTRODUCTION

Global transcriptome analysis is the most effective approach for identification of changes in gene activation profiles. In cancer and other complex pathologies, surveying the transcriptional landscape by the simultaneous interrogation of thousands of genes leads to the definition of key defective elements. cDNA microarrays have unquestionably contributed to understanding breast cancer (Perou et al. 2000; Sørlie et al. 2001; Veer et al. 2002; Brentani et al. 2005; Folgueira et al. 2005; Castro et al. 2008; Rozenchan et al, 2009; Koike Folgueira et al., 2009), allowing quantification of differences in the transcriptional repertoire. Despite some limitations, such as lack of sensitivity to detect small expression differences (Johnson et al . 2003; Brentani et a., 2005; Gardina et al. 2006; Norris and Kahn 2006; Castle et al 2008; Du et al. 2009) and analysis restricted to immobilized probes, the ability of cDNA microarrays to assess multiple samples and consequently estimate tumor biological variability has placed this methodology as a most useful tool for comparative transcriptome analyses in cancer.

Next-generation sequencing (NGS), providing hundreds of thousands of reads in a single run, is enabling the replacement of hybridization-based gene expression methods by sequence-based approaches, since digital transcript-counting approaches prevail over microarray restrictions in several aspects (Torres et al. 2008; Maher et al. 2009a; Tang et al. 2009; Yassour et al. 2009; Pleasance et al. 2010) and became appropriate for simultaneously evaluating multiple samples (Meyer et al. 2008). Moreover, the recently revealed inter-related and high complex structure of the mammalian transcriptome (Graveley BR 2001; Harrow et al. 2006), requires more sensitive investigations of both quantitative features and qualitative architectural aspects of gene expression profiles, which can only be provided by NGS methodology.

By combining *Dpn*II-restriction and parallel tagged sequencing, we report here a base-pair resolution method that permits the simultaneous comparison of whole transcriptomes of multiple samples. Parallel tagged sequencing of two human mammary cell lines HB4a (Stamps et al. 1994) and C5.2, its *ERBB2*-overexpressing counterpart (Harris et al. 1999) in the Roche-454 platform, allowed assessment of *ERBB2* driven transcriptional alterations. *ERBB2* (ENSG00000141736) is an oncogene highly expressed in 25-30% of human breast cancers and its amplification has long been considered to play a crucial role in the malignant transformation. No alteration in the qualitative aspects mediated by *ERBB2* level was observed. However expression of novel genes revealed to be influenced both in cell lines and tumor specimens by the *ERBB2* level. Moreover novel genes, splicing variants, Single Nucleotide Polymorphisms (SNPs) and gene fusions sequences were detected in the RNA-seq of both cell lines. Overall our approach seemed suitable for interrogating the whole transcriptome of multiple samples and also proper for amplified RNA, extending its applicability to conditions in which RNA occurs in limited quantities.

RESULTS

Whole transcriptome sequencing and analysis of poly A^+ and amplified RNA-derived datasets

Double stranded cDNA, converted from purified poly A^+ and amplified mRNA, from HB4a and C5.2 cell lines was prepared separately, *DpnII*-digested, specifically labeled with a 4nt- barcode and pooled together before sequencing. Two libraries, poly A^+ and amplified RNA (Fig. 1), were generated and submitted to Titanium and FLX 454-Roche platform runs,

respectively. A total of 1,189,693 sequences was generated; 802,214 from the poly A^+ library (800K dataset - SRA012436.3) with median size of 197bp; and 387,470 sequences (300K dataset) from the amplified RNA library with median size of 244bp (Fig. 2). The 800k dataset obtained from the poly A^+ library was filtered to exclude mitochondrial and ribosomal RNAs, as well as sequences without adapters, resulting in 731,628 sequences (91.2%) (Fig. 2A). From these, 651,058 (88.99%) were aligned to the Human Genome (coverage ≥ 85% and identity ≥ 90%), of which 614,434 (94.4%) were single-hit matches (Fig. 2B). By mapping to genomic coordinates of 15 gene-track annotations available at University of California Santa Cruz, 613,425 sequences (99.83%) aligned to transcriptional units (see methods). Of these, 476,337 transcripts (73.16%) (Fig. 2C) represented 17,887 RefSeq entries, corresponding to 38.93% (17,887 out of 45,946 entries) of the complete transcript repertoire from RefSeq (build 36). In terms of gene representation, this set aligned with 11,366 human genes (Fig. 2C), corresponding to 40.74% of its respective gene repertoire (27,827 genes). Given that 17,887 full-length mRNAs correspond to 60,500,115 nt, the base-pair representation was 23.48% (14,208,089 nt were covered). To estimate transcript coverage we used the dataset with 613,425 sequences (Fig. 2C) and verified the read distribution throughout the length of full-length mRNAs by calculating their relative position as described (Dias Neto et al. 2000). We achieved full gene coverage (Fig. S1A) with a slightly higher concentration of sequences in the central portion of the transcripts. Additionaly, the 5' end of full transcripts was very well represented, irrespective of the original transcript lengths (Fig. S1B). This result depicts high-quality RNA templates, efficient cDNA synthesis during library construction and no bias in the representation of short transcripts. To test whether our approach is theoretically applicable to any RNA source in which the amount of RNA is limited, we used the 300K dataset, produced from one-round amplified RNA (Fig. 1). After applying the filters used for the 800k set, 291,803 (75.31%) sequences remained. Alignment of this set against the Human Genome resulted in 232,552 (79.69%). From those, 157,005 (67.51%) could be aligned against RefSeq (build 36.3) representing 11,117 out of the complete transcript repertoire from RefSeq (24.20%), which in turn matched 7,169 out of 27,827 genes (25.76%). Representation at base-pair resolution covered 13.32% of the putative 42,505,099 nt. Using increasing numbers of randomly selected RNA-seqs from 300K and 800K a simulation of gene representation was performed. The particularly superimposed curves derived from both datasets (Fig. S2) indicated comparable ability in gene representation when one-round amplified RNA is used. Transcript coverage based on the distribution of sequences along their respective full-length mRNAs was also analyzed for the 300k set (Fig. S1C) showing good representation of the 5' end, irrespective of transcript size (Fig. S2D), extending our strategy to amplified RNA.Besides the success in reconstruction of the transcriptome derived from amplified RNA samples, some highly abundant spurious reads were observed. This issue was easily elucidated by our restriction site-based strategy, since *Dpn*II-sites, absent from the original transcripts, were detected in these spurious reads. The adjacent sequence of *Dpn*II-site in the TS primer used for cDNA library construction favored internal priming events (see Fig. 1 and Table S1).

Structural variants found in the 800K dataset: discovery of novel single nucleotide polymorphisms (SNPs) and gene fusions

The 614,434 sequences mapping at a single genome position were used for assessment of novel SNPs, whereas the 80,570 reads with low coverage alignment (coverage ≥ 40% and ≤ 90% and identity ≥ 99.9%) to genome sequences (Fig. 2B), were used to explore gene-fusion candidates. For identifying novel SNPs, parameters and criteria were tested that reached an estimated rate of 90% capability in identifying known SNPs (see methods), revealing 3,532 known SNPs and 138 (3.7%) potentially novel SNPs.

Potentially novel SNPs exhibiting more than one mismatch in a 50bp window or mapped to homopolymers and repetitive sequences were discarded. Eighteen SNPs mapping to exonic regions were subjected to validation by Sanger-sequencing using genomic DNA from both cell lines. A high validation rate (89%) was obtained, and revealed 16 new SNPs (Table S2 and Fig. S3). Fourteen SNPs were identified in both HB4a and C5.2 genomic DNA (12 SNPs in heterozygous and 2 in homozygous). Intriguingly one SNP was detected in heterozygosis in both cell lines, but with an apparent difference in DNA dosage between alleles (*DKK1* gene - ENSG0000010798). Nine of the 16 new SNPs were located in coding regions, and four resulted in non-synonymous amino acid substitutions (nsSNPs) all representing benign amino acid substitutions in protein structure, as assessed by PolyPhen, a tool that predicts putative effects of amino acid substitutions on the structure of a given protein (Ramensky et al. 2002).

From the set of 80,570 sequences with low-coverage genome alignment mapping at multiple locations (Fig. 2B), 40 potential gene fusion or trans-splicing events were detected, which showed perfect alignment against 2 distinct chromosome regions. Of these, 38 matched 2 different chromosomes (inter-chromosomal gene fusions) and 292 matched 2 different regions in the same chromosome (intra-chromosomal gene fusion) (Fig. 2B). Most intrachromosomal rearrangements were intragenic and only 8 were intergenic. From the 284 intrachromosomal intragenic, 18 and 266 reported exon order alteration and strand inversion, respectively. Translocations between strands can occur during replication process by a specific mechanism known as Fork Stalling and Template Swichting Model (FoSTeS), due to nucleotide similarities between DNAs (Gu et al., 2008). Another replication mechanism that can lead to chromosomal rearrangement is non-homologous end-joining (NHEJ). NHEJ is evidenced by detection of microhomology regions or insertion of short sequences in the junction of gene fusion events (Gu et al., 2008). We have detected 288 events with microhomology (31 interchromosomal and 267 intrachromosomal) and 32 events with short inserted sequences in the transcriptome of both cell lines (Fig. 3A).

In order to check whether *ERBB2* amplification could augment genome instability we compared the number of gene fusion events between *ERBB2*-basal (HB4a) and high expression (C5.2) cell lines using the subset of reads with more stringent criteria for tag assignment (see material and methods). A similar number of gene fusions was observed in both cell lines, where 80 events were detected for C5.2 (68 intra and 12 interchromosomal) and 73 for HB4a (62 intra and 11 interchromosomal), indicating no influence of *ERBB2* over expression in genome instability (Fig. 3B). For validation, 12 inter- and 2 intra-chromosomal gene fusions were evaluated via a specific probe-ligation-based technique (Schouten et al. 2002). Two probes designed for each gene fusion were hybridized with cDNA and genomic DNA from HB4a and C5.2 cell-lines, and in the presence of the gene fusion, the probes were joined to yield one single probe that was amplified by PCR. As a negative control we firstly performed hybridization of all pairs of probes in the absence of template, followed by ligation and PCR. Among the 14, two were validated in both cDNA and genomic DNA from HB4a and C5.2 (*FTH1*/chromosome 11 -*EIF5A*/chromosome 17 and *VAMP8*/chromosome 2 - *SACF1*/chromosome 19) (Fig. 3C; Table S3). We assumed that the low level of confirmation may be a consequence of sequestration of probes by non fusioned transcripts, probably vastly more abundant. To test this we counted the number of correspondent reads not reporting fusion events. Interestingly, the 4 genes involved in the 2 validated gene fusion events were among the lowest read counts (Table S3), strengthening our hypothesis.

RNA-seq assessment: discovery of splicing variants and novel genes

Mapping of the 614,434 sequences to the UCSC gene-tracks (Fig. 2B) produced two subsets: a group of 613,425 reads overlapping at least one gene-track and a small fraction of 1,009 reads mapping outside any gene-track position (Fig. 2C). The former subset, aligned against mRNA databases yielded 597,565 sequences, which were used to assess novel splicing variants. A total of 2,865 potentially novel alternative splicing (AS) variants were detected (Fig. 4). To explore the influence of *ERBB2* over expression on alternative splicing regulation we compared the number of AS events of each category identified for cell line, normalized by the total number of reads obtained for each tag (Table S4). Enrichment of alternative splicing events was observed for the alternative exon usage category, where exon skipping ($p<0.001$) and inclusion ($p<0.01$) reported higher number of events in C5.2 cells, suggesting that changes in alternative splicing patterns were mediated by *ERBB2* expression.

For validation we focused on the exon inclusion subcategory (Fig. 4C), since the feasibility of designing primers in the newly included exons leads to more reliable results regarding the expression of specific variants. We selected 20 exon inclusion events from the subset of 89 events containing two known flanking exons (Fig. 4C). Eighteen out of these 20 events (90%) were confirmed as new *bona fide* exon inclusion AS-variants (Fig. S4). For 14 validated events the new exon was positioned within the coding sequence (CDS) suggesting substantial functional alteration of the respective protein. Six resulted in premature stop codons, four lead to amino acid losses and one to amino acid gains. Of the remaining three, one presented changes in functional domains and two completely abolished the open reading frames. Eight that were only detected in C5.2 cells, indicating their over expression in this cell line, were evaluated by quantitative RT_PCR in both cell lines. Six (75%) were in accordance with RNA-seq showing higher expression in the C5.2 (fold >2). Interestingly, for five of them, belonging to the genes *CSRP2BP* (ENSG00000149474), *PRCC* (ENSG00000143294), *CLTC* (ENSG00000141367), *NR2C1* (ENSG00000120798) and *KIAA1033* (ENSG00000136051), only the splicing variants containing these novel exons, seem to be influenced by *ERBB2* over expression (Fig. S4 and Table S5) since the total number of whole gene-related sequences seemed not to be differentially expressed in both cell lines or showed even higher expression in HB4a.

Additionally, the 1,009 sequences mapping out of any gene-track position were explored to find novel human genes (Fig. 2C). The majority of these sequences was composed of single continuous hits (960 sequences, 95%), whereas the minority covered multiple-exon regions (49 sequences). From these 49 sequences, six exhibited canonical splice sites at their introns and were selected for validation confirming three (50%) new human transcripts (Fig. 5). The complete characterization of the novel transcripts remains to be determined.

RNA quantification: *ERBB2*-mediated effects in the transcriptome of a breast cell-line

We determined the number of reads representing each gene, independently of their relative position within the full-length mRNA. A list of 436 potentially differentially expressed genes was obtained, 192 and 244 of which were enhanced in C5.2 and HB4a cells, respectively (Fig. 6A). As a proof of concept, sequence tags representing *ERBB2* mRNA were counted for both cell lines giving a C5.2/HB4a ratio of 15, in agreement with its known over-expression status in C5.2. This result reinforces the accuracy of our quantitative analysis from two standpoints: the effectiveness of this methodology for parallel sequencing of two different tagged cDNA populations and the feasibility of our approach to determine gene activation profiles. To investigate the biological properties of genes modulated by *ERBB2*-mediated expression, the 436 differentially regulated genes were classified according to Biological Processes in Gene Ontology (GO) and also within KEGG database pathways. Using FunNet (Prifti et al. 2008) we identified 12 enriched GO and 7 KEGG categories in both cell lines (Fig. 6). To confirm *ERBB2*-mediated gene regulation, we assessed the expression levels of 88 genes by qRT-PCR; 46 genes (52.3%) presented concordant results in gene expression in C5.2 and HB4a cells (fold ≥2) by both methodologies (Table S6). Moreover, assessment of the expression of these 46 genes was also evaluated in 14 human breast ductal carcinoma samples representing 2 subsets: *ERBB2* over-expression (7 samples) and basal *ERBB2* (7 samples) (Fig. S5). Despite tumor and patient heterogeneity, as well as the gap between cell lines models and patient samples, 4 (8.6%) genes modulated by *ERBB2* in these cell lines could be confirmed in breast tumor samples with distinct *ERBB2* backgrounds: *ATP5L* was increased in the *ERBB2*-positive samples, whereas *LOX* (ENSG00000113083), *GALNT3* (ENSG00000115339) and *MME* (ENSG00000196549) were more abundant in basal *ERBB2* samples.

One of the most important signaling pathways driven by ERBB2 involves the PI3K/AKT pathway that activates the protein kinase mammalian target of rapamycin (mTOR), an important regulator of mRNA translation that controls cell proliferation (Janus et al. 2005). To evaluate whether the genes modulated by *ERBB2*-mediated expression identified in this study were related to mTOR, we treated both cell lines with rapamycin and investigated their expression after treatment. From the 46 validated genes, 19 (41.3%) showed reduction or inversion in relative fold-difference between C5.2/HB4a cells (Table S7). This set included genes that might reflect the effects of certain drugs, such as ribosomal proteins (RPL1, RPL8, RPL29, RPL31, RPL39 and RPS24), cytokeratins (KRT4, KRT6A and KRT7), and proteins participating in glycolysis (PFKP), electron transport chain (NDUFB2, NDUFB3 and UQCRB) and in the ubiquitin ligase system (PBXL6 and TXNIP). Some of these genes have already been reported to be modulated by *ERBB2* and sensitive to rapamycin-treatment (Creighton C J 2007; Heinonen et al. 2008; Akcakanat et al. 2009; Meric-Bernstam and Gonzalez 2009).

DISCUSSION

The approach presented here provided robust data for quantitative and architectural aspects of two mammary cell line transcriptomes, showing elevated coverage of the gene repertoire as well as of the full representation of transcripts. Restriction enzyme digestion revealed several advantages over physical methods for fragmentation of cDNAs, including reduction of overlapping reads (which leads to increased transcriptome representation) and prompt identification of artifacts produced during cDNA library construction. *MALAT1*, one internal priming-affected transcript, has been reported as highly expressed in cancer when RNA amplification methods were used (Loi et al. 2007; Guffanti et al. 2009). This apparently specious result may not be noticed by other approaches. Furthermore, the presence of a restriction site in the fusion boundary of both cDNAs improves the detection of spurious gene fusions. For exploring SNPs, we used stringent bioinformatics and manual inspection that resulted in a high confirmation rate (89%). Some validated SNPs apparently exhibited different dosages between alleles that might be sustained by genomic rearrangements (Davidson et al. 2000). High-throughput transcriptome sequencing has proven a good strategy to define genomic rearrangements (Maher et al. 2009b; Maher et al. 2009; Wang et al. 2009). Here we identified two *bona fide* gene fusions that to our knowledge are reported for the first time. The detection of alternative splicing by our method is enhanced by the longer fragments produced by the 454-platform, compared to other next-generation sequencing technologies. An extrapolation of 90% confirmation rate over the 1,704 novel AS events in multi-exon splicing variants identified results in 1,533 novel AS events, for which conserved splice sites were observed.

ERBB2 oncogene amplification is considered an important tumor driver alteration, rather than being a simple consequence (Di Fiore et al., 1987), and has been reported in

approximately 30% of breast cancers (Slamon et al, 1989). The quantitative transcriptional aspect of over expression of the oncogene has been assessed by 3' end sequence methodology (colocar Nagai e outros). However, only a single nucleotide resolution approach of whole transcriptome sequencing enables the assessment of its structural aspects. No increase in genomic instability was observed in the transcriptional repertoire mediated by *ERBB2* over expression. Instead, replication process disorders such as FosTeS (ref) and NHEJ (ref) were observed in both cells, as detected in many breast cancer samples (Stephens et al., 2009). Additionally, intrachromosomal events were more prevalent than interchromosomal events, in accordance with what was previously observed by a breast cancer rearrangement profiling study (Stephens et al, 2009). *ERBB2* influence was observed in quantitative aspects of breast cell line transcriptomes. Its influence was detected not only on gene expression but also on specific splicing variant transcripts. Enrichment of both exon skipping and inclusion of alternative exon usage by *ERBB2* over expression observed in C5.2 cells reinforced the potential action of the oncogene in influencing regulation of the splicing process. Evidences in the same direction have been shown where activation of proteins, such as *AKT* and *MAPK* by deregulation of the *ERBB2* signaling pathway may phosphorylate and activate specific splicing factors, changing the alternative splicing balance of cells (Srebrow et al., 2006).

To highlight the biological relevance of our findings, suffice it to say that a considerable number of over-expressed genes in C5.2 cells are from the glycolytic and pentose pathways, as well as the Krebs cycle and the oxidative phosphorylation chain. The Warburg effect purported to explain these findings, but a novel interpretation (Vander Heiden et al. 2009) attributes such over-expression to the increased demands on carbon chains, nitrogen and reducing power required by enhanced cell division, a hallmark of cancer cells. Furthermore, decades ago, we demonstrated that the nucleolus was involved in mRNA processing (Brentani et al. 1967), data recently confirmed by studies in nucleolar protein-deficient yeast mutants (Schneiter et al. 1995; Ideue et al. 2004), and that the hormonal induction of new enzymes required new ribosomes (Da Silva et al. 1974). The over-expression of genes coding for several ribosomal proteins can therefore be easily explained.

The intrinsic molecular heterogeneity found between distinct human tumor samples, as well as within a single breast tumor sample has been reported by many laboratories (Perou et al. 2000) (Stingl and Caldas 2007). Additionally these differences appear to be strongly dependent upon microenvironmental factors (Allinen et al. 2004; Rozenchan et al. 2009). Despite the difference in molecular characteristics between cells *in vivo* and cell lines *in vitro* our approach allowed us to identify 6 genes, the expression of which is likely mediated by *ERBB2*. We highlight *LOX* downregulated in C5.2 cells as well as in tumor samples over-expressing *ERBB2*. Furthermore, our data also revealed a higher level of *LOX* expression in C5.2 cells after exposure to rapamycin (4-fold changes), indicating that *LOX* is a target potentially regulated by the ERBB2/mTOR pathway. *LOX* encodes an extracellular copper-requiring enzyme that initiates collagens and elastin crosslinking and enhances tumor cell invasion and metastasis (Noblesse et al. 2004). The 18-kDa LOX propeptide was found to be an effective inhibitor of the more invasive phenotype of breast cancer cells driven by *ERBB2* and has been suggested as a target for therapy in this subtype of breast cancer (Min et al. 2007).

Altogether, the results presented here demonstrated that our approach is suitable for interrogation of the whole transcriptome of multiple samples in parallel sequencing by the 454-ROCHE platform, from which an accurate quantitative and qualitative portrait of complex transcriptomes can be generated. Furthermore, our approach is also suitable for amplified RNA, and is therefore especially important under conditions where limiting amounts of RNA are available.

METHODS

Cell lines and tumor samples: treatment and RNA purification

Two human breast cell lines, HB4a and C5.2, were cultured as described (Harris et al. 1999). HB4a is derived from human mammary luminal epithelial cells that express basal levels of *ERBB2*. C5.2 is derived from HB4a cells transfected with four copies of full-length *ERBB2,* which is expressed at high levels (Stamps et al. 1994). For rapamycin treatment, both cell lines were plated in 25 cm^2 flasks and maintained at 40-50% confluence. Cells were treated with vehicle 0.01 % (v/v) absolute ethanol/ medium (control) or 20 ηM rapamycin for 24 hours. Breast tumor samples were retrieved from the A. C. Camargo Hospital Biobank. Fresh-frozen tumor blocks were cut, fixed and stained with hematoxylin and eosin (H&E) and reviewed by a pathologist. The H&E-stained sections were used to select tumor areas isolated from the rest of the samples with surgical blades. All specimens used in this study contain signed informed consents and are in accordance with the Hospital A.C. Camargo Ethics Committee (#952/07). Total RNA was extracted with Trizol and treated with DNaseI (Ambion) as specified by the manufacturer. RNA quality was assessed by a Eukaryote Total RNA 2100 Bioanalyzer (Agilent Technologies). Samples were classified as ERBB2 high or basal expression according to protein and mRNA levels. Protein was evaluated by imunohistochemistry and signals 2+ or 3+ was considered ERBB2 high expression, and signals 0 or 1+ were considered basal expression (Signal 0: no staining or membrane staining in < 10% of tumor cells; signal 1+: faint/barely perceptible membrane staining in > 10% of tumor cells, and partially membrane staining; signal 2+: weak to moderate complete membrane staining in > 10% of tumor cells; signal 3+: strong complete membrane staining in > 30% of tumor cells). *ERBB2* transcript was evaluated by quantitative RT-PCR. Basal or high *ERBB2* expression was denoted when samples displayed relative expression level below or above the average value among all samples, respectively. Only samples with concordant results at both protein and transcript expression level were included in the study.

Whole-transcriptome libraries and Roche-platform sequencing

Poly A libraries: mRNA poly A^+ was purified from 40 µg DNA-free RNA from HB4a and C5.2 cells using µMACS™ mRNA Isolation Kits (Miltenyi Biotec, USA). Two hundred nanograms of mRNA poly A^+ was incubated with 0.5 µg oligo-dT containing a *Dpn*II restriction site [5'GAGGCGGGATCT(30)3']. First and second strand cDNA synthesis were carried out as described[11]. Purified dscDNA was digested with 25 units of *Dpn*II at 37^{o}C for 3 hours. Next, Y-shaped DNA adapers (Watahiki et al. 2004) were added to dscDNA fragments. The HB4a and C5.2 Y-shaped adaptors were formed by primers A and B and primers C and D, respectively (Primer A: 5'-GATCTCCCGAGTGGTCACCTGCTC-3'; Primer B: 5'-CTAGCAGCTACCACTCGGGA-3'; Primer C: 5'-GATCCCCTGAGTGGTCACCTGCTC-3', and Primer D: 5'-CTAGCAGCTACCACTCAGGG-3').

One hundred and fifty nanograms of each adapter were added to *DpnII*-restricted dscDNA with 2,000 units of T4 DNA ligase (New England Biolabs) at 16^{o}C overnight. Fragments ranging from 150bp to 600bp were size selected by 1.5% low point melting agarose gel electrophoresis. One tenth of each purified products was used as a template in a 20-cycle PCR amplification, with 2 units Platinum *Taq* DNA Polymerase High-Fidelity, 0.2 mM dNTPs, 2 mM $MgCl_2$, and 5 *p*mol of forward (5'GAGCAGGTGACCACTC3') and 5 *p*mol of reverse (5'CTAGCAGCTACCACTC3') primers. PCR products were quantified using Nano Drop 1000 and verified in 1% ethidium-bromide-agarose gel.

***Amplified cDNA libraries**:* Two micrograms of total RNA were incubated with 0.5 µg oligo dT-T7, containing the T7 RNA polymerase and *Dpn*II recognition sites 5'GGCCGATGAATTGTAATACGACTCACTATAGGGAGGCGG GATCT(30)3'), at 70^{o}C for 10 min. Reverse transcription was carried out as described (Castro et al. 2008), in the presence of 1.5 µg of the Template-Switch (Matz et al. 1999) (TS) primer containing *Dpn*II restriction site (5'AAGCAGTGGTAACAACGCAGAGATCGGGCGGG3').
Second strand synthesis was performed in 1X Advantage Polymerase Mix (Clontech Laboratories), 0.2 mM dNTPs, 2 units RNaseH (Invitrogen-Life Technology) and 1X Polymerase Buffer. The reaction was incubated successively at 37^{o}C for 10 min, 94^{o}C for 2 min, 62^{o}C for 3 min, 68^{o}C for 15 min and 73^{o}C for 30 min. Purified dscDNA was transcribed *in vitro* with the RiboMax^{TM} Large Scale RNA Production System T7 Kit (Promega Corporation) according to the manufacturer' instructions. Amplified RNA was purified with TRI Reagent (Sigma), quantified and verified by Eukaryote Total RNA 2100 Bioanalyzer. First-strand cDNA was synthesized as described above, with the TS primer for first-strand synthesis and oligodT for second-strand synthesis. The cDNA digestion, linker ligation and PCR amplification were carried out as described above for poly A^{+} libraries. Before 454-ROCHE sequencing, libraries were submitted to a validation in which 1% of the PCR products from the HB4a and C5.2 libraries were pooled and cloned in DH10B-ultra competent *E. coli*. About 1,500 individual clones were sequenced on an ABI3130 instrument.
Deep sequencing: 1.5 µg and 2.0 µg of each cDNA population (HB4a and C5.2 poly A^{+} and amplified cDNA libraries, respectively) were pooled together and were submitted to Titanium and FLX 454-Roche platform sequencing, respectively.

Bioinformatics analyses

454-Roche Titanium and 454 FLX reads were screened for the presence of adapters and reads lacking adapters and/or internal adapters were removed. Next, sequences with high similarity (E-value ≤ 1x10^{-20}, identity ≥ 85% and coverage ≥ 85%) to human ribosomal RNA or mitochondrial DNA were filtered by MEGABLAST. Remaining reads were aligned against the human genome (release hg18, March 2006) through BLAT (Kent W J 2002) and best alignments were selected by pslReps tool with the following parameters: minCover = 0.70, minAli = 0.96, nearTop = 0.005. Reads with significant hits at multiple genome locations were discarded.
The KnownGene (Hsu et al. 2006) annotation track coordinates from the University of California Santa Cruz genome database (UCSC) were used as a reference for mapping the reads in relation to annotated transcripts and exons. Overlapping reads with any base of a transcriptional unit (TU), defined here as the complete genomic sequence between the first and last base of a transcript. To identify splice-spanning reads and to build an exon-exon junction database, we looked at all gaps observed in 454 read alignments against the human genome. Gaps spanning more than 30 pb and having 5' and 3' dinucleotides following the GT-AG rule for donor-acceptor splice sites were considered true intron sequences and the corresponding genome coordinates of exon-exon junctions were annotated. These genome coordinates were used to identify putative splicing events not annotated in the KnownGene annotation track. Putative alternative events were classified into 3 major categories: : alternative exon usage (exon skipping and exon inclusion), intron retention and alternative donor/acceptor site.

For single nucleotide polymorphisms (SNPs) Blast-like BLAT alignment outputs were parsed by the use of a Perl script. A set of parameters such as base coverage, proximity to exon-intron boundaries, proximity to alignment ends, number of different sequenced bases for a specific genome location, and ratio between divergent base and reference base was used to select putative SNPs. As two distinct libraries were sequenced, identification of a putative SNP in both libraries was used as an additional criterion of confidence. Characterized and novel SNPs were identified with the dbSNP (build 129) (Sherry et al. 2001). For quantitative analysis stringent criteria were used for DNA barcoding tag identification to assure accurate sample assignment. Tag assignment was only valid when DNA was flanked by a 5'-adapter sequence and a 3'-restriction enzyme site. For all comparative analysis between HB4a and C5.2 transcriptomes a subset of properly tagged sequences generated from the 800K dataset was used. This set comprised 410,788 sequences, where 188,382 (45.8%) and 222,406 (54.14%) corresponded to HB4a and C5.2 respectively, indicated that pooling equivalent numbers of cDNA molecules from each cell line was efficient. To analyze the differential gene expression profile between HB4a and C5.2 cell lines, we first aligned confident reads against the RefSeq database with the MegaBlast tool. Reads with significant alignments (E-value ≤ 1x10-15, identity ≥ 96% and coverage ≥ 90%) to transcripts derived from different genes were excluded. The overall read count *per* gene was scaled to reads per million (RPM) and differential expression was calculated as the ratio of C5.2RPM/HB4aRPM. We used the SAGEbetaBin statistical approach to assign a significance cutoff value (SAGEbetaBin (Vêncio et al. 2004).

Validation of SNPs

Genomic fragments containing transcripts with putative SNPs or mutations were recovered after the alignment of the ESTs with the human genome using BLAT. Repetitive sequences were removed with RepeatMasker. Primers were designed by Primer3 and were used for PCR amplification with DNA from HB4 and C5.2 cell lines. The amplicons were evaluated in 3% agarose gels and were sequenced in an ABI 3130xL (Applied Biosystems, Foster City, CA, USA).

Validation of gene fusions

For gene fusion confirmation a probe-ligation-based approach was used. Two probes (left and right) were designed for each putative gene fusion events. The left probe was complementary to one of the genes involved in the event, exactly at the limit of the fusion; the right probe directly adjacent to the left probe, was complementary to the other gene involved in the putative event. In addition, the left probe contained at its 5' end a recognition sequence of the forward PCR primer (5'GGGTAGGCTAAGGGTAGGA3'). The right probe was phosphorylated at its 5' end and contained a recognition sequence of the reverse PCR primer (5'TCTAGATTGGATCTTGCTGGCAC3') at its 3' end (Table S3). The probes were hybridized to pre-heated double-stranded cDNA and genomic DNA from HB4a and C5.2 cells at 54^{o}C for 12 hours. The two probes hybridized to their target sequence, were subsequently ligated by Ligase-65 (MC Holland), to form a single probe. Next, only the ligated probes were PCR amplified. As a negative control, hybridization in the absence of any template was performed for all probes and the reaction was submitted to PCR. A control PCR reaction without template was also performed. Amplification products were analyzed on 8% acrylamide gels and were sequenced on the ABI3130 instrument (Applied Biosystems).

Validation of splicing variants

Primers were designed at the respective novel exon and at one adjacent exon (Table S7). cDNAs converted from 40 ng of DNaseI-treated (Ambion) total RNA from HB4a and C5.2 cells was used in each reaction. PCR was performed in a total volume of 20 µl, 1 X reaction buffer, 2.5 mM $MgCl_2$, 0.2 mM dNTP, 10 pmoles of each primer, and 1 unit Platinum Taq DNA Polymerase (Invitrogen-Life Technology) in 40 cycles at 95^{o}C for 30 sec, 60^{o}C for 30 sec, and 72^{o}C for 30 sec, followed by a final extension at 72^{o}C for 7 min. Amplification products were visualized on 8% polyacrylamide gels and were sequenced on an ABI3130 instrument (Applied Biosystems). For quantitative analyses, PCR amplification with the same pair of primers was performed with the ABI Prism 7900 Sequence Detection System (Applied Biosystems) in 20 µl 1 X SYBR Green PCR

MasterMix (Applied Biosystems), containing 2-8 pmoles of each primer and cDNA converted from 100 ng total RNA.

Validation novel genes

Primers for the validation of 6 putative novel genes were designed at two distinct exons with the 454-read as a reference sequence (Table S9). cDNA converted from 40 ng of DNaseI-treated (Ambion) total RNA from HB4a and C5.2 cells was used in each reaction. PCR reaction was performed in 20 µl containing 1 X buffer, 2.5 mM $MgCl_2$, 0.2 mM dNTP, 10 pmoles of each primer, and 1 unit Platinum Taq DNA Polymerase (Invitrogen-Life Technology) incubated at 95°C for 30 sec, 60°C for 30 sec, and 72°C for 30 sec for 40 cycles, followed by a final extension at 72°C for 7 min. Amplification products were visualized on 8% polyacrylamide gels and were sequenced on an ABI3130 instrument (Applied Biosystems).

Validation of differential gene expression

Two micrograms of DNase-treated total RNA from HB4a and C5.2 cells, exposed to Rapamycin or not, and from 14 breast tissue samples were reverse-transcribed with 0.5 µg oligo-dT in the presence of 400 units SuperscriptIII (Invitrogen). cDNA converted from 400 ng of total RNA was used as a template for the evaluation of 96 distinct transcripts (target genes and endogenous controls) in duplicate. Expression levels of selected genes were verified by quantitative RT-PCR with customized low-density TaqMan arrays (Applied Biosystems) in an ABI7900 instrument. A total of 91 genes (75 and 16 up-regulated genes in C5.2 and HB4a, respectively) were conducted; *GUSB* was selected, from the 5 endogenous genes tested, as a reference gene. Differential expression levels were considered significant that exhibited a fold-change>2, and determined by the $2^{-\Delta\Delta Ct}$ method. The list of selected genes is presented in Table S6.

Gene Ontology (GO) and Kegg pathways annotation

Functional annotation was performed on differentially expressed genes according to Gene Ontology Biological Process and KEGG databases. FunNet tools were used for computation of the enriched GO and KEGG categories (Prifti et al. 2008). Significant themes were calculated for up- and down-regulated genes, with the 11,366 represented genes as reference set. A decorrelated annotation procedure was performed by application of the Fisher exact test using corrected *p*-values.

FIGURE LEGENDS

Figure 1. Schematic representation of cDNA libraries. On the left panel, the approach used for double-stranded cDNA synthesis from purified poly A^+ RNA is shown. An oligodT containing the *Dpn*II restriction site was used for first strand synthesis. Second-strand synthesis was performed with RNase H, DNA polymerase and T4 DNA ligase. On the right panel, the methodology for mRNA amplification and double-stranded cDNA synthesis from total RNA is shown. An oligodT containing the *Dpn*II restriction site and also a T7 RNA Polimerase Recognition Site (T7 R.S.) was used for first strand synthesis, in the presence of TS primer. When the reverse transcriptase enzyme reaches the 5'end of the RNA it adds cytosine residues at the 3' end of the cDNA strand. The cytosine residues are complementary to the 3' end of the TS Primer allowing the hybridization of the TS primer with the cDNA first strand synthesis. The TS primer is then used for second strand cDNA synthesis. The resulting double stranded cDNA has a TS primer at the 5' end and *Dpn*II restriction site followed by a T7 RNA polymerase recognition site at its 3' end. The *in vitro* transcription is perfomed using the T7 RNA Polymerase, after recognition of the T7R.S. The amplified antisense RNA is reverse transcribed using the TS primer and the second strand synthesis is performed with oligodT primer containing *Dpn*II restriction site. The *Dpn*II digestion, coupling of linkers and PCR are common to both approaches.

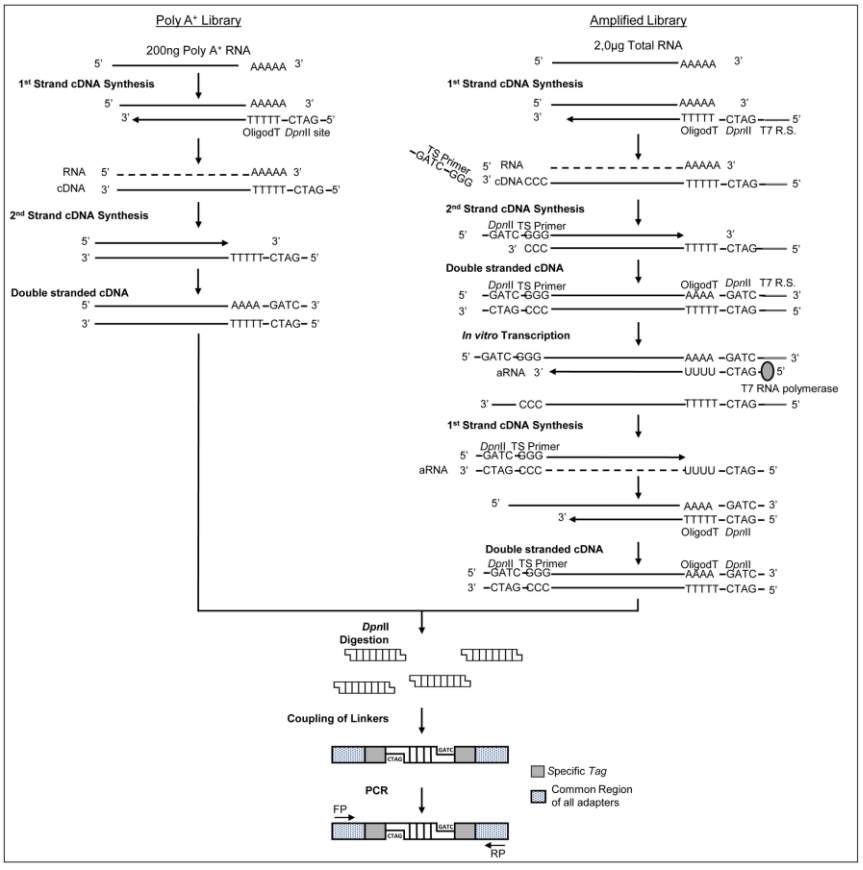


Figure 2. Flowchart of the bioinformatics analyzes. The reads from the 800K and 300K datasets were analyzed independently. High Coverage Genome Alignment corresponds to reads that aligned to Genome Sequence using the criteria of coverage ≥ 85% and identity ≥ 90%. Sequences aligning to more than one genome region following these criteria were discarded. Single-hit high coverage genome alignment sequences were used for discovery of novel SNPs. Low Coverage Genome Alignment corresponds to reads that aligned to Genome Sequence using the criteria of coverage ≥ 40% and ≤ 90% and identity ≥ 99.9. These reads were used for discovery of gene fusion events. For transcriptome analysis sequences from 800K dataset were aligned at 15 different gene tracks, and further aligned to Known gene databases for the discovery of novel splicing variants. Lastly, sequences from both datasets, 800K and 300K, were aligned to RefSeq databases for obtaining the number of transcripts (mRNAs) and genes identified by each dataset and also for analyzing the distribution of the reads throughout the length of full-length mRNA by calculating their relative position.

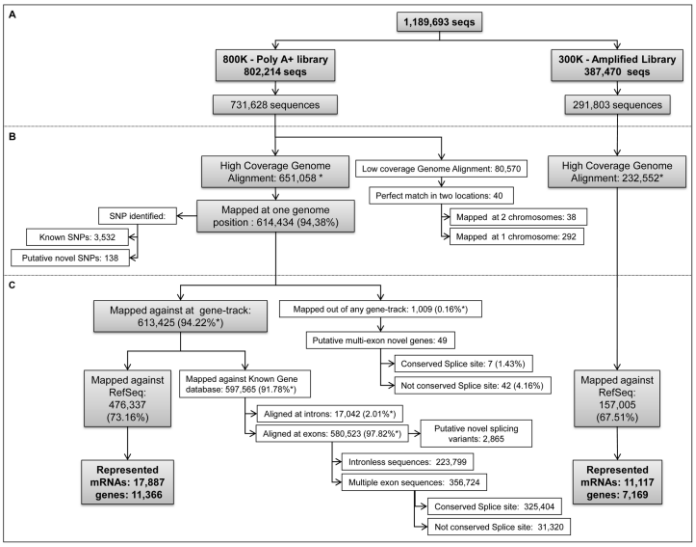


Figure 3. Identification and validation of gene fusion events. (A) The number of gene fusion events presenting microhomolgy or short inserted sequences for interchromossomal and interchromossomal fusions. (B) Number of intra and interchomossomal gene fusion identified for each cell line, C5.2 (left panel) and HB4a (right panel). (C) Validation of 2 gene fusions. The exon distribution of the original genes is represented by the numbered squares, and the regions involved in the fusion are represented by the colored lines. The grey lines represent the regions of the probes used for PCR primer recognition. Stuffer sequence: region of 38nts used to elongate the resultant probe.

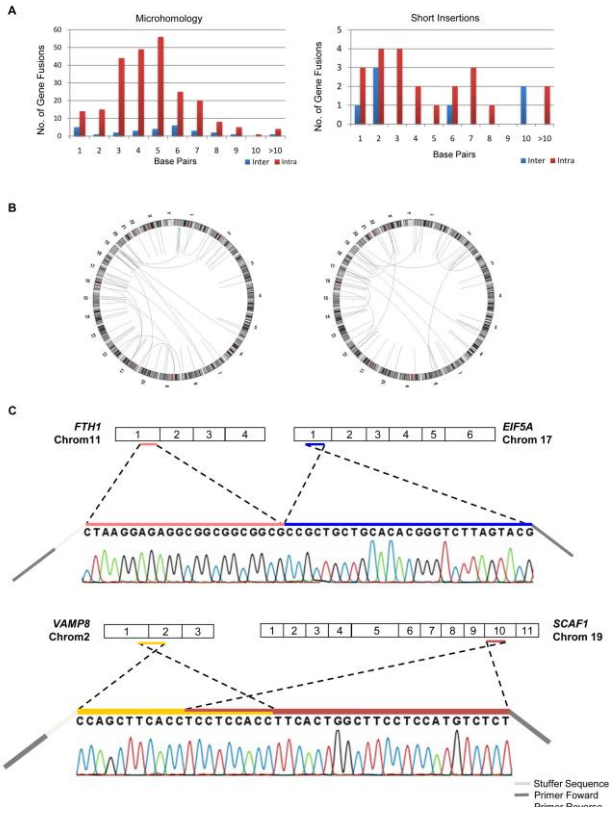


Figure 4. Discovery of alternative splicing variants. The 2,865 novel alternative splicing events detected in our approach are distributed according to the type of event reported. White squares represent the constitutive exons, and grey squares represent the alternative exons. The number of events is shown on the right side of each event type. (A) Intron retention showing the presence of one or more constitutive exons. (B) Alternative splice donor or acceptor site usage; (C) Alternative exon usage events were subclassified in exon skipping and exon inclusion events, showing the presence of one or both flanking constitutive exons

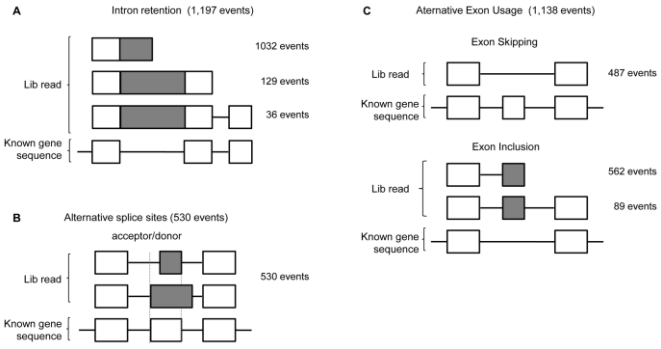


Figure 5. Partial sequences representation of the novel genes, its respective chromosome position and surrounding genes. Each validated candidate is shown individually. (A) Novel Gene 8 (NG8). (B) Novel Gene 7 (NG7). (C) Novel Gene 9 (NG9). The genomic coordinates of each novel gene are shown. Arrows represent the genomic localization of each gene and the direction in which it is transcribed. The red arrows represent the novel genes. In an expanded view, the genomic coordinates of NG are shown, as well as the conserved splice sites depicted in the introns and the *Dpn*II restriction sites.

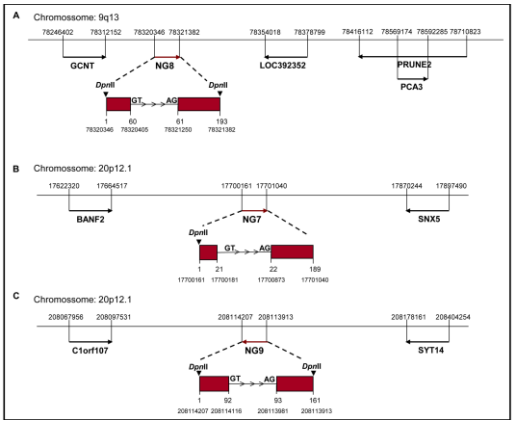
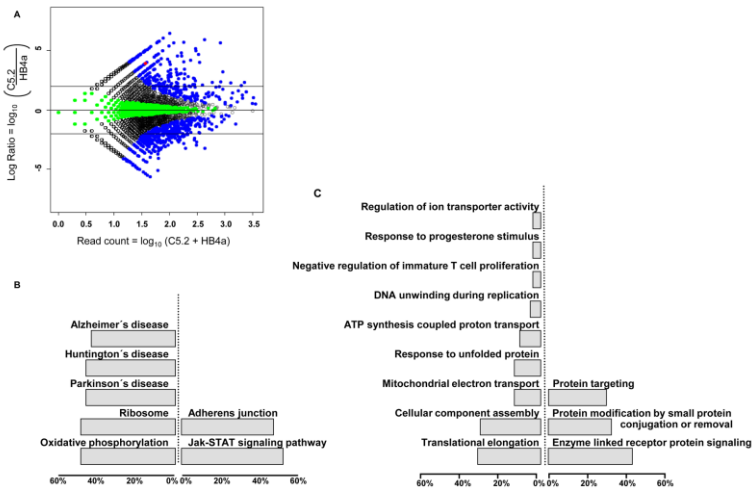


Figure 6. Differentially expressed genes between C5.2 and HB4a. (A) Relative gene expression between C5.2 and HB4a. The 2 black lines represent the cut-off value of $\log_2$ ratio ≥|2|- fold-change ≥|4|. The blue colored points correspond to genes with a BER equal to 0.0. *ERBB2* relative expression is identified by the red point. (B,C) Kegg (B) and GO Biological Process (C) enriched categories in differentially expressed genes between both cell lines. The bar corresponds to the percentage of differentially expressed genes in relation to all genes of our dataset annotated in the respective category.



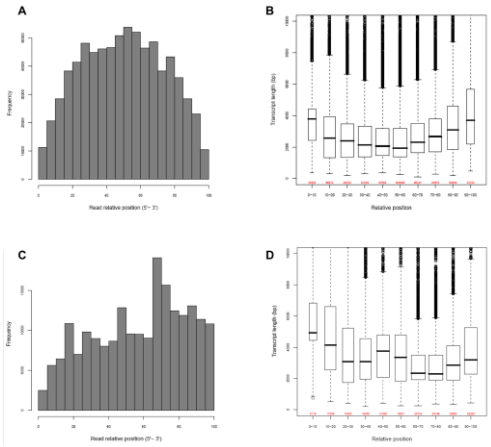
REFERENCES

1. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. (2000) Molecular portraits of human breast tumours. Nature 406: 747-752.
2. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 98: 10869-10874.
3. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression proffiling predicts clinical outcome of breast cancer. Nature 415: 530-535.
4. Brentani RR, Carraro DM, Verjovski-Almeida S, Reis EM, Neves EJ, et al. (2005) Gene expression arrays in cancer research: methods and applications. Crit Rev Oncol Hematol 54: 95-105.
5. Folgueira MA, Carraro DM, Brentani H, Patrão DF, Barbosa EM, et al. (2005) Gene expression profile associated with response to doxorubicin-based therapy in breast cancer. Clin Cancer Res 11: 7434-7443.
6. Castro NP, Osório CA, Torres C, Bastos EP, Mourão-Neto M, et al. (2008) Evidence that molecular changes in cells occur before morphological alterations during the progression of breast ductal carcinoma. Breast Cancer Res 10: R87. doi: 10.1186/bcr2157.
7. Rozenchan PB, Carraro DM, Brentani H, de Carvalho Mota LD, Bastos EP, et al. (2009). Reciprocal changes in gene expression profiles of cocultured breast epithelial cells and primary fibroblasts. I J Cancer 125: 2767-2777.
8. Koike FMA, Brentani H, Carraro DM, De Camargo BFM, Hirata KML, et al. (2009) Gene expression profile of residual breast cancer after doxorubicin and cyclophosphamide neoadjuvant chemotherapy. Oncol Rep 4:805-813.
9. Johnson JM, Castle J, Garrett-Engele P, Loerch PM, Armour CD, et al. (2003) Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. Science 302: 2141-2144.
10. Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, et al. (2006) Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics 7: 325. doi: 10.1186/1471-2164-7-325.
11. Norris AW, Kahn CR (2006) Analysis of gene expression in pathophysiological states: balancing false discovery and false negative rates. Proc Nat Acad Sci U S A 103: 649-653.
12. Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, et al. (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet 40: 1416-1425.
13. Du R, Tantisira K, Carey V, Bhattacharya S, Metje S, et al. (2009) Platform dependence of inference on gene-wise and gene-set involvement in human lung development. BMC Bioinformatics 10: 189. doi: 10.1186/1471-2105-10-189.
14. Torres TT, Metta M, Ottenwälder B, Schlötterer C (2008) Gene expression profiling by massively parallel sequencing. Genome Res 18: 172-177.
15. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, et al. (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Nat Acad Sci U S A. 106: 12353-12358.
16. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods 6: 377-382.
17. Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, et al. (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci U S A 106: 3264-3269.
18. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, et al. 2010. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463: 184-190.
19. Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. Nature protocols 3: 267-78.
20. Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. Trends Genet 17: 100-107.
21. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, et al. (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7 Suppl 1: S4.1-9.
22. Stamps AC, Davies SC, Burman J, O'Hare MJ (1994) Analysis of proviral integration in human mammary epithelial cell lines immortalized by retroviral infection with a temperature-sensitive SV40 T-antigen construct. I J Cancer. 57: 865-874.
23. Harris RA, Eichholtz TJ, Hiles ID, Page MJ, O'Hare MJ (1999). New model of ErbB-2 over-expression in human mammary luminal epithelial cells. Int J Cancer 80: 477-484.
24. Dias Neto E, Correa RG, Verjovski-Almeida S, Briones MR, Nagai MA, et al. (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc Natl Acad Sci U S A 97: 3491-3496.
25. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic AcidsRres 30: 3894-3900.
26. Gu W, Zhang F, Lupski JR (2008) Mechanisms for human genomic rearrangements.Pathogenetics. 1: 4.
27. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, et al. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30: e57.
28. Prifti E, Zucker J, Clement K, Henegar C (2008) FunNet: an integrative tool for exploring transcriptional interactions. Bioinformatics 24: 2636-2638.
29. Creighton CJ (2007) A gene transcription signature of the Akt/mTOR pathway in clinical breast tumors. Oncogene 26: 4648-4655.
30. Heinonen H, Nieminen A, Saarela M, Kallioniemi A, Klefström J, et al. (2008) Deciphering downstream gene targets of PI3K/mTOR/p70S6K pathway in breast cancer. BMC Genomics 9: 348. doi: 10.1186/1471-2164-9-348.
31. Akcakanat A, Zhang L, Tsavachidis S, Meric-Bernstam F (2009) The rapamycin-regulated gene expression signature determines prognosis for breast cancer. Mol Cancer 8: 75. doi: 10.1186/1476-4598-8-75.
32. Meric-Bernstam F, Gonzalez-Angulo AM (2009) Targeting the mTOR signaling network for cancer therapy. J Clin Oncol 27: 2278-2287.
33. Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, et al. (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. J Clin Oncol 25: 1239-1246.
34. Guffanti A, Iacono M, Pelucchi P, Kim N, Soldà G, et al (2009) A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics 10: 163. doi: 10.1186/1471-2164-10-163.
35. Davidson JM, Gorringe KL, Chin SF, Orsetti B, Besret C, et al (2000) Molecular cytogenetic analysis of breast cancer cell lines. B J Cancer 83: 1309-1317.
36. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, et al. (2009) Transcriptome sequencing to detect gene fusions in cancer. Nature 458: 97-101.
37. Wang XS, Prensner JR, Chen G, Cao Q, Han B, et al. (2009) An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. Nat Biotechnol 27: 1005-1011.
38. Di Fiore PP, Pierce JH, Kraus MH, Segatro O, King CR, et al. (1987) Erbb-2 is a potent oncogene when overexpressed in NIHI/3T3 Cells. Science 237: 178-181.
39. Slamon DJ, Godolphin W, Jones LA, Holt JA, Wong SG, et al. (1989) Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. Science, 244: 707-712.
40. Stankiewicz P, Lupski JR: Genome architecture, rearrangements and genomic disorders. Trends Genet 2002, 18:74-82.
41. Lee JA, Carvalho CM, Lupski JR: A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 2007, 131:1235-1247.
42. Srebrow A, Kornblihtt AR (2006) The connection between splicing and cancer. J Cell Sci. 119: 2635-2641.
43. Vander Heiden MG, Cantley LC, Thompson CB (2009) Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science 324: 1029-1033.
44. Brentani RR, Brentani M, Raw I (1967). Messenger activity

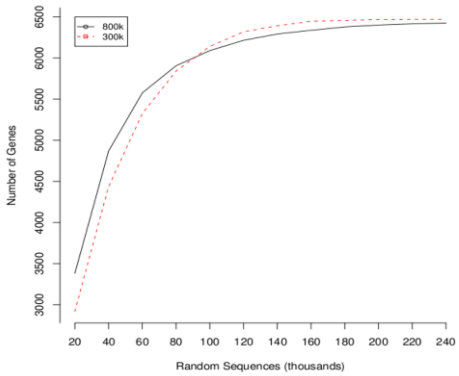
of purified RNA from rat liver nuclei. Nature 214: 1122-1123
45. Schneiter R, Kadowaki T, Tartakofft AM (1995) mRNA Transport in Yeast : Time to Reinvestigate the Functions of the Nucleolus. Mol Biol Cell 6: 357-370.
46. Ideue T, Azad AK, Yoshida J, Matsusaka T, Yanagida M, et al (2004) The nucleolus is involved in mRNA export from the nucleus in fission yeast. J Cell Sci 117: 2887-2895.
47. Da Silva A, Goldberg A, Barras E, Orlandi V, Salles J, et al (1974) Pharmacological inhibition of hormonal tyrosine amino transferase induction. Biochem Pharmacol 23: 2455-2457.
48. Stingl J, Caldas C. (2007) Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. Nat Rev Cancer 7: 791-799.
49. Allinen M, Beroukhim R, Cai L, Brennan C, Lahti-Domenici J, et al. (2004) Molecular characterization of the tumor microenvironment in breast cancer. Cancer Cell 6: 17-32.
50. Noblesse E, Cenizo V, Bouez C, Borel A, Gleyzal C, et al. (2004) Lysyl oxidase-like and lysyl oxidase are present in the dermis and epidermis of a skin equivalent and in human skin and are associated to elastic fibers. J Invest Dermatol 122: 621-630.
51. Min C, Kirsch KH, Zhao Y, Jeay S, Palamakumbura AH, et al. (2007) The tumor suppressor activity of the lysyl oxidase propeptide reverses the invasive phenotype of Her-2/neu-driven breast cancer. Cancer Res 67: 1105-1112.
52. Matz M, Shagin D, Bogdanova E, Britanova O, Lukyanov S, et al. (1999). Amplification of cDNA ends based on template-switching effect and step-out PCR. Nucleic Acids Res 27: 1558-1560.
53. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, et al. (2006) The UCSC Known Genes. Bioinformatics 22: 1036-1046.
54. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29: 308-311.
55. Vêncio RZ, Brentani H, Patrão DF, Pereira CA (2004) Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). BMC Bioinformatics 5: 119. doi: 10.1186/1471-2105-5-119

SUPPORTING INFORMATION

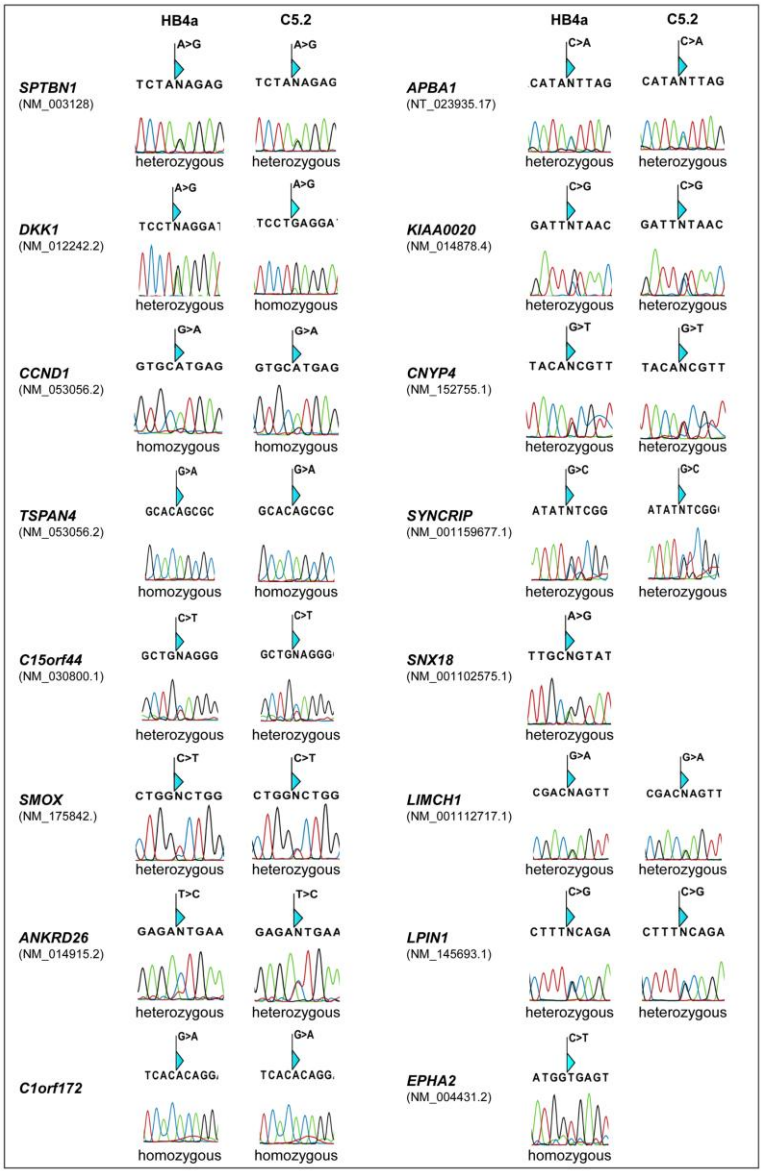
Supplementary Figure 1. Relative position frequency to RefSeq transcripts. (A) The frequency of reads distributed along transcript position from the Poly A+ library, where 0 is the 5'end and 100 is the 3'end of each corresponding transcript. (B) The relative transcript position of sequences from the Poly A+ library in relation to transcript size. The thickness of bars corresponds to the frequency of sequences in each group. (C,D) Data from the amplified library.



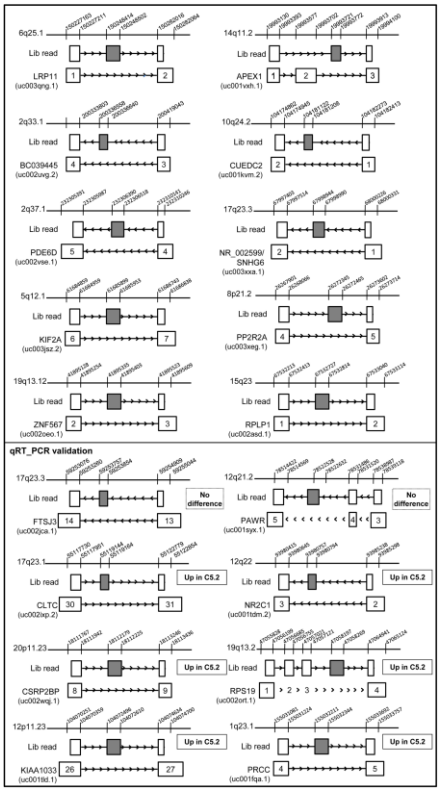
Supplementary Figure 2. Evaluation of overlap of genes represented by the 800K and 300K datasets. The number of genes represented by datasets of 20,000 randomly selected sequences from both datasets (300K and 800K) was compared. From the initial set of 20K sequences novel sets were generated by increasing 20,000 sequences each time, until 240,000 sequences were selected. X-axis: the number of randomly selected sequences in each dataset. Y-axis: the number of genes represented. The black solid line represents data from the 800K dataset and the red pointed line represents data from the 300K dataset.



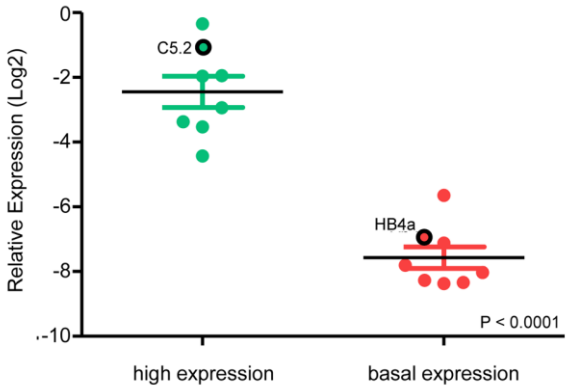
Supplementary Figure 3. Validation of novel SNPs. The eletropherogram represents the validation of the SNPs for each gene. The SNPs from the HB4a and C5.2 cell lines are shown separately and are classified as homozygous or heterozygous.



Supplementary Figure 4. Validation of alternative splicing variants by RT-PCR. Each validated AS event is represented by the genomic coordinates of each exon /intron border. The blank squares represent the constitutive exons and the grey squares represent the alternative exons. The gene symbol and corresponding RefSeq entry used as a reference are also shown. qRT_PCR validation: The 8 genes evaluated by RT_PCR are separated by the double line, and the results are shown inside the square as up- or down-regulation in the corresponding cell lines.



Supplementary Figure 5. Relative expression of *ERBB2* transcript in high expression and basal expression samples. Samples were classified as high or basal expression of ERBB2 according to protein and mRNA levels. Only samples with concordant results at both protein and transcript expression level were included in the study. We have compared the qRT-PCR expression results from high expression samples and C5.2 to basal expression samples and HB4a, confirming them as two distinct groups of samples concerning ERBB2 expression (p-value<0.0001). The black circuled points represent the expression level of the cell lines.



Supplementary Table 1. List of five transcripts (among the 15 most represented transcripts in amplified RNA-seq) and of two ribosomal RNAs in which an anchor region used for internal TS priming was found.

Gene	Number of reads	Anchor Region	Position
MALAT1	15.936	TS GGGCGGG I I I I I I Gene GGGCAGG	1322 - 1554
CYR61	8.809	TS GGGCGGG I I I I I I I Gene GGGCGGG	1446 - 1757
FN1	2.292	TS GGCGGG I I I I I Gene GGCAGG	4298 - 4528
gi|13994260	2.192	TS CGGGCGGG I I I I I I I Gene CGGCCGGG	150 - 345
TMEM49	1.902	TS GGGCGGG I I I I I I Gene GGGCAGG	1849 - 1973
18S	14.054	TS GGGCGGG I I I I I I I Gene GGGCGGG	268 - 592
28S	57.188	TS GGGCGGG I I I I I I I Gene GGGCGGG	1413 - 1705

Number of reads: RNA-seq reads related to respective gene identified in amplified library. Anchor Region: Alignment between TS-Primer and transcript sequence used as anchor region for internal priming. Position: Full-length relative position of anchor region.

Supplementary Table 2. Validation of SNPs.

Gene Symbol	mRNA location		Alteration	Sanger Sequencing	
	region/position	EXON	nts (aa)	Validation in HB4	Validation in C5.2
ANKRD26	CDS/4617	30	T>C (I>T)	T/C	T/C
APBA1	3' UTR/6186	13	C>A	C/A	C/A
C15ORF44	CDS/1524	10	C>T	C/T	C/T
C1ORF172	CDS/942	2	G>A	A/A	A/A
CCND1	3' UTR/3430	4	G>A	A/A?	A/A?
CNYP4	CDS/342	2	G>T (S>I)	G/T	G/T
DKK1	3' UTR/1037	4	A>G	G/A	G/G
EPHA2	CDS/2264	13	C>T	T/T	NA
EXOC6	CDS/1169	7	C>A (A>D)	C/C	C/C
KIAA0020	CDS/1036	9	C>G (L>V)	C/G	C/G
LIMCH1	3' UTR/4467	26	G>A	G/A	G/A
LPIN1	3' UTR/4826	20	C>G	C/G	C/G
SMOX	CDS/108	2	C>T	C/T	C/T
SNX18	3' UTR/4387	2	A>G	A/G	NA
SPG11	CDS/5587	30	A>G	A/A	A/A
SPTBN1	CDS/6734	33	A>G (K>R)	A/G	A/G
SYNCRIP	3' UTR/2309	11	T>G	T/G	T/G
TBC1D9B	3' UTR/4132	22	C>A	N/A	N/A
TSPAN4	CDS/765	7	G>A	A/A	A/A

The gene symbol is used to identify each selected SNP. The SNPs are localized according to untranslated region (5' and 3' UTR) or coding sequence (CDS) and mRNA coordinates. The nucleotide alteration is shown. The amino acid alteration is shown only for non-synonymous cases. Genotype identified for HB4a and C5.2 cell lines after Sanger sequencing is shown. ND: Not determined

Supplementary Table 3. Validation of gene fusion.

Chromosome	Gene Symbol	Microhomology	# of reads –total (hb4a/C5.2/undefined)
Chromosome 22	ATF4	No	144 (27/103/14)
Chromosome 12	STRAP		110 (42/57/11)
Chromosome X	KRT8	No	2610 (685/1653/272)
Chromosome 12	WDR45		9 (3/5/1)
Chromosome 11	RPLP2	No	228 (9/196/23)
Chromosome 8	PLEC1		106 (52/41/13)
Chromosome 11	FTH1	No	21 (0/17/4)
Chromosome 17	EIF5A		6 (0/5/1)
Chromosome 8	SDCBP	No	5 (2/0/3)
Chromosome 2	ATP5G3		294 (5/255/34)
Chromosome 8	PLEC1	No	106 (52/41/13)
Chromosome 11	FTH1		21 (0/17/4)
Chromosome 19	SLC27A5	Yes	13 (2/11/0)
Chromosome 1	RBM8A	8nts	0
Chromosome 14	C14orf147	Yes	70 (38/17/15)
Chromosome 17	CLTC	9nts	345 (128/175/42)
Chromosome 7	STK17A	No	9 (6/2/1)
Chromosome 19	RPS15		231 (67/148/16)
Chromosome 17	SHMT1	No	102 (39/59/4)
Chromosome 5	DIAPH1		120 (69/38/13)
Chromosome 2	VAMP8	Yes	19 (2/16/1)
Chromosome 19	SACF1	9nts	0
Chromosome 9	TXN	Yes	0
Chromosome 2	AFTPH	4nts	46 (24/19/3)
Chromosome 1	FAM36A	No	44 (10/24/10)
Chromosome 1	CAP1		226 (81/118/27)
Chromosome 17	CANT1	Yes	46 (17/27/2)
Chromosome 17	GEMIN4	7nts	25 (13/11/1)
Chromosome 16	MLYCD	Yes	20 (12/7/1)
Chromosome 16	CDH13	3nts	49 (2/46/1)

The gene fusions evaluated are characterized by the 2 chromosomes involved in the event as well as the corresponding genes. The specific hybridization sequence of the probes and the amplicon size expected after PCR are shown.

Supplementary Table 4. Alternative splicing events detected for each cell line.

AS event type	HB4a	C5.2
intron retention	511,21	494
exon inclusion	188,26	245
exon skipping	152,3	226
alternative aceptor donor	145,14	251

The number of alternative splicing events detected for each sample normalized by the total number of reads generated for each cell line.

Supplementary Table 5. Alternative splicing variants modulated by *ERBB2* expression.

Gene Symbol	Position	Inclusion between exons	Size of novel exon	Reported by EST	Fold-change C5.2/HB4a
CLTC	CDS	25 and 26	21bp	YES	3
CSRP2BP	CDS	8 and 9	47bp	NO	2
KIAA1033	CDS	26 and 27	115bp	NO	4
NR2C1	CDS	2 and 3	38bp	YES	2
RPS19	CDS	3 and 4	31bp	NO	3
PRCC	CDS	5 and 6	134bp	NO	3

The position of the novel exon identified is shown according to the number of the flanking exons. The expression level obtained by qRT-PCR is reported as fold-change between C5.2 and HB4a.

Supplementary Table 6. Validation of differential gene expression modulated by *ERBB2*.

	mRNA seq		qRT_PCR
Gene Symbol	**C5.2/Hb4a**		**C5.2/Hb4a**
ALDH2 *	14:0	14	41
ALDOA	1919:458	4	1
ANGPTL4 *	246:7	35	7
ANXA6 *	13:0	13	2
ATP5G3 *	685:14	49	2
ATP5L *	568:55	10	2
C12orf44 *	146:14	10	2
CAPG	720:96	8	1
CDC20	392:89	4	1
COL3A1 *	28:0	28	493
COPE	170:27	6	1
COX11 *	20:0	20	2
COX4I1 *	562:14	40	2
COX8A	1293:294	4	1
CSDA *	199:14	14	2
DST	35:212	-6	1
EEF1A1	4781:96	50	-1
EEF1B2	263:48	5	1
EIF4EBP1	199:41	5	1
ERBB2IP	18:137	-8	1
FAU	158:14	11	1
FBXL6 *	26:0	26	2
GALNT3 *	6:267	-45	-2
HMGA1 *	281:27	10	2
HMGB1 *	556:82	7	2
HMGB2	20:0	20	1
HMGN2	374:7	53	1
HRAS *	257:41	6	2
HSPA8 *	544:75	7	4
HSPE1 *	445:7	64	2
IL6ST	6:226	-38	1
JUP	23:68	-3	1
KRT15 *	1129:164	7	3
KRT19 *	9515:2345	4	2
KRT4 *	14:0	14	52
KRT6A *	492:96	5	3
KRT7 *	11141:2297	5	2
LAMB1	105:540	-5	1
LAMC1	222:971	-4	2
LAMC2	146:608	-4	1
LMNA *	339:48	7	2
LOX *	12:171	-14	-8
LRPAP1	35:0	35	1
MME *	0:27	-27	-11
NDUFA1	427:14	31	1
NDUFA13	948:157	6	1
NDUFA2	275:21	13	1
NDUFA5	13:0	13	1

NDUFB3 *	23:0	23	2
NDUFB8 *	1001:123	8	2
NDUFS7	334:14	24	1
NDUFS8	211:41	5	1
PDS5B	0:26	-26	1
PFKP *	749:123	6	4
PHB *	492:75	7	3
PHB2 *	4166:827	5	3
PTMA	46:0	46	1
PTMS	427:14	31	1
RAN *	351:21	17	2
RPL10A	2949:595	5	1
RPL29 *	1229:533	2	2
RPL31 *	152:14	11	2
RPL38	1006:96	10	1
RPL39 *	1375:321	4	2
RPL41 *	1042:21	50	2
RPL8 *	1691:48	35	3
RPLP1 *	2475:41	60	2
RPS13	860:103	8	1
RPS15A	714:68	11	1
RPS19	6267:1306	5	1
RPS24 *	363:62	6	2
RPS27A	35:0	35	1
RPS6	170:7	24	1
SEMA3C	25:0	25	1
SFN	374:7	53	1
SLC3A2 *	257:21	12	3
SOS2	6:301	-50	1
SOX15	15:0	15	1
STAT3	64:390	-6	1
TGFBR3	6:246	-41	1
TIMP1 *	357:7	51	2
TPI1 *	930:34	27	2
TUBB2C *	568:41	14	2
TXNIP *	275:2639	-10	-3
TXNRD2 *	146:14	10	2
UCRC *	30:0	30	2
UQCRB *	556:34	16	2
VEGFA	23:185	-8	1

The mRNA seq data is given as a fold-change between C5.2 and HB4a. When no reads were identified in the RNA-seq from one of the cell lines, we calculated fold-change by replacing “0” by “1”. Positive and negative values correspond to higher expression in C5.2 and HB4a, respectively. The qRT_PCR results are given as fold-change obtained by $2^{-\Delta\Delta CT}$. *genes validated in the qRT-PCR by the criterion for differentially expressed genes as Fold-change>2.

Supplementary Table 7. Effects of rapamycin treatment on genes influenced by *ERBB2*-mediated expression.

Gene Symbol	mRNA seq	Same cell lines	After Rapamycin treatment
ALDH2	14	41	nd
ANGPTL4	35	7	5 NO
ANXA6	13	2	2 NO
ATP5G3	49	2	-2 YES
ATP5L	10	2	6 NO
C12orf44	10	2	1 YES
COL3A1	28	493	Nd
COX11	20	2	Nd
COX4I1	40	2	-2 YES
CSDA	14	2	3 NO
FBXL6	26	2	1 YES
GALNT3	-45	-2	nd
HMGA1	10	2	1 YES
HMGB1	7	2	3 NO
HRAS	6	2	-2 YES
HSPA8	7	4	3 NO
HSPE1	64	2	2 NO
KRT15	7	3	-3 YES
KRT19	4	2	nd
KRT4	14	52	nd
KRT6A	5	3	-1620 YES
KRT7	5	2	-7 YES
LMNA	7	2	1 YES
LOX	-14	-8	-2 YES
MME	-27	-11	nd
NDUFB3	23	2	1 YES
NDUFB8	8	2	-2 YES
PFKP	6	4	1 YES
PHB	7	3	1 YES
PHB2	5	3	-3 YES
RAN	17	2	-1 YES
RPL29	2	2	-2 YES
RPL31	11	2	1 YES
RPL39	4	2	1 YES

The results from quantitative RT-PCR on rapamycin-treated cell lines for the 46 validated genes are given as fold-change between C5.2 and HB4a. Positive and negative values correspond to higher expression in C5.2 and HB4a, respectively. The fourth columns show the qRT_PCR results obtained from the C5.2 and HB4a cell lines and from the cells lines after rapamycin treatment. Response to rapamycin was considered when a decrease or inversion of fold-change between C5.2 and HB4a compared to non treated cell lines was observed. Yes and No represent response or no response to rapamycin, respectively. (nd) Ct not determined.

Supplementary Table 8. Novel alternative splicing variants.

Gene	Foward Primer	Reverse Primer	Amplicon
AET1G	CAGGTGTGTGCGAACAG	CTTCAGCTCAATCCCAATC	89
APEX1	GTTTGTCATTCCCTTGATG	CTCCTGCTGCCTCTTTGTC	83
BC039445	GTCGACCTCGCAACAG	CCTATGAAATAGTCTCGGC	116
CLTC	CCTAGAAACTGCATGGAG	CTTTTCCTTTATTGCATCAAC	103
CSRP2BP	GATTCATCCTGTTTTGCTTCTG	CCTTTGGCTTCATGGTTCC	131
CUEDC2	GGATTACAGGCATGAACC	GATGAGAGCTGCACCG	101
FLJ00150	CTCAGAAAGGGATAGTAGC	GTAGCCCAGGACAACCATG	104
FTJ3	GTCTGCTGCACTCATATCC	CCTGGCATCAAGCAATCC	95
KIAA1033	CATTCGTACATAGCCCATAGC	CAGTAAACAAAGTCTCTGTCC	92
KIF2A	GCAACAGCAAGAACTTAGAG	CTTCTAGGAAATAATACCACC	74
NR_002599	GCGAAGAGCCGTTAGTC	GAAAACAGAATTCAAGCTACTG	115
NR2C1	CAAGTGCTGTCACAATCTG	GTGGCAATAGAATCGGTAC	79
PAWR	CCACCTAGAACAGTTTCAG	CATTCTCTTCACCCTCCAAC	104
PDE6D	CCATGTGCCAAGTGAGTG	CCTAACTCCACAAATACCTG	103
PP2R2A	CTTTCAAGTTATACCCTTCTGG	GTATAGTGGAGAAGCCTGG	129
PRCC	CACCTAGTAGCTGAGAACAG	GTTGGCTGCTCACCTTTC	123
RPLP1	CATGGCCTCTGTCTCC	GGCAATTACACCGAAAGAG	108
RPS19	GTTCATCTTTCAGTCCTCAG	GCTTGCTCCCTACGATG	106
RWDD1	GAAAGCCAAGTTTGATG	CTCTTTTTCTGTGAATTTC	114
ZNF567	GCTCAGAAGACTCTATATATGG	CTGGGTAAGTGAAGACAC	124

The forward and reverse primer sequences used for each validation are shown with its corresponding gene symbol and the amplicon size.

Supplementary Table 9. Putative novel genes.

Chrom. location	Foward Primer	Reverse Primer	Amplicon
5p13.3	CCCACTTTGGTCTCCC	CTGCTTACAGTTCTTCATGC	147
6q23.2	TATATCGAATATTGTTAATAG	TTCACTGCAGTCTGG	182
20q13.13	CACGCCACTGCACTCC	CCTGACCTTTGTACATGCTG	144
9q21.13	CCTTCCATCTCAGCCTCC	CAGGAAGCTGGTATTCAAGAG	170
20p12.1	GATCAAAGAAGCCTCTGC	CACACCATACATGCTCTTC	240
1q32.2	GGTTCTAGTTTTGGTTCTTC	GGCTTATCTCTGTTGAATC	147

The chromosome localization of each putative novel gene is shown as well as the sequences of forward and reverse primers and the respective amplicon size.